2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Bridging the Gap: Protocol Towards Fair and Consistent Affect Analysis

Guanyu Hu[1,2⋆], Eleni Papadopoulou[3⋆], Dimitrios Kollias[2⋆], Paraskevi Tzouveli[3], Jie Wei[1]
and Xinyu Yang[1]
[1] The School of Computer Science and Technology, Xi'an Jiaotong University, China
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K
[3] National Technical University of Athens, Greece

Abstract— **The increasing integration of machine learning algorithms in daily life underscores the critical need for fairness and equity in their deployment. As these technologies play a pivotal role in decision-making, addressing biases across diverse subpopulation groups, including age, gender, and race, becomes paramount. Automatic affect analysis, at the intersection of physiology, psychology, and machine learning, has seen significant development. However, existing databases and methodologies lack uniformity, leading to biased evaluations. This work addresses these issues by analyzing six affective databases, annotating demographic attributes, and proposing a common protocol for database partitioning. Emphasis is placed on fairness in evaluations. Extensive experiments with baseline and state-of-the-art methods demonstrate the impact of these changes, revealing the inadequacy of prior assessments. The findings underscore the importance of considering demographic attributes in affect analysis research and provide a foundation for more equitable methodologies. Our annotations, code and pre-trained models are available at: https://github.com/dkollias/Fair-Consistent-Affect-Analysis**

I. Introduction

In recent years, the rapid integration of machine and deep learning algorithms into various aspects of our daily lives has prompted a growing awareness of the critical need for fairness and equity in their deployment. As these technologies play an increasingly pivotal role in decision-making processes, there is a pressing concern to ensure that their outcomes do not inadvertently reinforce or perpetuate existing societal biases. One of the pivotal dimensions in this pursuit of fairness is the examination of algorithmic behavior across diverse subpopulation groups, including but not limited to age, gender, and race. The ethical imperative of addressing potential biases and discriminatory impacts on different demographic cohorts is paramount, demanding a concerted effort to develop machine learning models that are not only accurate and efficient but also considerate of the diverse and nuanced characteristics within our society.

Automatic affect analysis has a long history of studies in the intersection of physiology, psychology, and machine/deep learning. The analysis involves automatic: a) recognition of categorical affect, represented via the so-called six basic expressions (e.g., anger, disgust, happiness), plus the neutral state [6]; b) detection of activation of facial action units (AUs) (i.e., specific movements of facial muscles) [6]; c) estimation of dimensional affect, represented via the emotion descriptors of valence (characterises an emotional state on a scale from positive to negative) and arousal (characterises an emotional state on a scale from active to passive) [24].

Affect analysis methods are being developed using existing databases with the goal of maximizing their performance on the test sets and surpassing the performance of other methods. In most of the cases, this is the only criterion and there is no checking whether these comparisons are fair and/or whether the methods are fair, unbiased and perform equally well across subjects of various demographic attributes such as race, gender and age (especially considering that the utilized databases do not contain an even distribution of subjects among the demographic groups). The methods, unless explicitly modified, are severely impacted by such bias since they are given more opportunities (training samples) for optimizing their objectives towards the majority demographic group in the database. This leads to lower performance for the minority groups, i.e., subjects represented with less number of samples, which is not desired.

In the past three decades, many affective databases have been constructed with a clear tendency from small-scale to large-scale and from lab-controlled to real-world unconstrained conditions (termed "in-the-wild"). Two of the most typically utilized in-the-wild databases for Facial Expression Recognition (FER) are RAF-DB [12] and AffectNet [20]. AffectNet is further annotated in terms of valence and arousal (VA). The most widely used databases for AU recognition are DISFA [18], GFT [8] and the in-the-wild EmotioNet [3] and RAF-AU [32]. All these databases have some problems that we present next.

Some databases (DISFA) do not have a pre-defined split into training, validation and test sets. Therefore many research works perform *k*-fold cross validation either with different *k* [19][11], or with the same *k* but not the same samples in each fold. Additionally, some of these works, in the *k*-fold cross-validation, perform a subject-dependent split and others a subject independent one (which means that the same subject can only appear in the same split) [22][21]. All these make the derived results of the methods non-comparable, as these methods have been developed with different training and validation samples (there is also no common test samples to compare their methods).

