Revisiting Representation Learning and Identity Adversarial Training for Facial Behavior Understanding

Anonymous FG2025 submission
Paper ID ****

***Abstract*— Facial Action Unit (AU) detection has gained significant attention as it enables the breakdown of complex facial expressions into individual muscle movements. In this paper, we revisit two fundamental factors in AU detection: diverse and large-scale data and subject identity regularization. Motivated by recent advances in foundation models, we highlight the importance of data and introduce Face9M, a diverse dataset comprising 9 million facial images from multiple public sources. Pretraining a masked autoencoder on Face9M yields strong performance in AU detection and facial expression tasks. More importantly, we emphasize that the Identity Adversarial Training (IAT) has not been well explored in AU tasks. To fill this gap, we first show that subject identity in AU datasets creates shortcut learning for the model and leads to sub-optimal solutions to AU predictions. Secondly, we demonstrate that strong IAT regularization is necessary to learn identity-invariant features. Finally, we elucidate the design space of IAT and empirically show that IAT circumvents the identity-based shortcut learning and results in a better solution. Our proposed methods, Facial Masked Autoencoder (FMAE) and IAT, are simple, generic and effective. Remarkably, the proposed FMAE-IAT approach achieves new state-of-the-art F1 scores on BP4D (67.1%), BP4D+ (66.8%), and DISFA (70.1%) databases, significantly outperforming previous work. We release the code and model.**

I. INTRODUCTION

The Facial Action Coding System (FACS) was developed to objectively encode facial behavior through specific movements of facial muscles, named Action Units (AU) [19]. Compared with facial expression recognition (FER) [37], [84], [35] and valence and arousal estimation [85], [51], [54], detecting action units offers a more nuanced and detailed understanding of human facial behavior capturing multiple individual facial actions simultaneously.

This problem attracted considerable interest within the deep learning community [13], [32], [72], [56], [61], [73]. Many works used a facial region prior [39], [56], [13], introduced extra modalities [80], [69], [81], or incorporated the inherent AU relationships [38], [46], [75] to solve the AU detection task and achieved significant advancements. Diverging from these approaches, which often necessitate complex model designs or depend heavily on prior AU knowledge, in this paper, we revisit two fundamental factors that significantly contribute to the AU detection task: diverse and large-scale data and subject identity regularization.

Recently, data has become pivotal in training foundation models [57], [2], [42], [64] and large language models [8], [1], [65], [15]. Following this trend, we introduce Face9M, a large-scale and diverse facial dataset curated and refined from publicly available datasets for pretraining. Different from contrastive learning methods [9], [13], [22], we propose to do facial representation learning using Masked Autoencoders (MAE) [27]. The underlying motivation is that most facial tasks require a fine-grained understanding of the face, and masked pretraining results in lower-level semantics than contrastive learning according to [5]. Our large-scale facial representation learning approach demonstrates excellent generalization and scalability in downstream tasks. Notably, our proposed Facial Masked Autoencoder (FMAE), pretrained on Face9M, sets new state-of-the-art benchmarks in both AU detection and FER tasks.

Similar to the importance of data, domain knowledge and task-prior knowledge can be incorporated into the model in the form of regularization [26], [31], [52], [55] to improve task performance. Our key observation is that popular AU detection benchmarks (i.e. BP4D [83], BP4D+ [86], DISFA [50]) include at most 140 human subjects and 192,000 images, meaning that each subject has hundreds of annotated images. This abundance can lead models to prefer simple, easily recognizable patterns over more complex but generalizable ones, as suggested by the **shortcut learning theory** [23], [29]. Therefore, we hypothesize that **AU detection models tend to learn the subject identity features to infer the AUs, resulting in learning a trivial solution that does not generalize well**. To verify our hypothesis, we employed the linear probing technique — adding a learnable linear layer to a trained AU model while freezing the network backbone —to measure identity recognition accuracy. The high accuracy (83%) we obtained in predicting the identities of the subjects clearly shows that the models effectively 'memorize' subject identities. To counteract the learning of identity-based features, we propose in this paper Identity Adversarial Training (IAT) for AU detection task by adding a linear identity prediction head and unlearning the identity feature using gradient reverse [20]. Further analysis shows that IAT significantly reduces the identity accuracy of linear probing and leads to better learning dynamics that avoid convergence to trivial solutions. This method further improves our AU models beyond the advantages brought by pretraining with a large-scale dataset.

Although Zhang et al. [87] first introduced identity-based adversarial training to AU detection tasks, the identity learning issue and its negative effect (identity shortcut learning) have not been explored. Also, the design space of IAT lacks illustration in [87]. We revisit the identity adversarial training method in depth to answer these unexplored questions. In contrast to the weak identity regularization used in [87],

This is the uncorrected author proof. Please do not distribute. Copyright with IEEE. Cite as Ning, M., A.A. Salah, I. Onal Ertugrul, "Revisiting Representation Learning and Identity Adversarial Training for Facial Behavior Understanding," Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, Clearwater, 2025.

we demonstrate that AU detection requires a strong identity regularization. To this end, the linear identity head and a large gradient reverse scaler are necessities for the AU detection task. Our proposed FMAE with IAT sets a new record of F1 score on BP4D (67.1%), BP4D+ (66.8%) and DISFA (70.1%) datasets, substantially surpassing previous work.

Overall, the main contributions of this paper are:

- We demonstrate the effectiveness of using a diverse dataset for facial representation learning.
- We highlight the identity shortcut learning issue and propose the use of a linear identity head and a large gradient reversal scalar in IAT to mitigate this issue for AU detection.
- We release the code and checkpoint of FMAE with various model sizes (small, base, large), aiming at facilitating all facial tasks.

