Norface: Improving Facial Expression Analysis by Identity Normalization

Hanwei Liu[1,2,5*], Rudong An[2*], Zhimeng Zhang[2], Bowen Ma[2,5], Wei Zhang[2], Yan Song[2], Yujing Hu[2], Wei Chen[3], and Yu Ding[2,4,5*]

[1]Tongji University [2]Netease Fuxi AI Lab [3]Hebei Agricultural University
[4]School of Computer Science, Hangzhou Dianzi University [5]Happy Elements



Fig. 1: Identity normalization. The figure demonstrates that various original images (i.e. in-the-lab and in-the-wild ones) with AUs and emotions are normalized to a common target identity with a consistent pose, background, etc.

Abstract. Facial Expression Analysis remains a challenging task due to unexpected task-irrelevant noise, such as identity, head pose, and background. To address this issue, this paper proposes a novel framework, called Norface, that is unified for both Action Unit (AU) analysis and Facial Emotion Recognition (FER) tasks. Norface consists of a normalization network and a classification network. First, the carefully designed normalization network struggles to directly remove the above task-irrelevant noise, by maintaining facial expression consistency but normalizing all original images to a common identity with consistent pose, and background. Then, these additional normalized images are fed into the classification network. Due to consistent identity and other factors (e.g. head pose, background, etc.), the normalized images enable the classification network to extract useful expression information more effectively. Additionally, the classification network incorporates a Mixture of Experts to refine the latent representation, including handling the input of facial representations and the output of multiple (AU or emotion) labels. Extensive experiments validate the carefully designed framework with the insight of identity normalization. The proposed method outperforms existing SOTA methods in multiple facial expression analysis tasks, including AU detection, AU intensity estimation, and FER tasks, as well as their cross-dataset tasks. For the normalized datasets and code please visit project page.

Keywords: Facial emotion recognition · Action Unit detection · AU intensity estimation · Identity normalization

* Equal contribution. Yu Ding is the corresponding author.

1 Introduction

Facial Expression Analysis (FEA) is a complex task in affective computing that often involves facial Action Units (AU) analysis (AU detection and AU intensity estimation) and Facial Emotions Recognition (FER). Despite the progress made so far [6,9,30,37,59,74], both FEA tasks still face challenges due to the inherent entanglement of expressions with unexpected factors like identity, head pose, and background [32, 46].

Specifically, FEA suffers from severe identity bias [32,33], which exacerbates the intricacies of refinement and representation of facial expressions [27, 78]. FEA model struggles to generalize effectively to unseen identities, this bias also leads the model to overfit seen identities and undermines the overall performance [25,26]. In addition, variations in pose and background compound the challenges in FEA tasks [2,30]. Hence, a crucial objective of FEA is to mitigate interference arising from irrelevant factors, including identity, pose, and background.

To address these challenges, some studies attempt to construct identity-based expression pairs [26, 27, 65, 68, 93] to obtain identity-invariant representations. Due to significant variations in identities, the issue of identity bias still persists. Alternatively, certain studies aim to incorporate generative techniques, such as generating 'neutral' emotions [25,78,80], 'average' identities [4], or diverse emotions [1,22,76], to disentangle expression from identity. Nonetheless, these methods either suffer from limited generation quality or heavily rely on controlled lab datasets, thereby restricting their generalizability in complex and diverse real-world scenarios. Moreover, these approaches often overlook non-identity noise factors, such as pose and background variations. Additionally, their frameworks are often tailored to either AU analysis or FER tasks, which cannot be unified for both tasks, despite task-irrelevant noise being common to both.

To address the above challenges, this paper introduces a novel framework called **Norface**, which is applicable to AU detection, AU intensity estimation, and FER tasks. The core concept behind Norface is **Identity normalization (Idn)**, which normalizes all images to a common identity with consistent pose and background, as shown in Fig.1. By doing so, the resulting normalized data is intended to retain only the facial expression variations that are relevant to the task. Consequently, Idn mitigates the influence of identity bias, pose variations, and background changes, which are commonly encountered in both AU analysis and FER tasks. This process can also be interpreted as the removal of the above task-irrelevant noise, which is carried out by a normalization network in the first stage. Then, in the second stage, normalized images, as complementary to the original images, are used to improve expression analysis (AU detection, AU intensity estimation, or emotion recognition) by a classification network.

Specifically, in the first stage, the normalization network achieves Idn from all original faces to a target face. Thus, our normalization network is not only suitable for in-the-wild data but also capable of meeting the high-quality requirement for expression consistency. Specifically, we employ a pre-trained Masked AutoEncoder (MAE) to extract facial features, effectively capturing both expression and other attributes. Then, we develop an Expression Merging Module

(EMM) to adaptively merge expression features from the original faces into the target face. Moreover, we introduce an expression loss and an eyebrow loss to enforce expression consistency between the normalized faces and the original faces.

In the second stage, the classification network relies on Mixture of Experts (MoE). MoE aims to learn several expert sub-branches which are automatically trained to dynamically activate task-specific experts. This work proposes Input and Output MoE modules, where the input MoE module can refine latent facial representation, and the output MoE module can facilitate the detection, estimation, or recognition of multiple AU or emotion labels.

Additionally, there are several multi-task methods [12, 29, 92, 103] that combine AU detection and FER tasks, requiring co-annotated AU and emotion labels, which are often unavailable in most datasets. Unlike these multi-task methods, our Norface offers a unified framework for these tasks individually, without the requirement for co-annotated AU and emotion labels, thus avoiding the issue of task imbalance [66, 98] being common in those multi-task methods.

Our work is the first to develop a unified framework to address noise issues in three tasks of AU detection, AU intensity estimation, and FER. Our Norface outperforms the existing results on the three tasks, as well as their cross-dataset tasks, showcasing its superior performance. To sum up, ours makes the following contributions:

- This work provides a novel insight and releases a normalized dataset to remove the task-irrelevant noise by identity normalization that normalizes all expression images to a common identity with consistent pose and background.
- This work proposes a new unified Facial Expression Analysis framework named Norface, designed for AU detection, AU intensity estimation, or FER tasks. This framework addresses identity normalization using Masked AutoEncoder and representation refinement based on Mixture of Experts.
- Extensive quantitative and qualitative experiments validate the effectiveness of Norface, surpassing existing methods in AU detection, AU intensity estimation, and FER tasks, as well as their cross-dataset tasks.

2 Related Work

Identity-Invariant FEA. Identity bias has always been a challenge for FEA tasks, as expressions and identities are inherently entangled. Some methods [26, 27, 65, 68, 93] construct pairs of identity-based expressions to obtain identity-invariant representations. DLN [93] and GLEE-Net [94] construct triples mapping identity to facial expression similarity for performing FER and AU detection tasks, respectively. IdenNet [65] leverages face clustering to extract identity-dependent features and perform AU detection. Moreover, synthetic images [1, 4, 22, 25, 76, 78, 80] have been employed to decouple identity and expression. DeRL [78] generates neutral facial images to represent lab facial images. Huang

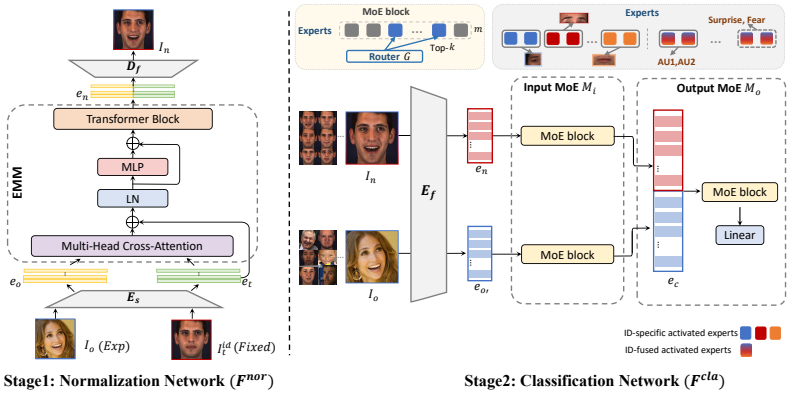


Fig. 2: Structure of the proposed Norface. Norface comprises two key stages: identity normalization and expression classification. In the first stage, the normalization network F^{nor} normalizes all original faces to a common identity with a fixed pose and background, thereby reducing the influence of task-irrelevant factors. In the second stage, a Mixture of Expert based classification network F^{cla} receives both normalized and original faces and performs AU detection, AU intensity estimation, or FER tasks.

et al. [22] synthesize basic emotions for each identity. Differing from these approaches, ours focuses on identity normalization to address identity bias.