Other databases are split into only two sets: RAF-DB and RAF-AU have a training and test set; AffectNet has a training

The work of Guanyu Hu was supported by the Innovation Leadership Talent Development Program of Xi'an Jiaotong University

⋆ equal contribution

and validation set (the test set has not been published); RAF-AU does not provide the exact training and test partitions. For RAF-AU, research works define their own partitions without making them available so that other works can use the same partitions while developing their method [1]. For the other databases, research works either use the one set as training set and the other one as both validation and test set [21][38][13][4], or split the existing training set into a new training and validation set (without specifying the samples that belong to each set) and then use these to develop their methods, evaluate their performance on the existing test set and compare with other methods [28][9]. However, the performance comparison between these methods is unfair, as the method in the first approach is developed using more training data compared to the method in the second approach.

Some databases consist of manual and automatic annotations (EmotioNet's training set consists of only automatic labels, whereas its validation and test sets consist of only manual labels; AffectNet also contains a part that is automatically annotated). Some research works utilize automatic annotations and develop their methods. Other works do not use them, because automatic annotations are very noisy. Instead, they opt to split the validation set into new training and validation sets, and leverage these for modeling; however, they partition in different percentages (e.g. [29] split in 90-10%, whereas [31] split in 80-20%) and they also do not report the exact splits. All these are problematic.

The databases' test set sizes are quite small, especially compared to the training set sizes. AffectNet's training set consists of around 290K images, whereas its test set of only 4K images; RAF-DB's training set consists of around 12K images and its test set of only 3K; Most works following the setting [27] for GFT, in which 78 subjects with around 110K frames in training set and 18 subjects with only 25K frames exist in testing set.

The evaluation of a method's performance on the test set is a critical aspect of assessing its efficacy and generalizability, and thus developed methods use this set to compare their performance to other methods. Thus having a small test set may result in wrong and inaccurate conclusions from these comparisons; if a method outperforms another in that small set, it does not mean that it would still outperform the other in a larger test set. We show this in the experimental section. After we re-partitioned the databases, we train various state-of-the-art methods and compared their performance; we show that methods that outperformed others in the old partitions, are not still outperforming them on the new partitions.

Some databases have inconsistent forms of annotations. AffectNet's VA annotations are in continuous [-1, 1], whereas AFEW-VA's annotations are integers in [-10, 10]. DISFA is annotated in terms of AU intensities (from a scale of 0-5). To convert the AU intensities into AU activation/non-activation, some research works associate an intensity higher than 0 to activation and an intensity of 0 to non-activation [13][15]; other works associate an intensity higher than 1 to activation and an intensity of 0 or 1 to non-activation [35][5]. DISFA contains videos from the left and right view of 27 subjects, but many research works only consider one view. What is more, for AU annotated databases, a different combination of AUs is being used in each database and in some of them not all the released AU annotations are used. DISFA contains annotations of 12 AUs but research works use only 8; GFT contains annotations of 32 AUs but works use only 10. All AUs should be used for consistency with the other databases, so that they contain more common AUs, which will enable and facilitate cross-database experiments.

The evaluation metrics of the databases are not common and/or are not the most appropriate ones. In FER, research works that utilize RAF-DB mainly report the total accuracy and less often the average accuracy (across all basic expressions), whereas works using AffectNet report the total accuracy and F1 score. Total accuracy, especially in imbalanced test sets, is not an ideal measure as it can be misleading; thus F1 score and average accuracy are more appropriate measures. In AU detection, research works that utilize DISFA and GFT, utilize the F1 score, works that use RAF-AU use AUC-ROC curve and F1 score, whereas works that use EmotioNet use the mean between the average accuracy and the F1 score. For VA estimation, research works that utilize AffectNet utilize the Concordance Correlation Coefficient (CCC), Pearson Correlation Coefficient (PCC) and Mean Squared Error (MSE).