II. Related Work

A. Action Unit Detection

Recent works have proposed several deep learning-based approaches for facial action unit (AU) detection. Some of them have divided the face into multiple regions or patches [88], [39], [56] to learn AU-specific representations and some have explicitly modeled the relationships among AUs [38], [46], [75]. The most recent approaches have focused on detecting AUs using vision transformers on RGB images [32] and on multimodal data including RGB, depth images, and thermal images [81]. Yin et al. [77] have used generative models to extract representations and a pyramid CNN interpreter to detect AUs. Yang et al. [74] jointly modeled AU-centered features, AU co-occurrences, and AU dynamics. Contrastive learning has recently been adopted for AU detection [41], [60]. Particularly, Chang et al. [13] have adopted contrastive learning among AU-related regions and performed predictive training considering the relationships among AUs. Zhang et al. [80] have proposed a weakly-supervised text-driven contrastive approach using coarse-grained activity information to enhance feature representations. In addition to fully supervised approaches, Tang et al. [63] have implemented a semi-supervised approach with discrete feedback. However, none of these approaches have made use of large-scale self-supervised pretraining.

B. Facial Representation Learning

Facial representation learning [9], [10], [89] has seen substantial progress with the advent of self-supervised learning [14], [28], [7], [27], [12]. For example, Mask Contrastive Face [68] combines mask image modeling with contrastive learning to do self-distillation, thereby enhancing facial representation quality. Similarly, ViC-MAE [30] integrates MAE with temporal contrastive learning to enhance video and image representations. MAE-face [47] uses MAE for facial pertaining by 2 million facial images. Additionally, ContraWarping [71] employs global transformations and local warping to generate positive and negative samples for facial representation learning. To learn good local facial representations, Gao et al. [22] explicitly enforce the consistency of facial regions by matching the local facial representations across views. Different from the above-mentioned work that mainly focuses on models, we emphasize the importance of data (diversity and quantity). Our collected datasets contain 9 million images from various public resources.

C. Adversarial Training and Gradient Reverse

Adversarial training [25] is a regularization technique in deep learning to enhance the model's robustness specifically against input perturbations that could lead to incorrect outputs. Although gradient reverse technique [20] aims to minimize domain discrepancy for better generalization across different data distributions, these two techniques share the same spirit of the 'Min-Max' training paradigm and are used to improve the model robustness [36], [48], [21], [66]. Gradient reverse has also been used for the regularization of fairness [58] or for meta-learning [4].

The most relevant research to our paper is [87], where the authors introduce identity-based adversarial training for the AU detection task. However, they did not thoroughly investigate the identity learning phenomenon and its detrimental impacts. Moreover, their empirical settings, the small gradient reverse scaler and the 2-layer MLP identity head, have been [87] verified as an inferior solution to AU detection. By contrast, we conduct a comprehensive examination for IAT to address these unexplored questions.

III. Methods

A. Large-scale Facial MAE Pretraining

While the machine learning community has long established the importance of having rich and diverse data for training, recent successes in foundation models and large language models illustrated the full potential of pretaining [57], [42], [8], [1], [65]. In line with this, our research pivots towards a nuanced exploration of data diversity and quantification in the context of facial representation learning. Unlike natural image datasets like ImageNet-1k, face datasets have low variance. Also, we observe that different facial datasets have domain shifts regarding the facial area, perspective and background. To increase the data diversity, we propose to collect a large facial dataset for pertaining from multiple data sources.

We first collect facial images from CelebA [44], FFHQ [33], VGGFace2 [11], CASIA-WebFace [76], MegaFace [34], EDFace-Celeb-1M [79], UMDFaces [6] and LAION-Face [89] datasets, because these datasets contain a massive number of identities collected in diverse scenarios. For instance, the facial images in UMDFaces also capture the upper body with various image sizes, while some datasets (FFHQ, CASIA-WebFace) mainly feature the center face. We then discard images whose width-height-ratio or height-width-ratio is larger than 1.5. Finally, the remaining images are resized to 224*224. The whole process yields 9 million facial images (termed Face9M) which will be used for self-supervised facial pertaining.

Regarding representation learning methods, we apply Masked Image Modeling (MIM) [27] as it tends to learn

more fine-grained features than contrastive learning according to the study in [5], which benefits facial behavior understanding. Specifically, we utilize Face9M to train a masked autoencoder (MAE) by the mean squared error between the reconstructed and original images in the pixel space. The resulting model is termed FMAE, and the decoder of FMAE is discarded for the downstream tasks.

B. Identity Adversarial Training

One of our key findings in this paper is that, the limited number of subjects in AU datasets makes identity recognition a trivial task and provides a shortcut learning path, resulting in a AU model that contains identity-related features and does not generalize well (see Section V). Motivated by the gradient reverse in domain adaption [20], we propose to apply gradient reverse on AU detection to learn identity-invariant features, aiming at better model generalization.

Our model architecture is presented in Figure 1, where the backbone is a vision transformer and parameterized by θ_f, the AU head predicts the AUs and the ID head outputs the subject identities, respectively. The input image $\boldsymbol{x}$ is first mapped by the backbone $G_f(\cdot;\theta_f)$ to a D-dimensional feature vector $\boldsymbol{f} \in \mathbb{R}^D$, then the feature vector $\boldsymbol{f}$ is fed into the AU head $G_f(\cdot;\theta_{au})$ and the ID head $G_f(\cdot;\theta_{id})$. simultaneously. Assume that we have data samples $(\boldsymbol{x}, y, d) \sim D_s$, parameters θ_{au} of the AU head are optimized to minimize the AU loss L_{au} given AU label y, and parameters θ_{id} of the ID head are trained to minimize the identity loss L_{id} given the identity label d.