AU analysis and FER tasks. Current AU analysis and FER methods often rely on auxiliary information to mitigate the interference of task-irrelevant noise, such as facial landmarks [35, 51, 74, 103], emotional priors [31, 64, 102], AU descriptions [8, 10, 75, 79], or patch-based learning methods [23, 38, 39, 57, 104]. SEV-Net [79] utilizes both semantic descriptions and visual features of AU to enhance AU detection. LA-Net [74] incorporates facial landmarks to augment expression features. However, they still encounter interference from such noise as they rarely explicitly eliminate task-irrelevant noise.

Expression reenactment. Identity normalization aims to normalize the expressions of all single-frame faces onto a fixed identity, pose, and background, which is distinct from popular expression reenactment. Currently, most expression reenactment methods [20, 45, 61, 71, 72, 101] cannot achieve identity normalization because they rely on consecutive frames, which cannot be applied to single-frame datasets. Additionally, while a few methods [44, 54, 81, 82, 91] can achieve identity normalization, the generated quality is often limited, making it challenging to meet the demand for high-quality expression consistency in identity normalization.

3 Method

As shown in Fig.2, Norface comprises two stages: the first stage employs a normalization network F^{nor} for Identity normalization (Idn), while the second stage utilizes a classification network F^{cla} for expression classification. Let I_o be the

original face belonging to a given AU or emotion dataset, and I_t^{id} be the given fixed target face I_t with identity id. Firstly, F^{nor} performs Idn on I_o and I_t^{id}, resulting in a normalized face I_n that maintains consistency in terms of identity, pose, background with I_t^{id}, and expression with I_o. Then, F^{cla} takes both I_n and I_o as input and then performs FEA tasks.

3.1 Normalization network

Normalization network F^{nor} performs Idn on the original face I_o and the target face I_t. As shown in Fig.2, a shared face encoder E_s first maps I_o and I_t into patch embeddings e_o and e_t. Then, an Expression Merging Module (EMM) is designed to adaptively fuse the expression information of I_o and other attribute information of I_t. Finally, a facial decoder D_f outputs the normalized face I_n.

Shared Face Encoder. We design a shared face encoder following MAE [19], to project original and target faces to a common latent representation. The MAE is pre-trained on a large-scale face dataset using a mask training strategy, which has been validated in many domains due to its superior feature extraction performance [3, 53]. By forcing the model to learn the topological and semantic information of masked image patches, the latent space of the MAE can better capture facial expressions and attributes. That is, based on the pre-trained face encoder E_s, we can project a original/target face $I_{o/t}$ into a latent presentation:

$$e_{o/t} = E_s(I_{o/t}) \tag{1}$$

where $e_{o/t} \in R^{N \times L}$. N and L denote the number of patches and the dimension of each embedding, respectively.

Expression Merging Module. We designed an Expression Merging Module (EMM) that adaptively integrates the expression information of the original face with the other attributes of the target face. As shown in Fig.2, EMM consists of a multi-head cross-attention block and two transformer blocks [15]. Given the original patch embeddings e_o and the target patch embeddings e_t, we first compute Q, K, V for each patch embedding in e_o and e_t. Then the cross-attention can be formulated as $\mathrm{CA}\left(Q_t, K_o\right) = \mathrm{softmax}\left(\frac{Q_t {K_o}^T}{\sqrt{d_k}}\right)$, where CA represents Cross Attention, Q_*, K_*, V_* are predicted by attention heads, and d_k is the dimension of K_*. Next, the expression information from the original face is aggregated based on the computed CA:

$$V_{fu} = \mathrm{CA} * V_o + V_t \tag{2}$$

Then, V_{fu} are normalized by a layer normalization (LN) and processed by multi-layer perceptrons (MLP). The fused embeddings e_{fu} are further fed into two transformer blocks to obtain the output e_n. Finally, we utilize a convolutional decoder D_f to generate the normalized face I_n from e_n.

Training Loss. The use of multiple loss functions is a pervasive way to constrain the image generation, e.g. those works [40, 42, 69, 73, 86–89] on face manipulation, as these loss functions carry out supervision from perspectives. Inspired

by them, we also employ multiple loss functions to ensure high expression consistency and fidelity in identity normalization. Specifically, the adversarial loss, the reconstruction loss, and the perceptual loss supervise the holistic image; an identity loss and a landmark loss are used to dominate the identity information; an expression loss and an eyebrow loss guide the expression transferring. They are formulated below.

$$\begin{aligned}\mathcal{L}^{nor} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{perc}\mathcal{L}_{perc} \\ + \lambda_{id}\mathcal{L}_{id} + \lambda_{lm}\mathcal{L}_{lm} + \lambda_{exp}\mathcal{L}_{exp} + \lambda_{eye}\mathcal{L}_{eye}\end{aligned} \tag{3}$$

where λ_{rec}, λ_{perc}, λ_{id}, λ_{lm}, λ_{exp}, λ_{eye} are hyperparameters for each term. During the training phase, the choice of I_t is as random as I_o, which forces the normalization network to adapt to the variations in facial attributes.

Adversarial Loss. The adversarial loss is used to make the normalized images more realistic. We apply the hinge version adversarial loss [36] for training:

$$\mathcal{L}_{adv} = -\mathbb{E}\left[D^{res}\left([I_t, I_n]\right)\right] \tag{4}$$

where D^{res} is the discriminator to distinguish between real and fake samples.

Reconstruction Loss. Since there is no ground-truth for face normalization results, we force $I_o = I_t$ with a certain probability during training, and introduce a pixel-level reconstruction loss:

$$\mathcal{L}_{rec} = \|I_n - I_t\|_2 \tag{5}$$

Perceptual Loss. Since high-level feature maps contain semantic information, we employ the feature maps from the last two convolutional layers of pre-trained VGG [5] as the facial attribute representation. The loss is formulated as:

$$\mathcal{L}_{perc} = \|VGG(I_t) - VGG(I_n)\|_2 \tag{6}$$

Identity Loss. The identity loss is to constrain I_n's identity information to be consistent with I_t:

$$\mathcal{L}_{id} = 1 - cos\left(E_{id}\left(I_t\right), E_{id}\left(I_n\right)\right) \tag{7}$$

where E_{id} denotes a face recognition model [13] and *cos* denotes the cosine similarity.

Landmark Loss. To enhance facial contour consistency of I_t and I_n, we first use the pre-trained facial landmark detector [63] to predict the facial landmarks of I_t and I_n, and then only apply loss to the facial contour, as follows:

$$\mathcal{L}_{lm} = \|P_t - P_n\|_2 \tag{8}$$

where P_t and P_n are the embeddings of the facial contours of I_t and I_n, respectively.