The databases have inconsistent label distributions in each set partition, which along with the heavy label imbalance and the small sizes constitute a bad combination. A crucial assumption in machine learning is that the data distributions and their labels are the same for the training, validation and test sets. However, this is not the case in many databases (e.g., AffectNet, EmotioNet); this phenomenon, known as label shift, severely affects methods' performance, since minimizing the objective on the validation set no longer results in good performance on the test set.

The databases do not necessarily contain an even distribution of subjects in terms of demographic attributes such as race, gender and age and/or such attributes have not been taken into consideration when the databases' partitions were performed. For instance, GFT includes only subjects aged 21-28, but its test set does not include subjects aged 25-27; GFT does not contain any Asian subjects in the test set; the large-scale AffectNet had only 27 test images of people aged 70 or older. *Finally, most databases do not contain labels regarding these attributes, making it difficult to assess bias and methods' performance in each attribute.*

All things considered, the contributions of this work are:

- annotating six affective databases in terms of demographic attributes (age, gender and race);
- partitioning these affective databases according to a common protocol that we define; in that protocol we pay particular attention to the demographic attributes and to fairness of evaluations;
- conducting extensive experimental study of various baseline and state-of-the-art methods in each database using the new protocol; various performance metrics are utilized, including ones measuring fairness and bias

II. Materials & Methods

Here, we describe the annotation of the 7 affective databases for demographic attributes. We also define and describe a common protocol for partitioning these databases, taking into account the problems presented in the introduction section and the demographic annotations. Finally, we present the new partition of the databases and provide relevant statistics.

A. Annotation

We annotated all affective databases (AffectNet, RAF-DB, DISFA, EmotioNet, GFT and RAF-AU) in terms of demographic attributes of age, gender, and race. Let us note that although, in practice, race and ethnicity terms are used interchangeably, they are not the same. Race and ethnicity are different categorizations of humans. Race is defined based on physical traits and ethnicity is based on cultural similarities [26]. Therefore we decided to annotate the databases in terms of race (facial appearance is easier to judge and annotate). We adopted a commonly accepted race classification from the U.S. Census Bureau and thus we defined the following 5 race groups (in alphabetical order): Asian, Black (or African American), Indian (or Alaska Native), Native Hawaiian (or Other Pacific Islander), and White (or Caucasian). During the annotation process, we did not find any cases of Hawaiian or Other Pacific Islanders.

We further annotated all databases in terms of gender, with the options being: Male, Female, and Other/Uncertain. The “Uncertain” category mainly includes cases that are challenging to determine the gender, e.g., infants below the age of 3 or adolescents aged 4-19 with significant obstructions in the central part of the image, such as hats or hands.

Finally, we annotated all databases for age. We divide the age into nine categories, i.e. 0–2, 3–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70 and over. We made this categorization: i) to be consistent with age ranges in other databases (e.g. FairFace[10], RAF-DB); ii) to provide more fine-grained categories (rather than ones like ‘young’, ‘middle-aged’, ‘old’); and iii) to make the annotation easier and the process less noisy (it is easier to tell if a person is between 40 and 49 years old, rather than if he is 42 or 49 years old).

We developed our own annotation software that enabled the annotation of each image and video independently in terms of race, gender, and age group for each subject. Two labelers annotated around 20,000 images from each image-database and all videos from each video-database. If both agreed on their judgments, we took the values as ground truth. The annotation process involved a final step. We trained a model with our ground truths and tested it on the rest of the images of each image-database so that we create annotations for these images as well.

B. Protocol

In this subsection, we define the protocol for creating new partitions for each of the 7 affective databases. Each database has been split in the same way, with the same rationale (for consistency), which involves: i) partitioning into 3 sets (training, validation and test); ii) each set having an adequate amount of data and subjects (to a possible extent given the whole dataset size); iii) each set having similar distributions in terms of affect labels (basic expressions, action units, or valence-arousal), age groups, race groups and genders; iv) each set being subject independent; v) utilization of only manual annotations of affect labels; vi) usage of specific performance evaluation metrics; vii) usage of a consistent annotation form.