To make the feature vector $\boldsymbol{f}$ invariant to subject identity, we seek the parameters θ_f of the backbone that **maximize** the identity loss L_{id} (Equation 3), so that the backbone excludes the identity-based features. In the meantime, the backbone $G_f(\cdot;\theta_f)$ is expected to minimize the AU loss L_{au}. Formally, we consider the following functional loss:

$$L_{au} = \mathbb{E}_{(\boldsymbol{x},y)\sim D_s}[CE(G_y(G_f(\boldsymbol{x};\theta_f);\theta_{au}),\ y)] \quad (1)$$

$$L_{id} = \mathbb{E}_{(\boldsymbol{x},d)\sim D_s}[CE(G_d(G_f(\boldsymbol{x};\theta_f);\theta_{id}),\ d)] \quad (2)$$

where CE denotes the cross entropy loss function. We seek the parameters θ_f^*, θ_{au}^*, θ_{id}^* that deliver a solution:

$$(\theta_f^*, \theta_{au}^*) = arg \min_{\theta_f,\theta_{au}} L_{au}(D_s;\theta_f,\theta_{au}) - \lambda L_{id}(D_s;\theta_f,\theta_{id}^*) \quad (3)$$

$$\theta_{id}^* = arg \min_{\theta_{id}} L_{id}(D_s;\theta_f^*,\theta_{id}) \quad (4)$$

where the parameter λ controls the trade-off between the two objectives that shape the feature $\boldsymbol{f}$ during learning. Comparing the identity loss L_{id} in Equation 3 and Equation 4, θ_f is optimized to maximize to increase L_{id} while θ_{id} is learned to reduce L_{id}. To achieve these two opposite optimizations through regular gradient descent and backpropagation, the gradient reverse layer is designed to reverse the identity partial derivative $\frac{\partial L_{id}}{\partial \theta_{id}}$ before it is propagated to the backbone. The resultant derivative $-\lambda\frac{\partial L_{id}}{\partial \theta_f}$, together with $\frac{\partial L_{au}}{\partial \theta_f}$, are used to update the backbone parameter θ_f.

Intuitively, the backbone is still optimized to learn the AU-related features, but under the force of reducing the identity-related features. The 'Min-Max' training paradigm in gradient reverse (see Equation 3) resembles the adversarial training [49] and Generative Adversarial Networks (GANs) [24], so we name our method 'Identity Adversarial Training' for the AU detection task.

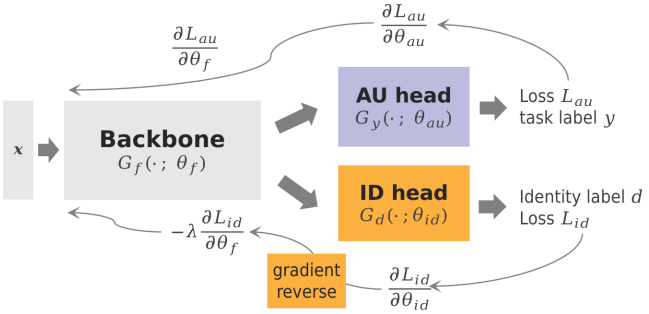


Fig. 1: Architecture of Identity Adversarial Training. The AU head and ID head both are a linear classifier predicting the AUs and identity, respectively. The backbone $G_f(\cdot;\theta_f)$ is the encoder of the pretrained FMAE. During training, the AU head is optimized by $\frac{\partial L_{au}}{\partial \theta_f}$ and the ID head is optimized by $\frac{\partial L_{id}}{\partial \theta_f}$. The gradient reverse layer multiplies the gradient by a negative value $-\lambda$ to unlearn the features capable of recognizing identities. Finally, the parameters of the backbone are optimized by the two forces: $-\lambda\frac{\partial L_{id}}{\partial \theta_f}$ and $\frac{\partial L_{au}}{\partial \theta_f}$.

Importantly, we reveal the key design of identity adversarial training for AU detection: **a strong adversarial regularization (large magnitude of $-\lambda\frac{\partial L_{id}}{\partial \theta_f}$) is required to learn identity-invariant features for the backbone**. Specifically, we propose to use a large λ and a linear projection layer for the ID head. The former scales up the $\frac{\partial L_{id}}{\partial \theta_f}$ and the latter ensures a large L_{id}, leading to a large $||-\lambda\frac{\partial L_{id}}{\partial \theta_f}||$ during training. In Section V-C, we will show that the small λ and 2-layer MLP ID head used by [87] would lead to a weak identity regularization (small magnitude of $-\lambda\frac{\partial L_{id}}{\partial \theta_f}$) and inferior AU performance. We defer more details and analysis to Section V-C

IV. Experiments

We test the performance of FMAE and FMAE-IAT on AU benchmarks, using the F1 score. To illustrate the representation learning efficacy of FMAE, we also report its facial expression recognition (FER) accuracy on RAF-DB [59] and AffectNet [53] databases, and compare FMAE with previous face models pretrained based on contrastive learning.

A. Datasets

BP4D [83] is a manually annotated database of spontaneous behavior containing videos of 41 subjects. There are 8 activities designed to elicit various spontaneous emotional responses, resulting in 328 video clips. A total of 140,000

TABLE I: F1 scores (in %) achieved for 12 AUs on BP4D dataset. The best and the second-best results of each column are indicated with bold font and underline, respectively.

Methods	Venue	AU												Avg.
		1	2	4	6	7	10	12	14	15	17	23	24	
HMP-PS [62]	CVPR'21	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
SEV-Net [73]	CVPR'21	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
FAUT [32]	CVPR'21	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
PIAP [63]	ICCV'21	55.0	50.3	51.2	80.0	79.7	84.7	**90.1**	65.6	51.4	63.8	50.5	50.9	64.4
KSRL [13]	CVPR'22	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
ANFL [46]	IJCAI'22	52.7	44.3	60.9	79.9	80.1	85.3	89.2	69.4	55.4	64.4	49.8	55.1	65.5
CLEF [80]	ICCV'23	55.8	46.8	**63.3**	79.5	77.6	83.6	87.8	67.3	55.2	63.5	53.0	57.8	65.9
MCM [81]	WACV'24	54.4	48.5	60.6	79.1	77.0	84.0	89.1	61.7	**59.3**	64.7	53.0	**60.5**	66.0
AUFormer [78]	ECCV'24	-	-	-	-	-	-	-	-	-	-	-	-	66.2
MDHR [69]	CVPR'24	58.3	50.9	58.9	78.4	80.3	84.9	88.2	**69.5**	56.0	**65.5**	49.5	59.3	66.6
FMAE	(ours)	59.2	50.0	62.7	80.0	79.2	84.7	89.8	63.5	52.8	65.1	**55.3**	56.9	66.6
FMAE-IAT	(ours)	**62.7**	**51.9**	62.7	79.8	80.1	84.8	89.9	64.6	54.9	65.4	53.1	54.7	**67.1**

frames are annotated by expert FACS annotators. Following [39], [80], [69], we split all annotated frames into three subject-exclusive folds for 12 AUs.