Expression Loss. To ensure that I_o and I_n express similar movements of full face, we use a novel fine-grained expression loss [93] to penalize the L_2 distance of expression embeddings of I_o and I_n:

$$\mathcal{L}_{exp} = \left\|E_{exp}\left(I_o\right) - E_{exp}\left(I_n\right)\right\|_2 \tag{9}$$

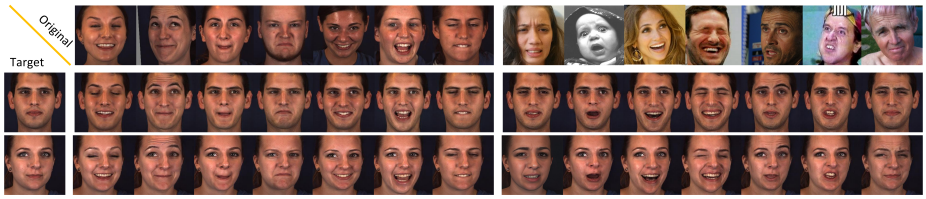


Fig. 3: Identity normalization of AU and emotion datasets as original faces. The figure presents results on target faces with two identities.

where $\| * \|$ denotes the euclidean distance.

Eyebrow Loss. Since the network tends to be insensitive to subtle changes in facial eyebrows, to further supervise the expression consistency, we use expression blend shapes from 3DMM [100], and penalize the L_2 distance in the channel associated with eyebrow movements:

$$\mathcal{L}_{eye} = \|E_{eye}(I_o) - E_{eye}(I_n)\|_2 \tag{10}$$

Inference. During inference, we set I_o from a given AU or emotion dataset and provide a fixed target face I_t^{id}. As a result, we obtain normalized face I_n that maintains consistent facial expressions as I_o and ensures consistency with the identity, pose, and background of I_t^{id}. Fig.3 intuitively showcases Idn results on some in-the-lab and in-the-wild AU and emotion data, and all normalized data* has been released.

3.2 Classification Network

The classification network F^{cla} is designed to perform AU detection, AU intensity estimation, or emotion recognition. As shown in Fig.2, F^{cla} takes I_n as input, as well as I_o. Specifically, F^{cla} consists of a facial feature extractor E_f, an Input MoE module M_i, and an Output MoE module M_o. First, E_f maps I_n and I_o to their latent presentations e_f, where $e_f \in \{e_n, e_{o\prime}\}$. Then, e_f are fed into M_i, which consists of two identity-specific Mixture of Expert (MoE) [43] blocks, resulting in the identity-specific facial representation e_h. Finally, all of e_h are input into M_o, which refines an identity-fused representation from fixed and original identities for the final classification. E_f, M_i, and M_o will be further described below, as well as the MoE block.

E_f aims to map I_n and I_o to a latent representation e_f. The design of E_f is similar to E_s and it is also pre-trained on the same large-scale face datasets. That is, based on E_f, we can project $I_{n/o}$ into a latent presentation e_f, that is $e_f = E_f(I_{n/o})$.

M_i consists of two identity-specific MoE blocks, with each MoE block receiving the individual latent representations e_f of I_n and I_o, respectively, that is $e_h = M_i(e_f)$.

* Dataset link. https://norface-fea.github.io/.

M_o combines the latent representations e_h of all identities to induce label-related information. First, all e_h are concatenated to obtain the embedding e_c. Then, e_c is input into an MoE block to obtain the identity-fused latent representation. Finally, we use a linear layer to output the AU detection, AU intensity estimation, or FER result.

Mixture of Experts. As mentioned above, MoE is used multiple times in M_i and M_o. The MoE divides a model into a group of experts, with each expert learning individual roles, allowing for collaboration when it contributes to the task. The sparsity of these experts enhances the model's inductive capabilities [43, 58], thereby effectively refining the latent representations. Specifically, the MoE in M_i focuses on refining the facial representation from identity, while the MoE in M_o fuses these representations from identities and focuses on refining representations relevant to different task labels.

In our design, MoE block learns m experts $E_{xp} \in \{E_{xp_1}, E_{xp_2}, ..., E_{xp_m}\}$ and a routing network G, where each expert is an MLP layer and multiple experts consist into a group of parallel MLP layers. By training, each expert plays individual roles but collaboratively performs their common task. In inference, G assigns weights for each expert according to the MoE input.

Formally, for given input presentation x, the noisy [58] Top-k routing network G, with parameters W_g and W_{noise}, predicts the gating weights for x and selects the Top-k experts from the m available experts to contribute to the final representation, that is $G(x) = \text{Topk}(\text{Softmax}(xW_g + \mathcal{N}(0,1)\text{Softplus}(xW_{noise})))$.

The final representation y of the k activated experts are linearly combined according to the gating weights, that is $y = \sum_{k=1}^{m} G^k(x) E_{xp_k}(x)$.

Training Loss. To implement expression analysis tasks, we employ four loss functions to train our F^{cla}:

$$\mathcal{L}^{cla} = \mathcal{L}_{cla} + \lambda_{imp}\mathcal{L}_{imp} + \lambda_{g\&l}\left(\mathcal{L}_{global} + \mathcal{L}_{local}\right) \tag{11}$$

where λ_{imp} and $\lambda_{g\&l}$ are hyperparameters for each term. $\mathcal{L}_{cla}$ is the cross entropy loss for AU detection and emotion recognition, or the L2 loss for AU intensity estimation. The importance loss $\mathcal{L}_{imp}$ is a classical solution [18, 58] in MoE, used to prevent the collapse caused by the Top-k strategy. The global & local loss $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$ are employed to enhance the optimal utilization of experts [49].

Importance Loss. The importance loss $\mathcal{L}_{imp}$, promotes a balanced distribution of the gating weights among the experts. More formally, for a given input x in the batch B, the loss $\mathcal{L}_{imp}$ is defined by the squared coefficient of variation:

$$\mathcal{L}_{imp} = \left(\frac{\text{std}(G(x))}{\text{mean}(G(x))}\right)^2 \tag{12}$$

Global & local Loss. To utilize each expert optimally, the global loss $\mathcal{L}_{global}$, maximizes the marginal entropy, aiming for wide expert use from a global perspective. On the other hand, the local loss $\mathcal{L}_{local}$, encourages low entropy in focused routing weights, fostering specialization of each expert in specific tasks

from a local perspective. Formally, $p(E_{xp}|x) \in \mathbb{R}^E$ is the probability distribution of the router over E_{xp} experts, $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$ are defined as:

$$\mathcal{L}_{global} = -\mathcal{H}\left(\widetilde{p}\left(E_{xp}|x\right)\right), \mathcal{L}_{local} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{H}\left(p\left(E_{xp}|x_i\right)\right) \tag{13}$$

where $\widetilde{p}(E_{xp}|x) = \frac{1}{B}\sum_{i=1}^{B}(p(E_{xp}|x_i))$ is the expert probability distribution averaged over the input x, and $\mathcal{H}(p)$ denotes the entropy.