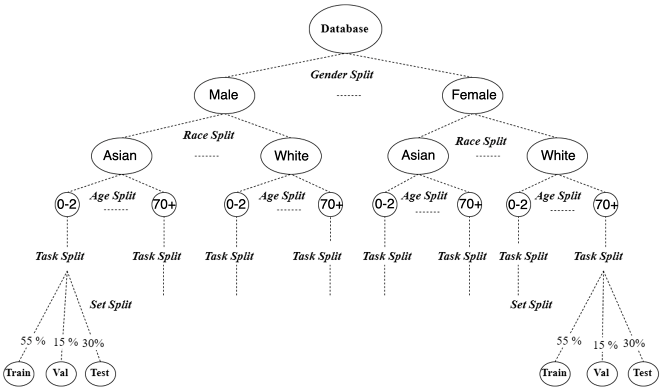


Fig. 1. The proposed protocol and subsequent partition of a database

To give some more details on the above points: for point (vi), the F1 score is selected for evaluating AU detection and basic expression recognition (average accuracy is also good in the later case); CCC is selected for evaluating VA estimation (average CCC of valence and arousal). For point (vii), the VA label values of AFEW-VA are scaled to the range [-1,1]. In DISFA, all 12 AUs are utilized; AU labels with an intensity higher than 0 correspond to AU activation, while an intensity of 0 corresponds to AU non-activation; both subjects’ views are included in the new partition. All 14 AUs in GFT are used.

Finally, for point (iii), the training, validation and test sets follow the 55%-15%-30% rule, according to which the training set consists of 55% of the data (spanning all affect labels and all race, all age groups and all genders), the validation set consists of 15% of the data and the test set consists of the remaining 30% of the data. Figure 1 illustrates the proposed protocol and database partition. Figure 2 illustrates the ’Task Split’ part of the proposed protocol and database partition in the case of basic expressions. In the case of VA, at first we convert the continuous 2D space into distinct regions (one such region is when valence takes values in [-1,-0.8] and arousal takes values in [-0.8,-0.6]). Figure 3 illustrates the ’Task Split’ part of the proposed protocol and database partition in the case of VA.

C. Databases: Original vs New Partition

Tables I, II and III, as well as Figure 4 present information (Label, Gender, Race, Age) regarding the original and new partitions (according to the previously mentioned protocol) of all utilized affective databases.

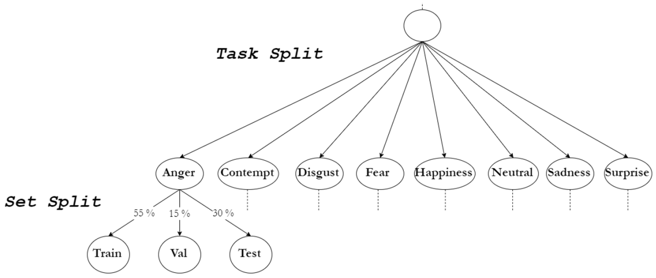


Fig. 2. 'Task Split' part of protocol & partition in case of basic expressions

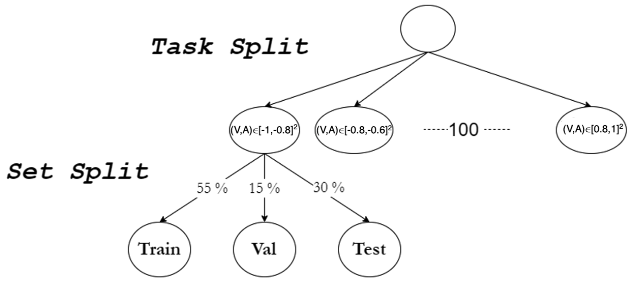


Fig. 3. 'Task Split' part of proposed protocol and partition in case of VA

D. Performance Measures

In the following, we present the performance measures that we selected for validating the performance and fairness of the models on each task (Expression Recognition, AU Detection and VA Estimation).