BP4D+ [86] is an extended dataset of BP4D and features 140 participants. For each subject, 20 seconds from 4 activities are manually annotated by FACS annotators, resulting in 192,000 labelled frames. We divide the subjects into four folds as per guidelines in [82], [80] and 12 AUs are used for AU detection.

DISFA [50] contains left-view and right-view videos of 27 subjects. Similar to [73], [80], we use 8 of 12 AUs. We treat samples with AU intensities higher or equal to 2 as positive samples. The database contains 130,000 manually annotated images. Following [80] we perform subject-exclusive 3-fold cross-validation.

RAF-DB [59] contains 15,000 facial images with annotations for 7 basic expressions namely neutral, happiness, surprise, sadness, anger, disgust, and fear. Following the previous work [61], [80], we use 12,271 images for training and the remaining 3,068 for testing.

AffectNet [53] is currently the largest FER dataset with annotations for 8 expressions (neutral, happy, angry, sad, fear, surprise, disgust, contempt). AffectNet-8 includes all expression images with 287,568 training samples and 4,000 testing samples. In practice, we only use 37,553 images (from Kaggle) for training as training on the whole training set is expensive.

B. Implementation details

Regarding facial representation learning, we pretrain FMAE with Face9M for 50 epochs (including two warmup epochs) using four NVIDIA A100 GPUs. The remaining parameter settings follow [27] without any changes. After the pertaining, we finetune FMAE for FER tasks with cross-entropy loss, and fine-tune FMAE and FMAE-IAT for AU detection with binary cross-entropy loss. In most cases, we finetune the model for 30 epochs with a batch size of 64 and a base learning rate of 0.0005. Following MAE [28], we use a weight decay of 0.05, AutoAugmentation [16] and Random Erasing 0.25 [90] for regularization. By default, we apply ViT-large for FMAE and FMAE-IAT throughout this paper, if not specified otherwise. The complete code, hyperparameters and training/testing protocols are posted on our GitHub repository for reproducibility.

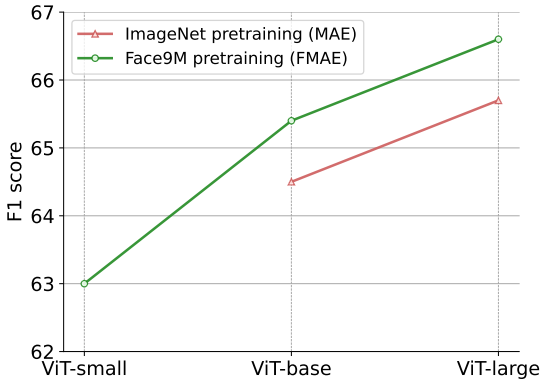


Fig. 2: F1 results of FMAE using different model sizes on 12 AUs of the BP4D. Models pretrained on Face9M are better than the ones pretrained on ImageNet-1k. MAE paper does not train ViT-small on ImageNet-1k, thus this entry is missing.

C. Result of FMAE

We first show the F1 score of FMAE on the BP4D dataset in Table I. FMAE achieves the same average F1 (66.6%) with the state-of-the-art method MDHR [69] which utilizes a two-stage model to learn the hierarchical AU relationships. Here, we see the effectiveness of data-centric facial representation learning, and demonstrate that a simple vision transformer [18], which is the architecture of FMAE, is capable of learning complex AU relationships. FMAE surpasses all previous work on BP4D+ and DISFA by achieving 66.2% and 68.7% F1 scores, respectively (see Table II and Table III).

TABLE II: F1 scores (in %) achieved for 12 AUs on BP4D+ dataset. The best and the second-best results of each column are indicated with bold font and underline, respectively. MFT* uses extra depth modality.

Methods	Venue	AU												**Avg.**
		1	2	4	6	7	10	12	14	15	17	23	24	
ViT [18]	ICLR'21	45.6	38.2	35.5	85.9	88.3	90.3	89.0	81.9	45.8	48.8	57.2	34.6	61.6
CLIP [57]	ICML'21	49.4	39.7	38.9	85.7	87.6	90.6	89.0	80.6	44.9	50.3	56.1	32.8	62.1
SEV-Net [73]	CVPR'21	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
MFT [82]	FG'21	48.4	37.1	34.4	85.6	88.6	90.7	88.8	81.0	47.6	51.5	55.6	36.9	62.2
MFT* [82]	FG'21	49.6	42.0	43.5	85.8	88.6	90.6	89.7	80.8	49.8	52.2	59.1	38.4	64.2
CLEF [80]	ICCV'23	47.5	39.6	40.2	86.5	87.3	90.5	**89.9**	81.6	47.0	46.6	54.3	41.5	63.1
GLTE-Net [3]	Intelli'24	51.5	46.6	43.5	86.8	**89.6**	91.0	89.8	82.3	46.8	49.3	60.9	**50.9**	65.7
FMAE	(ours)	53.9	45.5	45.9	86.2	88.3	**91.2**	**89.9**	82.3	**51.3**	**56.3**	60.7	42.7	66.2
FMAE-IAT	(ours)	**54.2**	**47.0**	**53.9**	85.7	88.4	**91.2**	89.7	82.4	50.3	54.4	**61.0**	43.4	**66.8**

TABLE III: F1 scores (in %) achieved for 8 AUs on DISFA dataset. The best and the second-best results of each column are indicated with bold font and underline, respectively.