4 Experiment

Datasets. Previous works on collecting datasets (e.g., BP4D [95], BP4D+ [99], DISFA [47], AffectNet [48] and RAF-DB [34]) make significant contributions to academic communities. We evaluate the AU analysis task on BP4D, BP4D+, DISFA and the FER task on AffectNet and RAF-DB. **BP4D** consists of 328 video clips from 41 participants. About 140,000 frames are annotated with the occurrence or absence of 12 AUs for AU detection and the intensity of 5 AUs for AU intensity estimation. **BP4D+** contains 1,400 video clips from 140 participants. About 198,000 frames are annotated with the same AUs in BP4D for AU detection. **DISFA** contains about 131,000 frames from 27 video clips. Each frame of DISFA is annotated with the occurrence or absence of 8 AUs for AU detection and the intensity of 12 AUs for AU intensity estimation. **AffectNet**, we use 7 basic emotions, selecting approximately 280,000 and 3,500 images in total for training and testing, respectively. **RAF-DB**, the experiment's training set and test set sizes are 12,271 and 3,068, respectively.

For fair comparisons, our experiments use the same training and testing datasets, as well as evaluation criteria as other methods. In the *AU detection* task, our evaluation follows the prior works [35, 57], using three-fold cross-validation, and the report results are the F1 averages from three-fold experiments. For *AU intensity estimation* task, we evaluate the model performance with intra-class correlation (ICC) [60], mean squared error (MSE), and mean absolute error (MAE). We report the accuracy in the *FER* task.

To pick up the appropriate target face, we intentionally select 10 individuals from BP4D&BP4D+, ensuring diversity in age, gender, and ethnicity, and used the individual with the best AU detection performance as the fixed target face I_t^{id} that can be seen in Fig.1.

Implementation Details. The normalization network F^{nor} is first pre-trained on CelebA-WebFace [41] and Emo135 [7] datasets. The face images are aligned and cropped to a size of 256×256. Afterward, we fine-tune the network on the target AU and emotion datasets. We adopt Adam [28] optimizer with $\beta_1 = 0$, $\beta_2 = 0.99$, a learning rate of 0.0001, and a batch size of 8. We set $\lambda_{rec} = 10$, $\lambda_{perc} = 5$, $\lambda_{id} = 10$, $\lambda_{lm} = 5000$, $\lambda_{exp} = 5000$, and $\lambda_{eye} = 10$. For the classification network F^{cla}, we performed fine-tuning for 40 epochs, and set $\lambda_{imp} = 0.001$, $\lambda_{g\&l} = 0.001$. The settings for each MoE block are $m = 4$ and $k = 2$. The base learning rates are set to $1e-4$, $1e-3$, and $2e-5$ for AU detection, AU intensity estimation, and FER tasks, respectively.

Table 1: Comparison of AU detection results on BP4D and BP4D+ in terms of F1 scores (%).

Dataset	Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg.	Δ ↑
BP4D	LP-Net [51]	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0	0.0
	JAA-Net [57]	53.8	47.8	58.2	78.5	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4	+1.4
	HMP-PS [62]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	[49.9]	54.5	63.4	+2.4
	SEV-Net [79]	58.2	[50.4]	58.3	**81.9**	73.9	**87.8**	87.5	61.6	52.6	62.2	44.6	47.6	63.9	+2.9
	MHSA-FFN [23]	51.7	49.3	[61.0]	77.8	79.5	82.9	86.3	[67.6]	51.9	63.0	43.7	56.3	64.2	+3.2
	CaFNet [9]	55.1	49.3	57.7	78.3	78.6	85.1	86.2	67.4	52.0	64.4	48.3	56.2	64.9	+3.9
	Chang et al. [6]	53.3	47.4	56.2	79.4	**80.7**	85.1	[89.0]	67.4	[55.9]	61.9	48.5	49.0	64.5	+3.5
	GLEE-Net [94]	[60.6]	44.4	61.0	[80.6]	78.7	85.4	88.1	64.9	53.7	[65.1]	47.7	[58.5]	[65.7]	+4.7
	Norface (Ours)	**60.9**	**55.3**	**67.4**	79.7	[79.9]	[87.5]	**90.7**	**69.8**	**58.1**	**66.2**	**54.9**	**61.8**	**69.3**	**+8.3**
BP4D+	ML-GCN [11]	40.2	36.9	32.5	84.8	88.9	89.6	89.3	81.2	[53.3]	43.1	55.9	28.3	60.3	0.0
	MS-CAM [83]	38.3	37.6	25.2	85.0	**90.9**	[90.9]	89.0	81.5	**60.9**	40.6	58.2	28.0	60.5	+0.2
	SEV-Net [79]	47.9	40.8	31.2	[86.9]	87.5	89.7	88.9	[82.6]	39.9	**55.6**	[59.4]	27.1	61.5	+1.2
	GLEE-Net [94]	**54.2**	**46.3**	[38.1]	86.2	87.6	90.4	[89.5]	81.3	46.3	47.4	57.6	[39.6]	[63.7]	+3.4
	Norface (Ours)	[52.2]	[46.0]	**51.7**	**88.1**	[89.0]	**91.3**	**90.1**	**83.3**	50.4	[51.1]	**61.6**	**45.4**	**66.7**	**+6.4**

Table 2: Comparison of AU detection results on DISFA in terms of F1 scores (%).

Method	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Avg.	Δ ↑
LP-Net [51]	29.9	24.7	72.7	[46.8]	49.6	72.9	93.8	56.0	56.9	0.0
JAA-Net [57]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5	+6.6
GLEE-Net [94]	61.9	54.0	**75.8**	45.9	[55.7]	[77.6]	92.9	60.0	65.5	+8.6
LGRNet [17]	[62.6]	[64.4]	72.5	46.6	48.8	75.7	[94.4]	**73.0**	[67.3]	+10.4
Norface (Ours)	**76.4**	**66.1**	[74.2]	**58.5**	**57.2**	**81.7**	**97.6**	[69.6]	**72.7**	**+15.8**

4.1 Comparison with State-of-the-Art Methods

We verify the effectiveness of our method on AU detection, AU intensity estimation, and facial emotion recognition tasks, as well as their cross-dataset tasks.

AU detection task. Tab.1 and 2 report comparison results with those previous methods. Our method outperforms those baselines on all datasets with the average of AUs detection. Specifically, ours exceeds the SOTA of GLEE-Net [94] for over 3% on both BP4D and BP4D+ datasets, and exceeds the SOTA of LGRNet [17] for over 5% on the DISFA dataset, which validates our method on the AU detection task.

AU intensity estimation. Tab.3 report comparison results with those previous methods. Ours surpasses all existing approaches in terms of MSE and MAE, and exceeds the SOTA of APs [56] by 0.19 in terms of ICC on the DISFA dataset, which validates our method on the AU intensity estimation task.

Facial emotion recognition task. Tab.4 reports comparison results with previous methods. Ours outperforms these baselines in terms of accuracy on both datasets. Specifically, ours exceeds the SOTA of LA-Net [74] for 1.41% and 1.09% on RAF-DB and AffectNet, respectively, validating our method on the FER task.

Cross-dataset task. Tab.5 and 6 report comparison results of AU detection and FER with previous methods on cross-dataset tasks, respectively. Ours achieves the best performance, surpassing the SOTA of GLEE-Net [94] by 5.1% on AU detection, and outperforming RANDA [24] by 6.09% and 11.2% on the 'R→A' and 'A→R' tasks of FER, respectively. Due to the normalization of task-irrelevant noise, domain barriers between datasets are reduced, which highlights the benefits of our method for cross-domain tasks.

The above experiments validate our framework being unified for those tasks, as identity normalization addresses the interference of task-irrelevant noise, such as identity, pose, and background, etc. More experiments on identity normalization are described below.

Table 3: Comparison of AU intensity estimation results on BP4D and DISFA.