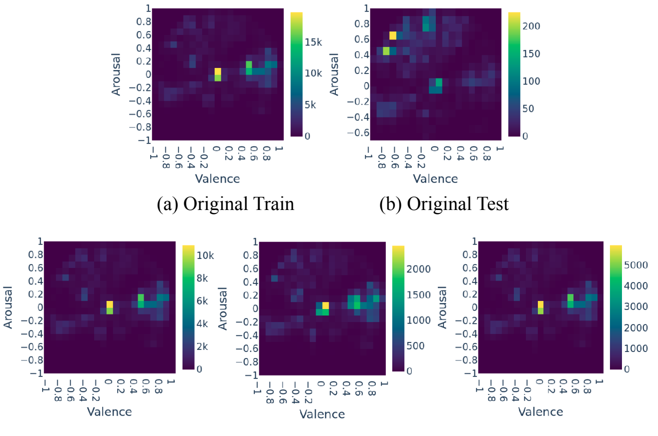
1) Expression Recognition

F1 Score: It represents the average F1 Score across all expression categories (i.e., macro F1 Score). It takes values in [0, 1]. Higher values are desirable in evaluation.

Statistical Parity (SP): It quantifies the extent of disparity in model predictions across different subgroups within a demographic attribute. It is the average of absolute differences in predicted probabilities between any distinct pairs of subgroups with the same class prediction. For each demographic attribute, SP is:

$$\frac{2\sum_{c=1}^{k}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left|p\left(\hat{\mathbf{y}}=c \mid \mathbf{s}=s_i,\mathbf{x}\right)-p\left(\hat{\mathbf{y}}=c \mid \mathbf{s}=s_j,\mathbf{x}\right)\right|}{n(n-1)},$$

where:
s_i and s_j represent distinct subgroups of a demographic attribute (e.g. s_i is 'Asian', s_j is 'Black' and the demographic attribute is 'Race');
$p(\hat{\mathbf{y}}=c \mid \mathbf{s}=s_i,\mathbf{x})$ denotes the probability of predicting class c with sample $\mathbf{x}$ belonging to subgroup s_i;
k represents the number of expression categories;
n represents the number of subgroups of a demographic attribute.
SP takes values in [0, 1]. A value of 0 is desirable; values in $[0, 0.1]$ indicate fair models.



(a) Original Train (b) Original Test

(c) New Train (d) New Valid (e) New Test

Fig. 4. 2D Valence-Arousal Histogram: A Comparison Between the Original and New Partitions of AffectNet

TABLE I
DATA STATISTICS FOR THE ORIGINAL AND NEW PARTITIONS OF DATABASES ANNOTATED IN TERMS OF BASIC EXPRESSIONS

Database		AffectNet					RAF-DB				
Partition		Original		New			Original		New		
Set		Train	Test	Train	Valid	Test	Train	Test	Train	Valid	Test
Total		287651	3999	159540	43330	87710	12271	3068	8327	2206	4806
Gen.	Female	144422	1794	80281	21808	44127	6562	1620	4455	1189	2538
	Male	142172	2192	79259	21522	43583	4957	1249	3366	891	1949
Race	Asian	25573	311	14168	3823	7893	1912	483	1285	329	781
	Black	23578	321	13072	3518	7309	968	234	631	156	415
	Indian	17201	234	9526	2552	5357	/	/	/	/	/
	White	220242	3120	122774	33437	67151	9391	2351	6411	1721	3610
Age	$\leq$ 2	15418	288	8607	2327	4772	1283	329	865	219	528
	03-09	17794	257	9895	2684	5472	2171	486	1436	377	844
	10-19	15042	177	8337	2249	4633					
	20-29	117914	1464	65624	17881	35873	6531	1662	4480	1212	2501
	30-39	56462	821	31476	8564	17243					
	40-49	28280	427	15758	4272	8677	1920	502	1312	344	766
	50-59	22950	363	12790	3467	7056					
	60-69	9395	162	5227	1407	2923					
	$\geq$ 70	3339	27	1826	479	1061	366	89	234	54	167
Expr.	Neutral	74874	500	41252	11223	22588	2524	680	1744	463	997
	Happiness	134415	500	73958	20144	40432	4772	1185	3260	877	1820
	Sadness	25459	500	14174	3842	7816	1982	478	1332	353	775
	Surprise	14090	500	7951	2148	4425	1290	329	872	229	518
	Fear	6378	500	3729	997	2113	281	74	184	44	127
	Disgust	3803	500	2318	611	1355	717	160	471	119	287
	Anger	24882	500	13859	3755	7649	705	162	464	121	282
	Contempt	3750	499	2299	610	1332	/	/	/	/	/