Methods	Venue	AU								**Avg.**
		1	2	4	6	9	12	25	26	
FAUT [32]	CVPR'21	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	61.5
PIAP [63]	ICCV'21	50.2	51.8	71.9	50.6	54.5	79.7	94.1	57.2	63.8
ANFL [46]	IJCAI'22	54.6	47.1	72.9	54.0	55.7	76.7	91.1	53.0	63.1
KSRL [13]	CVPR'22	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5
KS [40]	ICCV'23	53.8	59.9	69.2	54.2	50.8	75.8	92.2	46.8	62.8
CLEF [80]	ICCV'23	64.3	61.8	68.4	49.0	55.2	72.9	89.9	57.0	64.8
SACL [43]	TAC'23	62.0	**65.7**	74.5	53.2	43.1	76.9	**95.6**	53.1	65.5
MDHR [69]	CVPR'24	**65.4**	60.2	75.2	50.2	52.4	74.3	93.7	58.2	66.2
AUFormer [78]	ECCV'24	-	-	-	-	-	-	-	-	66.4
GPT-4V [45]	CVPRW'24	52.6	56.4	**82.9**	**64.3**	55.3	75.4	91.2	66.4	67.3
FMAE	(ours)	62.7	59.5	67.3	55.6	**61.8**	77.9	95.0	69.8	68.7
FMAE-IAT	(ours)	64.7	61.3	70.8	58.1	59.4	**79.9**	95.2	**71.3**	**70.1**

To further verify the importance of the Face9M dataset, we compare FMAE pretrained on Face9M with FMAE pretrained on ImageNet-1k [17], using BP4D as the test set. Figure 2 shows that FMAE pretrained on Face9M always outperforms the one pretrained on ImageNet-1k given the same model size (ViT-base or ViT-large). Also, we empirically demonstrate that FMAE benefits from the scaling effect of model size on AU detection tasks (see the green line in Figure 2).

TABLE IV: Results of accuracy on FER benchmarks. FMAE surpasses all previous contrastive-related work.

Model	Contrastive	MIM	AffectNet-8	RAF-DB
MCF [68]	√	√	60.98	86.86
FaRL [89]	√	√	-	88.31
CLEF [80]	√		62.77	90.09
FRA [22]	√		-	90.76
LA-Net [70]	√		64.54	91.78
FMAE (ours)		√	**65.00**	**93.09**

In addition to AU detection, we benchmark FMAE on the downstream facial task of FER to verify the effectiveness of masked image representation learning. We present the results of FMAE on AffectNet-8 and RAF-DB in Table IV and compare FMAE with other contrastive learning-based models. FMAE sets a new state-of-art accuracy on both datasets (65% on AffectNet-8 and 93.09% on RAF-DB). Note that, we did not test FMAE-IAT on FER tasks because these datasets do not include the identity labels and do not suffer from identity shortcut learning due to the large number of subjects.

D. Results of FMAE-IAT

Although FMAE has already achieved superior results on AU benchmarks, we highlight that the Identity Adversarial Training could further boost the performance of FMAE across all AU datasets. Specifically, we compare FMAE-IAT with the most recent state of the art methods on BP4D, BP4D+ and DISFA datasets. Table I suggests that FMAE-IAT shows superior performance by achieving an average F1 Score of 67.1% and FMAE-IAT ranks as the best or second-best performer in several individual AUs, notably AU 1, 2, 4, 12, 17 and 23. Similarly, FMAE-IAT also stands out on BP4D+ dataset with the highest average F1 score of 66.8% shown in Table II. Our results on the DISFA benchmark given in Table III are even more distinguishing, FMAE-IAT gains the best or the second-best performance on 6 out of 8 AUs, pushing the average F1 score beyond the 70% mark.

For the gradient reverse layer, we use $\lambda = 2.0$ for BP4D, $\lambda = 1.0$ for BP4D+ and $\lambda = 0.5$ for DISFA. Differing from the setting $\lambda \in [0.008, 0.08]$ used in [87], we emphasize that a strong IAT regularization is necessary for AU tasks and we defer the in-depth discussion throughout Section V.

V. Analysis of Identity Adversarial Training

In this section, we elucidate IAT by first showing the identity learning issue in AU tasks (Section V-A), then demonstrating the learning dynamics refined by IAT (Section V-B), and finally illustrating the importance of ensuring a strong regularization of IAT (Section V-C).

A. Linear probing for identity recognition

Motivated by the shortcut learning theory [23], [29], we hypothesize that each subject of AU datasets is exposed to the neural network hundreds of times in a single training epoch, which provides an identity shortcut for the model to learn the subject identity. This identity learning issue is undesired, as the model is supposed to generalize to unseen subjects.

To demonstrate identity learning in AU detection, we quantitatively and qualitatively evaluate the identity features via linear probing [14] and t-SNE [67] technique, respectively. In detail, we apply linear probing on a trained AU detection model (FMAE) and evaluate the identity recognition accuracy on the BP4D dataset, which contains 41 subjects with the identity labels. Specifically, we freeze the backbone $G_f(\cdot;\theta_f)$ of a well-trained FMAE and add a learnable linear classifier on top of the backbone to predict the identity label. For each subject in BP4D, we randomly draw 70 samples for training and 30 samples for testing. The resultant accuracy under linear probing is shown in Figure 3, the red line indicates that FMAE can recognize more than half of people among the 41 subjects even though the model is only trained for one epoch. Given enough training, the identity recognition accuracy can be as high as 83%. By contrast, IAT significantly alleviates this identity learning issue with 4.6% accuracy after one epoch of training and 27.9% accuracy at epoch 19. An interesting phenomenon is that even under the strong identity unlearning regularization, FMAE-IAT seems still to partially learn the identity-based features, by showing 27.9% accuracy (higher than the random guess accuracy 2.4%). We believe that the inherent high correlation between training and testing images for each subject provides the possibility for the model to infer the identity by looking at the non-face area.

We also visualize the feature output from the backbone of FMAE and FMAE-IAT using t-SNE and see how these features are clustered according to the identity label. Figure 4 presents the t-SNE results for 20 subjects in BP4D (41 subjects in total), given trained FMAE and FMAE-IAT models. It is clear that the identity-based feature clusters in FMAE become less linearly distinguishable (the ID head is a linear layer) under the effect of IAT.