Metric	Method	AU on BP4D						AU on DISFA												
		6	10	12	14	17	Avg.	1	2	4	5	6	9	12	15	17	20	25	26	Avg.
ICC ↑	ISIR [50]	.79	.80	.86	.71	.44	.72	-	-	-	-	-	-	-	-	-	-	-	-	-
	HR [52]	.82	.80	.86	.69	.51	.73	.35	.19	.78	.73	.52	.65	.81	.49	.61	.28	.92	.67	.58
	SCC-Heatmap [16]	.74	.82	.86	.68	.51	.72	.73	.44	.74	.06	.27	.51	.71	.04	.37	.04	.94	.78	.47
	APs [56]	.82	.80	.86	.69	.51	**.74**	.35	.19	.78	.73	.52	.65	.81	.49	.61	.28	.92	.67	.58
	Norface (Ours)	.81	.74	.90	.50	.74	**.74**	.72	.68	.77	.68	.59	.56	.87	.54	.65	.34	.96	.70	**.67**
MSE ↓	ISIR [50]	.83	.80	.62	1.14	.84	.85	-	-	-	-	-	-	-	-	-	-	-	-	-
	HR [52]	.68	.80	.79	.98	.61	.78	.41	.37	.70	.08	.44	.30	.29	.14	.26	.16	.24	.39	.32
	APs [56]	.72	.84	.60	1.13	.57	.77	.68	.59	.40	.03	.49	.15	.26	.13	.22	.20	.35	.17	.30
	Norface (Ours)	.71	.95	.50	1.02	.47	**.73**	.22	.19	.46	.03	.41	.30	.24	.09	.17	.14	.15	.29	**.22**
MAE ↓	BORMIR [97]	.85	.90	.68	1.05	.79	.85	.88	.78	1.24	.59	.77	.78	.76	.56	.72	.63	.90	.88	.79
	KBSS [96]	.65	.65	.48	.98	.63	.66	.48	.49	.57	.08	.26	.22	.33	.15	.44	.22	.43	.36	.33
	SCC-Heatmap [56]	.61	.56	.52	.73	.50	.58	.16	.16	.27	.03	.25	.13	.32	.15	.20	.09	.30	.32	.20
	Norface (Ours)	.48	.56	.39	.82	.40	**.53**	.14	.12	.30	.03	.29	.17	.22	.09	.14	.08	.17	.26	**.17**

Table 4: Comparison of facial emotion recognition results (%).

Method	RAF-DB		AffectNet	
	Acc.	Δ ↑	Acc.	Δ ↑
LDL-ALSG [8]	85.53	0.00	59.35	0.00
DLN [93]	86.40	+0.87	63.70	+4.35
SCN [70]	87.03	+1.53	-	-
PSR [67]	88.98	+3.45	63.77	+4.42
DMUE [59]	89.42	+3.89	63.11	+3.76
FDRL [55]	89.47	+3.94	-	-
TransFER [77]	90.91	+5.38	66.23	+6.88
Meta-Face2Exp [85]	88.54	+3.01	64.23	+4.88
Latent-OFER [30]	89.60	+4.07	63.90	+4.55
LA-Net [74]	[91.56]	+6.03	[67.60]	+8.25
Norface (Ours)	**92.97**	**+7.44**	**68.69**	**+9.34**

Table 5: Comparison of AU detection on cross-dataset (trained on BP4D and evaluated on BP4D+) in terms of F1 scores (%).

Method	EAC-Net [35]	JAA-Net [57]	GLEE-Net [94]	Norface (Ours)
AU1	38.0	39.7	[39.8]	**48.2**
AU2	37.5	35.6	[37.9]	**39.1**
AU4	32.6	30.7	[41.6]	**47.1**
AU6	82.0	82.4	[83.4]	**83.5**
AU7	83.4	84.7	[88.2]	**89.7**
AU10	87.1	88.8	**90.2**	[89.4]
AU12	85.1	87.0	**86.9**	[86.3]
AU14	62.1	62.2	[76.6]	**81.1**
AU15	[44.5]	38.9	**48.3**	43.2
AU17	43.6	**46.4**	42.9	[45.4]
AU23	45.0	[48.9]	47.7	**63.2**
AU24	32.8	**36.6**	29.8	[35.1]
Avg.	56.1	56.8	[57.5]	**62.6**

4.2 Analysis on Identity Normalization

We address the following Idn questions: **A1.** Is Idn effective? **A2.** How does the normalization network perform? **A3.** What is the difference between Idn and data augmentation? **A4.** Why use normalized images instead of expression features for the classification network?

A1. Impact of Idn. Tab.7 reports the results of different FEA tasks *w/* and *w/o* Idn. The results show that the performance *w/* Idn consistently outperforms that *w/o* Idn in all tasks. For example, *w/* Idn, there are improvements of 2.1%, 1.9%, and 2.5% in AU detection on BP4D, BP4D+, and DISFA, 3.73% and 1.4% increase in FER on RAF-DB and AffectNet, highlighting the benefits of Idn.

In addition, Fig.4 provides a detailed presentation of the normalization and AU detection results for different target identities. It is clear that using several target identities are validated to consistently yields better results than *w/o* Idn. Moreover, our study indicates that the selection of target identities contributes differently to performance enhancement, and certain identities are unsuitable as target faces. For example, *w/* 'ID3' resulted in a performance decrease of 1.1% compared to *w/* 'ID6', which we attribute to the *confusion* in facial expressions. In Fig.4, the non-candidate target faces in the blue box are labeled as having no AUs present, but these faces still exhibit raised eyebrows or lip corner depressors, leading to confusion in the expressions of their normalized faces. In contrast, as shown in the orange box in Fig.4, these faces have minimal confusion in facial

Table 6: Comparison of FER on cross-dataset (%). R→A (A→R) indicates trained on RAF-DB (AffectNet) and evaluated on AffectNet (RAF-DB).

Method	FDRL [55]	RANDA [24]	TransFER [77]	CA-FER [21]	Norface
R→A	41.57	[52.34]	44.57	47.2	**58.43**
A→R	-	62.48	-	-	**73.68**

Table 7: Results of Norface w and w/o Idn. AU det: AU detection.

Method	AU det(*F1*)			FER(*Acc.*)	
	BP4D	BP4D+	DISFA	RAF-DB	AffectNet
w/o Idn	67.2	64.8	70.2	89.24	67.29
$w/$ Idn	**69.3**	**66.7**	**72.7**	**92.97**	**68.69**

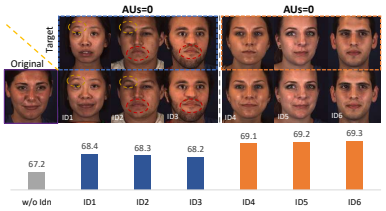


Fig. 4: Normalization results for different identities as target face (top) and AU detection results based on BP4D (bottom).

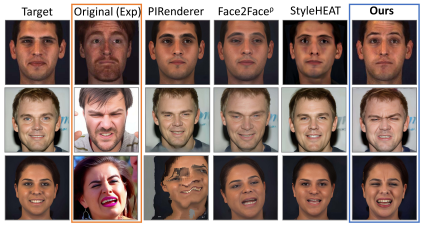


Fig. 5: Visualization of different methods for normalization.

expressions, ensuring high discriminability and performance enhancement. They can be considered as candidate target faces.

Our method is compared to multi-task methods, based on BP4D and AffectNet for fair comparison, ours exceeds JPML [103] for 23.4% and MT-VGG [29] for 8.69% under BP4D and AffectNet, respectively. Compared to identity-invariant methods, ours exceeds the SOTA of DLN [93] for 4.99% under AffectNet and exceeds the SOTA of IPD-FER [25] for 4.16% under RAF-DB. This clearly illustrates the significant benefits of Idn for AU analysis and FER tasks.