2) Action Unit Detection

F1 Score: It represents the average F1 Score across all AUs (i.e., binary F1 score). It takes values in [0, 1]. Higher values are desirable in evaluation.

Demographic Parity Difference(DPD): It measures the disparity in model predictions across different demographic subgroups. For each demographic attribute, DPD is:

$$\frac{\sum_{c=1}^{k}\left(\max_{i\in[1,n]}\left[p_c\left(\hat{\mathbf{y}}=1 \mid \mathbf{s}=s_i,\mathbf{x}\right)\right]-\min_{i\in[1,n]}\left[p_c\left(\hat{\mathbf{y}}=1 \mid \mathbf{s}=s_i,\mathbf{x}\right)\right]\right)}{k},$$

where: k represents the number of AUs; n denotes the number of subgroups within a demographic attribute; $p_c(\hat{\mathbf{y}}=1 \mid \mathbf{s}=s_i,\mathbf{x})$ denotes the probability of predicting

TABLE II

Data Statistics for the Original and New Partitions of Databases Annotated for AUs ('Org.' denotes Original partition)

Database		DISFA				EmotioNet					GFT					RAF-AU			
Partition		**Org.**	**New**			**Org.**		**New**			**Org.**		**New**			**Org.**	**New**		
Set			**Train**	**Valid**	**Test**	**Train**	**Test**	**Train**	**Valid**	**Test**	**Train**	**Test**	**Train**	**Valid**	**Test**		**Train**	**Valid**	**Test**
Total Amount		261628	141041	29070	84442	24644	20361	24674	6705	13549	108549	24645	69618	23685	39891	4601	2469	645	1426
Gender	**Female**	116278	63521	9690	38723	11716	10132	11996	3261	6591	45736	11120	30452	9778	16626	2301	1250	328	723
	Male	145350	77520	19380	45719	12892	10188	12678	3444	6958	62813	13525	39166	13907	23265	2239	1219	317	703
Race	**Asian**	29070	9690	/	19296	1410	1043	1340	359	754	2641	/	/	/	2641	453	243	58	152
	Black	9690	/	/	9690	1979	1935	2144	579	1191	7989	4891	9050	/	3830	274	144	34	96
	Indian	9690	/	/	9690	1442	1210	1450	388	814	/	/	/	/	/	252	131	29	92
	White	213178	131351	29070	45766	19777	16132	19740	5379	10790	97919	18505	60568	23685	32171	3561	1951	524	1086
Age	$\leq$ **2**	/	/	/	/	583	262	460	123	262	/	/	/	/	/	352	191	48	113
	03-09	/	/	/	/	939	488	782	210	435	/	/	/	/	/	585	318	83	184
	10-19	9690	/	/	9690	1047	754	986	266	549	/	/	/	/	/	222	117	30	75
	20-29	213178	131351	29070	48413	10101	8912	10454	2849	5710	108549	24645	69618	23685	39891	1662	912	246	504
	30-39	19380	/	/	19333	5026	4206	5075	1380	2777	/	/	/	/	/	1068	584	156	328
	40-49	19380	9690	/	7006	3128	2581	3136	852	1721	/	/	/	/	/	385	208	53	124
	50-59	/	/	/	/	2484	2093	2513	683	1381	/	/	/	/	/	161	86	20	55
	60-69	/	/	/	/	1041	834	1026	278	571	/	/	/	/	/	55	27	5	23
	$\geq$ **70**	/	/	/	/	259	190	242	64	143	/	/	/	/	/	50	26	4	20
AU1		17556	9425	2082	5868	1452	1170	1973	371	587	4011	23540	2999	431	1686	1076	595	164	307