B. IAT mitigates identity shortcut learning

After showing that a regular AU model (FMAE) learns the subject identity, we now illustrate that the identity shortcut learning leads to a trivial AU prediction solution that is inferior to the solution delivered by IAT. Concretely, we observe that FMAE and FMAE-IAT have totally different learning dynamics in terms of AU predictions (indicated by F1 score). Figure 5 shows the F1 score of both models along the training epochs, where the two models share the same learning rate, batch size and initial training states. It is clear from Figure 5 that FMAE is optimized quickly and converges at the third epoch with an F1 score of 65.45% under the identity shortcut. In contrast, FMAE-IAT learns the AU decision boundary progressively and converges only at epoch 15 with an F1 score of 66.66%. One can infer that IAT explicitly pushes the backbone $G_f(\cdot;\theta_f)$ away from the identity-related solution region and delivers a better solution for AU detection tasks.

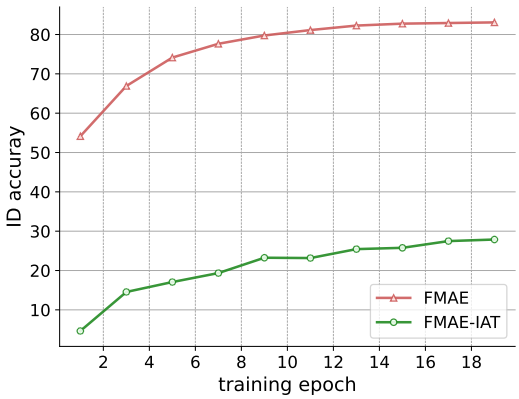
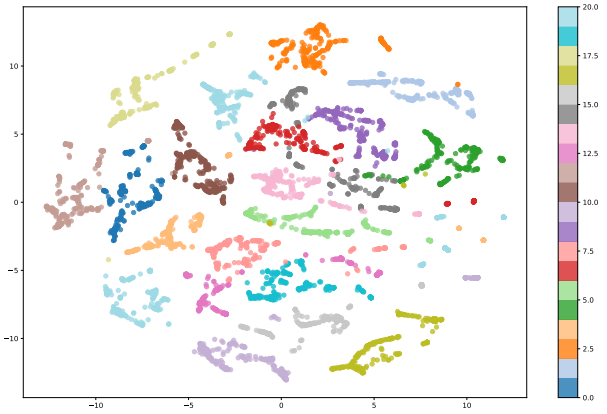


Fig. 3: Identity recognition accuracy (%) evaluated by linear probing on the BP4D dataset. IAT greatly reduces the identity-related features learned by the network backbone $G_f(\cdot;\theta_f)$.

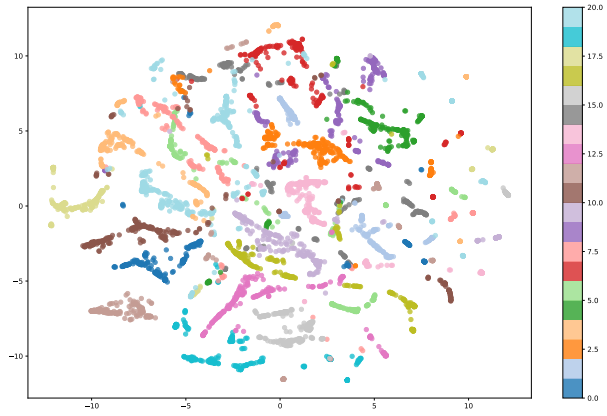
C. Large $||-\lambda\frac{\partial L_{id}}{\partial \theta_f}||$ is necessary

In Section III-B, we have mentioned the key design space of IAT: a linear projection layer for the ID head and a large λ for the gradient reverse layer. These two factors together ensure the large magnitude of $-\lambda\frac{\partial L_{id}}{\partial \theta_f}$ during the adversarial training of the backbone $G_f(\cdot;\theta_f)$. We elaborate here on the specifics of the IAT design space. We postulate that learning the subject identity is relatively easy, since there are many facial components and non-facial cues that can be used for identity recognition. Therefore, a strong regularization of IAT (i.e., a large $||-\lambda\frac{\partial L_{id}}{\partial \theta_f}||$) is required to counteract the identity-related learning tendency.

We first show the effect of using different λ on the fold-2 of the BP4D dataset. All models share the same training settings except for λ. In Table V, 'Epoch' indicates the training convergence point in terms of the F1 score and $\lambda = 0$ represents the group without IAT. We see that a small λ ($\lambda = 0.02$), such as the one used in [87], has little gain of F1 score, whereas the large λ ($\lambda = 1, 2, 3$) yields significant improvement of AU prediction. Moreover, the larger λ we



(a) FMAE



(b) FMAE-IAT

Fig. 4: t-SNE visualization of the backbone features on BP4D dataset regarding the identity labels, each color stands for a subject. Only 20 subjects are visualized for readability even though BP4D contains 41 subjects. FMAE features are more identity-clustered than FMAE-IAT features

TABLE V: The effect of different λ on BP4D. F1 is reported on fold-2 of BP4D and Epoch means the convergence epoch during training.

λ	0	0.02 (used in [87])	1	2	3
F1	68.33	68.60	69.26	**69.57**	69.47
Epoch	2	10	20	21	27

use, the more training epochs are required to reach a better optimization point, which is consistent with the phenomenon in Figure 5. Additionally, we perform the ablation study on λ using the AU datasets to demonstrate that λ is an easy hyper-parameter to tune in practice. Table VI shows that λ values within the set of [0.5, 1, 2], which are used across all AU datasets in this paper, consistently result in an improvement in the F1 score.

Furthermore, we show that recognizing identity is a trivial task since we find that a non-linear ID head $G_f(\cdot;\theta_{id})$ can still recognize the subjects given the identity-invariant features (regularized by IAT). To investigate this in more

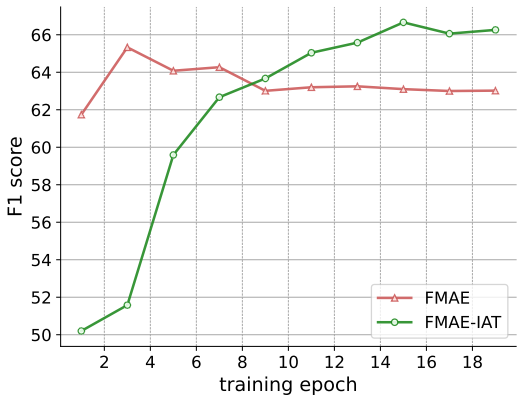


Fig. 5: F1 dynamics of FMAE and FMAE-IAT on BP4D+ during training. Fold-2 of BP4D+ is used for visualization.