A2. Normalization performance. Fig.5 visualizes the results of our F^{nor} and other available methods enabling identity normalization. Our method offers better expression consistency between output and original faces than PIRenderer [54], Face2Face$^{\rho}$ [81], and StyleHEAT [82]. Additionally, a subjective evaluation with 100 volunteers rating 50 randomly selected normalized images on a scale of 0 to 5 shows our method (4.6) significantly outperforms PIRenderer (2.2), Face2Face$^{\rho}$ (1.8), and StyleHEAT (1.6). The conducted Paired Samples T-Test shows significant differences in expression consistency between ours (M = 4.6, SD = 0.49) and PIRenderer (M = 2.2, SD = 0.61), t(49) = 21.00, p< 0.0001; between ours and Face2Face$^{\rho}$ (M = 1.8, SD = 0.61), t(49) = 26.19, p< 0.0001; between ours and StyleHEAT (M = 1.6, SD = 0.49), t(49) = 33.20, p< 0.0001.

A3. Data augmentation. Tab.8 shows results of recent data augmentation methods such as mixup [90], cutout [14], and cutmix [84]. Ours outperforms these on both AU detection and FER tasks, e.g., outperforming mixup by 2.1%, 1.9%, 1.9%, 3.73%, and 1.4% on the five datasets, respectively. Unlike traditional data augmentation methods that merely boost the diversity of the training samples, our method directly impacts the test samples by additional normalized images.

Table 8: Comparison with data augmentation methods (%). Idn is replaced with these methods.

Method	AU det(*F1*)			FER(*Acc.*)	
	BP4D	BP4D+	DISFA	RAF-DB	AffectNet
Random init	66.4	64.0	70.2	89.15	66.26
Mixup [90]	67.2	64.8	70.8	89.24	67.29
Cutout [14]	66.7	64.2	70.3	89.28	66.49
CutMix [84]	67.4	64.5	70.4	89.57	66.74
Norface	**69.3**	**66.7**	**72.7**	**92.97**	**68.69**

Table 9: Comparison with expression features (%). Baseline means w/o E_{exp}&E_{eye} and w/o normalized images.

Method	AU det(*F1*)			FER(*Acc.*)	
	BP4D	BP4D+	DISFA	RAF-DB	AffectNet
Baseline	67.2	64.8	70.2	89.24	67.29
GLEE-Net [94]	65.7	63.7	65.5	-	-
GLEE-Net [94] ($w/$ E_{exp}&E_{eye})	65.6	63.9	65.8	-	-
$w/$ E_{exp}&E_{eye}	67.1 (↓)	64.6 (↓)	70.1 (↓)	89.77	67.06 (↓)
$w/$ Idn (Ours)	**69.3**	**66.7**	**72.7**	**92.97**	**68.69**

However, these additional normalized images aim to reduce the diversity of task-irrelevant factors, thereby enhancing facial expression analysis performance.

A4. Images v.s. Expression features for classification. As mentioned before, our classification network takes normalized images as input. These images result from the Idn that is supervised by two types of expression feature, E_{exp}&E_{eye} for constraining the expression consistency. Intuitively, there exists another possibility to directly use E_{exp}&E_{eye} for classification, instead of normalized images. (1) Tab.9 shows the results of $w/$ E_{exp}&E_{eye}. In this setup, normalized images are omitted, and the used features are $w/$ E_{exp}&E_{eye} in replace of $e_{n_{id*}}$ in the classification network. For a fair comparison, we maintain $e_{o\prime}$. These features are fed into the Input and Output MoE modules. Our method outperforms $w/$ E_{exp}&E_{eye} approach by 2.2%, 2.1%, 2.6%, 3.20%, and 1.63% on the five datasets, respectively. (2) Moreover, our method is also compared with GLEE-Net [94], which is the SOTA method designed for AU detection using features from E_{exp} and additional 3DMM. Tab.9 shows that, on average, ours also performs better than GLEE-Net by 4.6%. (3) For a fair comparison, we replace GLEE-Net's features with E_{exp}&E_{eye}, ours still leads by 4.5%. The above results show the superiority of normalized images for classification.

It is interesting to discuss the disparities between normalized images and expression features. Compared to highly abstract expression features, images encompass more structured details at pixel level. Furthermore, the used MAE can refine the facial representation from those normalized images due to its training on large-scale datasets. These factors contribute to the superior performance of normalized images over expression features.

4.3 Ablation Study

Ablation studies are conducted to validate expression consistency loss in the normalization network and each component in the classification Network.

Expression consistency loss. The expression loss $\mathcal{L}_{exp}$ and eyebrow loss $\mathcal{L}_{eye}$ in F^{nor} are used to ensure expression consistency. Tab.10 validates their effectiveness, presenting the results for $w/$ both $\mathcal{L}_{exp}$ and $\mathcal{L}_{eye}$, as well as only $w/$ $\mathcal{L}_{exp}$ in the AU detection task. The result shows that, in the eyebrow regions, using both $\mathcal{L}_{exp}$ and $\mathcal{L}_{eye}$ yields an average 0.5% higher than merely using $\mathcal{L}_{exp}$. While $\mathcal{L}_{exp}$ guides the full face, $\mathcal{L}_{eye}$ enhances the eyebrow regions. Moreover, without $\mathcal{L}_{exp}$, Idn would fail, as $\mathcal{L}_{eye}$ solely concentrates on the eyebrow regions. Fig.6 further visualizes their impacts by a few samples.

Table 10: Results of AU detection $w/$ and w/o $\mathcal{L}_{eye}$ in Eyebrow Region on BP4D in terms of F1 scores (%).

Losses	AU1	AU2	AU4	AU6	AU7	Avg.
$w/\ \mathcal{L}_{exp}$	60.2	54.9	67.1	79.8	79.8	68.3
$w/\ \mathcal{L}_{exp}\&\mathcal{L}_{eye}$	**60.7**	**55.1**	**67.7**	**79.9**	**80.4**	**68.8**

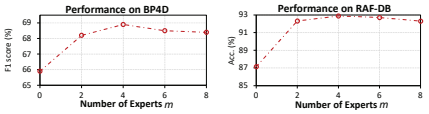


Fig. 7: Comparison of different numbers of experts m.

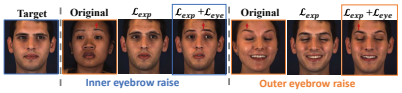


Fig. 6: Visualization of the effects of the expression loss $\mathcal{L}_{exp}$ and eyebrow loss $\mathcal{L}_{eye}$ in normalization network.



(a) AUs relevance. (b) Emotion relevance.

Fig. 8: Router decision statistics for the Output MoE module with Top-2 on both tasks. Dark colors represent high weight.

Classification Network. Fig.7 shows the results of different numbers of experts m under AU detection and FER tasks. As m increases, the performance of MoE improves, peaking at $m = 4$ with a performance boost of 3.0% and 5.74% compared to $m = 0$. In addition, Tab.11 reports the results $w/$ and w/o the Input MoE module M_i and the Output MoE module M_o, demonstrating the contribution of each module. Furthermore, Fig.8 visualizes the statistics of the Top-2 experts selected by the Output MoE module. For example, in the AU detection task, Expert 1 and Expert 4 primarily focus on detecting AU1 and AU2. These results imply that the Output MoE module effectively focuses on specific tasks, thus enhancing task-related inductive learning. Tab.12 shows the effectiveness of each loss in F^{cla}.

Table 11: Ablation study with MoE modules (%).