AU2		14728	7482	512	6480	689	713	1051	247	383	14527	3009	8817	2920	5799	795	412	121	254
AU4		49188	25692	3118	19623	2857	610	3941	675	1207	3989	836	3671	466	688	1817	994	235	576
AU5		5458	4081	302	954	875	1287	1252	243	424	2600	397	1214	166	1617	985	599	140	225
AU6		38968	24160	2866	11067	4572	4777	7289	1395	2809	30787	6826	20954	7037	9622	450	218	60	166
AU7		/	/	/	/	/	/	/	/	/	49403	11108	34042	10580	15889	/	/	/	/
AU9		14264	7510	1668	4606	505	143	494	62	34	1519	24267	1289	95	513	774	385	88	294
AU10		/	/	/	/	/	/	/	/	/	26740	6079	19032	5786	8001	1390	661	205	509
AU11		/	/	/	/	/	/	/	/	/	14926	3710	11117	3022	4497	/	/	/	/
AU12		61588	33754	8828	18233	7546	6641	12430	2397	5009	32021	7186	21578	7426	10203	1268	584	178	488
AU15		15724	8631	1878	5206	/	/	/	/	/	11536	2350	8617	2231	3038	/	/	/	/
AU16		/	/	/	/	/	/	/	/	/	/	/	/	/	/	720	302	107	299
AU17		25860	12127	2290	10931	492	184	515	57	58	32723	7984	20395	6502	13810	541	321	85	135
AU20		9064	5216	484	2909	134	146	132	17	7	/	/	/	/	/	/	/	/	/
AU23		/	/	/	/	/	/	/	/	/	27009	6281	17051	5666	10573	/	/	/	/
AU24		/	/	/	/	/	/	/	/	/	15477	3480	10108	1899	6950	/	/	/	/
AU25		92104	45461	18938	26691	11590	9473	18155	3500	6880	/	/	/	/	/	2829	1478	408	901
AU26		211676	26465	12532	10440	2058	1778	3146	651	1156	/	/	/	/	/	1089	578	160	334
AU27		/	/	/	/	/	/	/	/	/	/	/	/	/	/	810	382	98	313

TABLE III

Age Distribution of the Original and New Partitions of GFT

Part.	Set	Age							
		21	**22**	**23**	**24**	**25**	**26**	**27**	**28**
Org.	Train	57179	18644	14004	1486	3896	5797	3274	4269
	Test	9967	10300	916	1249	/	/	/	2213
New	Train	35408	14918	8698	1486	2237	2570	1618	2683
	Valid	14145	4818	3096	/	/	1626	/	/
	Test	17593	9208	3126	1249	1659	1601	1656	3799

activation of AU_c with sample $\mathbf{x}$ belonging to the subgroup s_i; $\max\limits_{i\in[1,n]}$ and $\min\limits_{i\in[1,n]}$ represent the maximum and minimum probability across n demographic subgroups, respectively. DPD takes values in [0, 1]. A value of 0 is desirable; values in $[0, 0.1]$ indicate fair models.

3) Valence-Arousal Estimation

Concordance Correlation Coefficient (CCC): It quantifies the agreement between predicted values and their annotations. Mathematically, CCC is:

$$\frac{2 \cdot s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2},$$

where: s_x and s_y represent the variances of the valence/arousal annotations and predicted values, respectively; $\bar{x}$ and $\bar{y}$ denote the mean values of the corresponding annotations and predictions; s_{xy} represents their covariance. CCC takes values in [-1, 1]. Higher values are desirable in evaluation. We compute the average of the CCC values for valence and arousal.