TABLE VI: Ablation study of λ on BP4D, BP4D+ and DISFA. F1 scores are reported for one fold of each dataset, with the numbers in parentheses indicating the absolute improvement in F1 compared to the baseline $\lambda = 0$.

Dataset	λ			
	0	0.5	1	2
BP4D	68.81	69.22 (+0.41)	69.26 (+0.45)	**69.57 (+0.76)**
BP4D+	65.45	66.58 (+1.13)	**66.66 (+1.21)**	66.62 (+1.17)
DISFA	71.06	**73.48 (+2.42)**	73.23 (+2.17)	73.01 (+1.95)

detail, we increase the model capacity of the ID head $G_f(\cdot;\theta_{id})$ given the backbone trained with a large λ, and measure the identity loss. Table VII shows the results of using different MLP layers for FMAE-IAT under the same regularization strength ($\lambda = 2$). The ID loss in Table VII suggests that the model gradually learns the identity given some model capacity. By contrast, using the 1-layer MLP (linear projection layer) for the ID head leads to a large ID loss L_{id}, thus ensuring the large magnitude of $-\lambda\frac{\partial L_{id}}{\partial \theta_f}$. Therefore the linear projection layer is another necessity for IAT in AU detection. The convergence epoch and F1 score in Table VII also imply that the 2-layer MLP and 3-layer MLP both converge fast and learn a sub-optimal solution to the AU tasks, which is consistent with our previous observations.

TABLE VII: The effect of different MLPs for the ID head. Epoch in the table shows the convergence epoch during training and ID loss indicates the average identity loss at the convergence epoch using the training set. A higher ID loss implies a lower ID accuracy.

ID head	1-layer MLP	2-layers MLP (used in [87])	3-layers MLP
F1	**69.57**	69.00	68.90
ID loss	0.152	0.096	0.085
Epoch	21	7	6

VI. Conclusion

In conclusion, we have proposed to use a masked autoencoder (FMAE) with diverse pre-training for the AU detection task in this paper. We have leveraged a vast and diverse dataset (Face9M) for pretraining, combined with masked image modeling to significantly improve AU detection performance. Moreover, we have demonstrated the identity learning issue and its harmful effect on AU prediction models. The use of Identity Adversarial Training helped in mitigating the model's learning of identity-related features. Also, we elucidated the two design factors of IAT, and our experiments consistently demonstrated superior performance over previous methods, achieving new SOTA results on AU benchmarks like BP4D, BP4D+ and DISFA.

We also noticed that the scaling effect of FMAE pretrained on Face9M has not converged even using the ViT-large model. The potential of using ViT-huge and distilling it into a smaller model for practical use is promising, and we leave this for future work.

ETHICAL IMPACT STATEMENT

Our work on Facial Action Unit (AU) detection aims to advance the understanding of human facial behavior while adhering to ethical standards. We provide our trained models and our codebase as open-source to facilitate further research and application development. We utilize publicly available databases to train our models. Note that some ethnicities or age groups may not be represented well in the large datasets. By incorporating identity removal techniques, our work seeks to mitigate biases in facial behavior analysis and promote fairness across diverse populations. However, the representations obtained using our models may still contain residual identity information. Therefore, caution should be exercised while using our pretrained models for downstream tasks. Our models should not be used in applications to disadvantage minorities (e.g., develop systems to automatically hire employees by looking at their facial behavior during the interviews).

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35:23716–23736, 2022.

[3] R. An, A. Jin, W. Chen, W. Zhang, H. Zeng, Z. Deng, and Y. Ding. Learning facial expression-aware global-to-local representation for robust action unit detection. *Applied Intelligence*, 54(2):1405–1425, 2024.

[4] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas. Learning to learn by gradient descent by gradient descent. *NeurIPS*, 29, 2016.

[5] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pages 15619–15629, 2023.

[6] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE international joint conference on biometrics (IJCB)*, pages 464–473. IEEE, 2017.

[7] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *ICLR*, 2021.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

[9] A. Bulat, S. Cheng, J. Yang, A. Garbett, E. Sanchez, and G. Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. In *ECCV*, pages 107–125. Springer, 2022.

[10] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat. Marlin: Masked autoencoder for facial video representation learning. In *CVPR*, pages 1493–1504, 2023.

[11] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Face and Gesture Recognition*, pages 67–74. IEEE, 2018.

[12] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021.

[13] Y. Chang and S. Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *CVPR*, pages 20417–20426, 2022.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[15] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[16] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, pages 113–123, 2019.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020.

[19] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[20] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189. PMLR, 2015.

[21] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[22] Z. Gao and I. Patras. Self-supervised facial representation learning with facial region awareness. In *CVPR*, pages 2081–2092, 2024.

[23] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[26] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *NeurIPS*, 30, 2017.

[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[29] K. L. Hermann, H. Mobahi, T. Fel, and M. C. Mozer. On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*, 2023.

[30] J. Hernandez, R. Villegas, and V. Ordonez. Vic-mae: Self-supervised representation learning from images and video with contrastive masked autoencoders. *arXiv preprint arXiv:2303.12001*, 2023.

[31] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. pmlr, 2015.

[32] G. M. Jacob and B. Stenger. Facial action unit detection with transformers. In *CVPR*, pages 7680–7689, 2021.

[33] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

[34] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882, 2016.

[35] D. Kim and B. C. Song. Emotion-aware multi-view contrastive learning for facial emotion recognition. In *European Conference on Computer Vision*, pages 178–195. Springer, 2022.

[36] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[37] I. Lee, E. Lee, and S. B. Yoo. Latent-ofer: detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In *ICCV*, pages 1536–1546, 2023.

[38] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8594–8601, 2019.

[39] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *Face and Gesture Recognition*, pages 103–110. IEEE, 2017.