Strategy	$w/o\ M_i\ \&\ M_o$	$w/o\ M_i$	$w/o\ M_o$	Ours
BP4D (*F1*)	65.9	68.4	68.8	**69.3**
RAF-DB (*Acc.*)	87.15	91.92	92.57	**92.97**

Table 12: Ablation study with different losses (%).

Losses	$w/o\ \mathcal{L}_{imp}$	$w/o\ \mathcal{L}_{global}$	$w/o\ \mathcal{L}_{local}$	Ours
BP4D (*F1*)	68.5	68.7	68.9	**69.3**
RAF-DB (*Acc.*)	91.66	92.08	92.57	**92.97**

5 Conclusion

This paper provides a novel insight and a unified framework called Norface for AU detection, AU intensity estimation, and FER tasks, aiming to remove the task-irrelevant noise by identity normalization that normalizes all images to a common identity with consistent pose and background, and improves expression classification. Extensive quantitative and qualitative experiments demonstrate the superior performance of Norface for AU analysis and FER tasks as well as their cross-dataset tasks.

References

1. Ali, K., Hughes, C.E.: Facial expression recognition by using a disentangled identity-invariant expression representation. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9460–9467. IEEE (2021) 2, 3
2. An, R., Jin, A., Chen, W., Zhang, W., Zeng, H., Deng, Z., Ding, Y.: Learning facial expression-aware global-to-local representation for robust action unit detection. Applied Intelligence pp. 1–21 (2024) 2
3. Bai, Y., Wang, Z., Xiao, J., Wei, C., Wang, H., Yuille, A.L., Zhou, Y., Xie, C.: Masked autoencoders enable efficient knowledge distillers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24256–24265 (2023) 5
4. Cai, J., Meng, Z., Khan, A.S., O'Reilly, J., Li, Z., Han, S., Tong, Y.: Identity-free facial expression recognition using conditional generative adversarial network. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1344–1348. IEEE (2021) 2, 3
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018) 6
6. Chang, Y., Wang, S.: Knowledge-driven self-supervised representation learning for facial action unit recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20417–20426 (2022) 2, 10
7. Chen, K., Yang, X., Fan, C., Zhang, W., Ding, Y.: Semantic-rich facial emotional expression recognition. IEEE Transactions on Affective Computing **13**(4), 1906–1916 (2022) 9
8. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y.: Label distribution learning on auxiliary label space graphs for facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13984–13993 (2020) 4, 11
9. Chen, Y., Chen, D., Wang, Y., Wang, T., Liang, Y.: Cafgraph: Context-aware facial multi-graph representation for facial action unit recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1029–1037 (2021) 2, 10
10. Chen, Y., Joo, J.: Understanding and mitigating annotation bias in facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14980–14991 (2021) 4
11. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5177–5186 (2019) 10
12. Cui, Z., Song, T., Wang, Y., Ji, Q.: Knowledge augmented deep neural networks for joint facial expression and action unit recognition. Advances in Neural Information Processing Systems **33**, 14338–14349 (2020) 3
13. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019) 6
14. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017) 12, 13
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 5

16. Fan, Y., Lam, J., Li, V.: Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12701–12708 (2020) 11
17. Ge, X., Wan, P., Han, H., Jose, J.M., Ji, Z., Wu, Z., Liu, X.: Local global relational network for facial action units recognition. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 01–08. IEEE (2021) 10
18. He, J., Qiu, J., Zeng, A., Yang, Z., Zhai, J., Tang, J.: Fastmoe: A fast mixture-of-expert training system. arXiv preprint arXiv:2103.13262 (2021) 8
19. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 5
20. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3397–3406 (2022) 4
21. Huang, P.J., Xie, H., Huang, H.C., Shuai, H.H., Cheng, W.H.: Ca-fer: Mitigating spurious correlation with counterfactual attention in facial expression recognition. IEEE Transactions on Affective Computing (2023) 12
22. Huang, W., Zhang, S., Zhang, P., Zha, Y., Fang, Y., Zhang, Y.: Identity-aware facial expression recognition via deep metric learning based on synthesized images. IEEE Transactions on Multimedia **24**, 3327–3339 (2021) 2, 3, 4
23. Jacob, G.M., Stenger, B.: Facial action unit detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7680–7689 (2021) 4, 10
24. Ji, Y., Hu, Y., Yang, Y., Shen, H.T.: Region attention enhanced unsupervised cross-domain facial emotion recognition. IEEE Transactions on Knowledge and Data Engineering (2021) 10, 12
25. Jiang, J., Deng, W.: Disentangling identity and pose for facial expression recognition. IEEE Transactions on Affective Computing **13**(4), 1868–1878 (2022) 2, 3, 12
26. Kim, D., Song, B.C.: Contrastive adversarial learning for person independent facial emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 5948–5956 (2021) 2, 3
27. Kim, D., Song, B.C.: Optimal transport-based identity matching for identity-invariant facial expression recognition. Advances in Neural Information Processing Systems **35**, 18749–18762 (2022) 2, 3
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
29. Kollias, D., Zafeiriou, S.: Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. arXiv preprint arXiv:2103.15792 (2021) 3, 12
30. Lee, I., Lee, E., Yoo, S.B.: Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1536–1546 (2023) 2, 11
31. Li, G., Zhu, X., Zeng, Y., Wang, Q., Lin, L.: Semantic relationships guided representation learning for facial action unit recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8594–8601 (2019) 4
32. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE transactions on affective computing **13**(3), 1195–1215 (2020) 2

33. Li, S., Deng, W.: A deeper look at facial expression dataset bias. IEEE Transactions on Affective Computing **13**(2), 881–893 (2020) 2
34. Li, S., Deng, W., Du, J.: 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. pp. 2852–2861 (2017) 9
35. Li, W., Abtahi, F., Zhu, Z., Yin, L.: Eac-net: Deep nets with enhancing and cropping for facial action unit detection. IEEE transactions on pattern analysis and machine intelligence **40**(11), 2583–2596 (2018) 4, 9, 11
36. Lim, J.H., Ye, J.C.: Geometric gan. arXiv preprint arXiv:1705.02894 (2017) 6
37. Liu, H., Cai, H., Lin, Q., Zhang, X., Li, X., Xiao, H.: Feda: Fine-grained emotion difference analysis for facial expression recognition. Biomedical Signal Processing and Control **79**, 104209 (2023) 2
38. Liu, H., Cai, H., Lin, Q., Li, X., Xiao, H.: Adaptive multilayer perceptual attention network for facial expression recognition. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 6253–6266 (2022) 4
39. Liu, H., Cai, H., Lin, Q., Li, X., Xiao, H.: Learning from more: Combating uncertainty cross-multidomain for facial expression recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5889–5898 (2023) 4
40. Liu, R., Ma, B., Zhang, W., Hu, Z., Fan, C., Lv, T., Ding, Y., Cheng, X.: Towards a simultaneous and granular identity-expression control in personalized face generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2114–2123 (2024) 5
41. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015) 9
42. Luo, H., Nagano, K., Kung, H.W., Xu, Q., Wang, Z., Wei, L., Hu, L., Li, H.: Normalized avatar synthesis using stylegan and perceptual refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11662–11672 (2021) 5
43. Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., Chi, E.H.: Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1930–1939 (2018) 7, 8
44. Ma, T., Li, B., He, Q., Dong, J., Tan, T.: Gafet: Learning geometry-aware facial expression translation from in-the-wild images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7115–7125 (2023) 4
45. Mallya, A., Wang, T.C., Liu, M.Y.: Implicit warping for animation with image sets. Advances in Neural Information Processing Systems **35**, 22438–22450 (2022) 4
46. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE transactions on affective computing **10**(3), 325–347 (2017) 2
47. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing **4**(2), 151–160 (2013) 9
48. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017) 9
49. Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., Houlsby, N.: Multimodal contrastive learning with limoe: the language-image mixture of experts. Advances in Neural Information Processing Systems **35**, 9564–9576 (2022) 8

50. Nicolle, J., Bailly, K., Chetouani, M.: Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In: 2015 11th ieee international conference and workshops on automatic face and gesture recognition (FG). vol. 6, pp. 1–6. IEEE (2015) 11
51. Niu, X., Han, H., Yang, S., Huang, Y., Shan, S.: Local relationship learning with person-specific shape regularization for facial action unit detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11917–11926 (2019) 4, 10
52. Ntinou, I., Sanchez, E., Bulat, A., Valstar, M., Tzimiropoulos, G.: A transfer learning approach to heatmap regression for action unit intensity estimation. IEEE Transactions on Affective Computing **14**(1), 436–450 (2021) 11
53. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4088–4099 (2023) 5
54. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021) 4, 12
55. Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., Wang, H.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7660–7669 (2021) 11, 12
56. Sanchez, E., Tellamekala, M.K., Valstar, M., Tzimiropoulos, G.: Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9074–9084 (2021) 10, 11
57. Shao, Z., Liu, Z., Cai, J., Ma, L.: Jaa-net: joint facial action unit detection and face alignment via adaptive attention. International Journal of Computer Vision **129**, 321–340 (2021) 4, 9, 10, 11
58. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017) 8
59. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6248–6257 (2021) 2, 11
60. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86**(2), 420 (1979) 9
61. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019) 4
62. Song, T., Cui, Z., Zheng, W., Ji, Q.: Hybrid message passing with performance-driven structures for facial action unit detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6267–6276 (2021) 10
63. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019) 6

64. Sun, X., Zeng, J., Shan, S.: Emotion-aware contrastive learning for facial action unit detection. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 01–08. IEEE (2021) 4
65. Tu, C.H., Yang, C.Y., Hsu, J.Y.j.: Idennet: Identity-aware facial action unit detection. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–8. IEEE (2019) 2, 3
66. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3614–3633 (2021) 3
67. Vo, T.H., Lee, G.S., Yang, H.J., Kim, S.H.: Pyramid with super resolution for in-the-wild facial expression recognition. IEEE Access **8**, 131988–132001 (2020) 11
68. Wang, C., Wang, S., Liang, G.: Identity-and pose-robust facial expression recognition through adversarial feature learning. In: Proceedings of the 27th ACM international conference on multimedia. pp. 238–246 (2019) 2, 3
69. Wang, J., Zhao, K., Zhang, S., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13844–13853 (2023) 5
70. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6897–6906 (2020) 11
71. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10039–10049 (2021) 4
72. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. arXiv preprint arXiv:2203.09043 (2022) 4
73. Wu, S., Yan, Y., Li, Y., Cheng, Y., Zhu, W., Gao, K., Li, X., Zhai, G.: Ganhead: Towards generative animatable neural head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 437–447 (2023) 5
74. Wu, Z., Cui, J.: La-net: Landmark-aware learning for reliable facial expression recognition under label noise. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20698–20707 (2023) 2, 4, 10, 11
75. Xie, H.X., Lo, L., Shuai, H.H., Cheng, W.H.: Au-assisted graph attention convolutional network for micro-expression recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2871–2880 (2020) 4
76. Xie, S., Hu, H., Chen, Y.: Facial expression recognition with two-branch disentangled generative adversarial network. IEEE Transactions on Circuits and Systems for Video Technology **31**(6), 2359–2371 (2020) 2, 3
77. Xue, F., Wang, Q., Guo, G.: Transfer: Learning relation-aware facial expression representations with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3601–3610 (2021) 11, 12
78. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2168–2177 (2018) 2, 3
79. Yang, H., Yin, L., Zhou, Y., Gu, J.: Exploiting semantic embedding and visual feature for facial action unit detection. In: Proceedings of the IEEE/CVF Con-

ference on Computer Vision and Pattern Recognition. pp. 10482–10491 (2021) 4, 10
80. Yang, H., Zhang, Z., Yin, L.: 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). pp. 294–301. IEEE (2018) 2, 3
81. Yang, K., Chen, K., Guo, D., Zhang, S.H., Guo, Y.C., Zhang, W.: Face2face ρ: Real-time high-resolution one-shot face reenactment. In: European conference on computer vision. pp. 55–71. Springer (2022) 4, 12
82. Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: European conference on computer vision. pp. 85–101. Springer (2022) 4, 12
83. You, R., Guo, Z., Cui, L., Long, X., Bao, Y., Wen, S.: Cross-modality attention with semantic graph embedding for multi-label classification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12709–12716 (2020) 10
84. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019) 12, 13
85. Zeng, D., Lin, Z., Yan, X., Liu, Y., Wang, F., Tang, B.: Face2exp: Combating data biases for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20291–20300 (2022) 11
86. Zeng, H., Zhang, W., Chen, K., Zhang, Z., Li, L., Ding, Y.: Paste you into game: Towards expression and identity consistency face swapping. In: 2022 IEEE Conference on Games (CoG). pp. 1–8. IEEE (2022) 5
87. Zeng, H., Zhang, W., Chen, K., Zhang, Z., Li, L., Ding, Y.: Face identity and expression consistency for game character face swapping. Computer Vision and Image Understanding **236**, 103806 (2023) 5
88. Zeng, H., Zhang, W., Fan, C., Lv, T., Wang, S., Zhang, Z., Ma, B., Li, L., Ding, Y., Yu, X.: Flowface: Semantic flow-guided shape-aware face swapping. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3367–3375 (2023) 5
89. Zhang, B., Qi, C., Zhang, P., Zhang, B., Wu, H., Chen, D., Chen, Q., Wang, Y., Wen, F.: Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22096–22105 (2023) 5
90. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 12, 13
91. Zhang, J., Zeng, X., Wang, M., Pan, Y., Liu, L., Liu, Y., Ding, Y., Fan, C.: Freenet: Multi-identity face reenactment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5326–5335 (2020) 4
92. Zhang, W., Guo, Z., Chen, K., Li, L., Zhang, Z., Ding, Y.: Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. arXiv preprint arXiv:2107.03708 (2021) 3
93. Zhang, W., Ji, X., Chen, K., Ding, Y., Fan, C.: Learning a facial expression embedding disentangled from identity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6759–6768 (2021) 2, 3, 6, 11, 12

94. Zhang, W., Li, L., Ding, Y., Chen, W., Deng, Z., Yu, X.: Detecting facial action units from global-local fine-grained expressions. IEEE Transactions on Circuits and Systems for Video Technology (2023) 3, 10, 11, 13
95. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014) 9
96. Zhang, Y., Dong, W., Hu, B.G., Ji, Q.: Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2314–2323 (2018) 11
97. Zhang, Y., Zhao, R., Dong, W., Hu, B.G., Ji, Q.: Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7034–7043 (2018) 11
98. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering **34**(12), 5586–5609 (2021) 3
99. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3438–3446 (2016) 9
100. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021) 7
101. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3657–3666 (2022) 4
102. Zhao, K., Chu, W.S., Martinez, A.M.: Learning facial action units from web images with scalable weakly supervised clustering. In: Proceedings of the IEEE Conference on computer vision and pattern recognition. pp. 2090–2099 (2018) 4
103. Zhao, K., Chu, W.S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2207–2216 (2015) 3, 4, 12
104. Zhao, K., Chu, W.S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3391–3399 (2016) 4