[40] X. Li, X. Zhang, T. Wang, and L. Yin. Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity. In *ICCV*, pages 20979–20989, 2023.

[41] Y. Li and S. Shan. Contrastive learning of person-independent representations for facial action unit detection. *IEEE Transactions on Image Processing*, 32:3212–3225, 2023.

[42] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *NeurIPS*, 36, 2024.

[43] X. Liu, K. Yuan, X. Niu, J. Shi, Z. Yu, H. Yue, and J. Yang. Multi-scale promoted self-adjusting correlation learning for facial action unit detection. *arXiv preprint arXiv:2308.07770*, 2023.

[44] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[45] H. Lu, X. Niu, J. Wang, Y. Wang, Q. Hu, J. Tang, Y. Zhang, K. Yuan, B. Huang, Z. Yu, et al. Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing. In *CVPR*, pages 322–331, 2024.

[46] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *IJCAI*, 2022.

[47] B. Ma, R. An, W. Zhang, Y. Ding, Z. Zhao, R. Zhang, T. Lv, C. Fan, and Z. Hu. Facial action unit detection and intensity estimation from self-supervised representation. *IEEE Transactions on Affective Computing*, 2024.

[48] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[49] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[50] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[51] L. Meng, Y. Liu, X. Liu, Z. Huang, W. Jiang, T. Zhang, C. Liu, and Q. Jin. Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild. In *CVPR*, pages 2345–2352, 2022.

[52] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

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

[54] M. Ning, I. O. Ertugrul, D. S. Messinger, J. F. Cohn, and A. A. Salah. Automated emotional valence estimation in infants with stochastic and strided temporal sampling. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2023.

[55] M. Ning, E. Sangineto, A. Porrello, S. Calderara, and R. Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *ICML*, pages 26245–26265. PMLR, 2023.

[56] İ. Onal Ertugrul, L. Yang, L. A. Jeni, and J. F. Cohn. D-pattnet: Dynamic patch-attentive deep network for action unit detection. *Frontiers in computer science*, 1:11, 2019.

[57] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[58] E. Raff and J. Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198. IEEE, 2018.

[59] L. Shan and W. Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018.

[60] Z. Shang, B. Liu, F. Teng, and T. Li. Learning contrastive feature representations for facial action unit detection. *arXiv preprint arXiv:2402.06165*, 2024.

[61] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *CVPR*, pages 6248–6257, 2021.

[62] T. Song, Z. Cui, W. Zheng, and Q. Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *CVPR*, pages 6267–6276, 2021.

[63] Y. Tang, W. Zeng, D. Zhao, and H. Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *ICCV*, pages 12899–12908, 2021.

[64] C. Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[65] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[66] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017.

[67] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[68] Y. Wang, J. Peng, J. Zhang, R. Yi, L. Liu, Y. Wang, and C. Wang. Toward high quality facial representation learning. In *ACM MM*, pages 5048–5058, 2023.

[69] Z. Wang, S. Song, C. Luo, S. Deng, W. Xie, and L. Shen. Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition. In *CVPR*, pages 1270–1280, 2024.

[70] Z. Wu and J. Cui. La-net: Landmark-aware learning for reliable facial expression recognition under label noise. In *ICCV*, pages 20698–20707, 2023.

[71] F. Xue, Y. Sun, and Y. Yang. Unsupervised facial expression representation learning with contrastive local warping. *arXiv preprint arXiv:2303.09034*, 2023.

[72] H. Yang, U. Ciftci, and L. Yin. Facial expression recognition by de-expression residue learning. In *CVPR*, pages 2168–2177, 2018.

[73] H. Yang, L. Yin, Y. Zhou, and J. Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *CVPR*, pages 10482–10491, 2021.

[74] J. Yang, Y. Hristov, J. Shen, Y. Lin, and M. Pantic. Toward robust facial action units' detection. *Proceedings of the IEEE*, 111(10):1198–1214, 2023.

[75] J. Yang, J. Shen, Y. Lin, Y. Hristov, and M. Pantic. Fan-trans: Online knowledge distillation for facial action unit detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6019–6027, 2023.

[76] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[77] Y. Yin, D. Chang, G. Song, S. Sang, T. Zhi, J. Liu, L. Luo, and M. Soleymani. Fg-net: Facial action unit detection with generalizable pyramidal features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6099–6108, 2024.

[78] K. Yuan, Z. Yu, X. Liu, W. Xie, H. Yue, and J. Yang. Auformer: Vision transformers are parameter-efficient facial action unit detectors. *ECCV*, 2024.

[79] K. Zhang, D. Li, W. Luo, J. Liu, J. Deng, W. Liu, and S. Zafeiriou. Edface-celeb-1 m: Benchmarking face hallucination with a million-scale dataset. *IEEE TPAMI*, 45(3):3968–3978, 2022.

[80] X. Zhang, T. Wang, X. Li, H. Yang, and L. Yin. Weakly-supervised text-driven contrastive learning for facial behavior understanding. In *ICCV*, pages 20751–20762, 2023.

[81] X. Zhang, H. Yang, T. Wang, X. Li, and L. Yin. Multimodal channel-mixing: Channel and spatial masked autoencoder on facial action unit detection. In *WACV*, pages 6077–6086, 2024.

[82] X. Zhang and L. Yin. Multi-modal learning for au detection based on multi-head fused transformers. In *Face and Gesture Recognition*, pages 1–8. IEEE, 2021.

[83] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[84] Y. Zhang, C. Wang, X. Ling, and W. Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *ECCV*, pages 418–434. Springer, 2022.

[85] Y.-H. Zhang, R. Huang, J. Zeng, and S. Shan. M 3 f: Multi-modal continuous valence-arousal estimation in the wild. In *Face and Gesture Recognition*, pages 632–636. IEEE, 2020.

[86] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, pages 3438–3446, 2016.

[87] Z. Zhang, S. Zhai, L. Yin, et al. Identity-based adversarial training of deep cnns for facial action unit recognition. In *BMVC*, page 226. Newcastle, 2018.

[88] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016.

[89] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, pages 18697–18709, 2022.

[90] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *AAAI*, volume 34, pages 13001–13008, 2020.