Average CCC: Evaluates the alignment between predicted values and annotations across diverse demographic subgroups. It is computed by averaging the mean CCC values for valence and arousal across all subgroups:

$$\text{Average CCC} = \frac{1}{2n} \sum_{i=1}^{n} \left(CCC_{\text{valence}}^{(i)} + CCC_{\text{arousal}}^{(i)} \right)$$

where: i indicates the index of n subgroups within a demographic attribute; $CCC_{\text{valence}}^{(i)}$ and $CCC_{\text{arousal}}^{(i)}$ are computed following the method of CCC defined previously.

TABLE IV

Performance comparison (in %) between various baseline and state-of-the-art works for FER; for F1 higher values are wanted; for SP, a value of 0 is wanted, values in $[0, 10]$ indicate fair models

Model	**AffectNet-7**							**AffectNet-8**							**RAF-DB**						
	Test	**Age**		**Gender**		**Race**		**Test**	**Age**		**Gender**		**Race**		**Test**	**Age**		**Gender**		**Race**	
	F1	**SP**	**F1**	**SP**	**F1**	**SP**	**F1**	**F1**	**SP**	**F1**	**SP**	**F1**	**SP**	**F1**	**F1**	**SP**	**F1**	**SP**	**F1**	**SP**	**F1**
ResNet18	58.8	1.5	55.9	0.9	56.2	1.9	56.6	51.7	16.4	49.5	18.0	51.5	7.2	50.5	65.7	24.1	57.5	8.9	64.8	6.4	63.2
ResNet50	58.8	1.8	57.4	0.8	57.7	1.6	58.4	51.4	17.0	49.5	17.7	51.1	7.3	50.9	59.5	23.5	51.1	10.0	57.3	7.0	54.0
ResNext50	58.6	2.1	58.1	0.7	58.6	1.7	59.0	51.8	15.9	49.1	16.5	51.0	6.3	50.0	56.1	22.9	51.3	8.8	56.0	5.0	53.8
DenseNet121	59.4	1.9	59.1	0.3	59.4	1.3	60.2	52.8	15.9	51.1	16.2	52.6	6.5	52.0	64.1	23.6	56.8	9.6	63.1	7.0	60.8
ViT_B_16	58.0	1.8	58.1	0.7	58.1	2.0	58.5	51.3	14.7	49.2	17.0	50.9	6.8	49.8	71.4	24.6	64.8	11.4	71.1	7.7	68.5
VGG16	58.3	1.8	57.2	0.5	57.5	2.0	57.6	51.1	16.5	49.1	17.4	50.9	7.0	50.1	67.1	23.3	59.7	9.7	65.9	8.5	63.9
EffNet_B0	59.3	2.0	59.3	0.9	59.3	1.6	59.8	53.1	15.3	49.4	17.5	51.7	6.2	51.2	66.5	23.5	59.5	9.9	65.4	7.1	63.6
EffNet_B7	59.8	1.9	59.3	0.7	59.8	1.6	60.2	52.9	15.7	49.7	17.2	51.7	6.3	50.9	70.1	22.5	62.1	10.8	68.0	7.5	66.8
Swin_B	59.7	2.1	59.0	0.5	58.8	1.6	59.2	52.1	16.1	50.9	17.7	52.4	7.4	51.6	74.2	23.6	67.5	11.5	73.7	8.0	72.7
Swin_V2_B	59.2	2.0	58.7	0.6	58.3	1.5	58.9	52.0	15.9	50.5	18.2	51.9	7.5	51.4	74.1	24.1	66.9	10.7	72.6	9.2	72.0
ConvNeXt_Base	60.8	2.0	60.2	1.0	60.5	1.4	60.9	53.9	15.4	51.7	18.9	53.7	7.4	52.9	73.2	24.0	67.8	12.2	72.4	6.9	70.3
iResNet101	59.6	1.7	59.5	0.6	59.6	1.6	59.9	52.0	15.2	49.1	18.2	51.3	6.3	49.9	67.9	24.0	61.0	9.7	67.4	7.5	65.3
POSTER++ [17]	61.4	2.0	60.9	0.7	61.3	1.8	61.6	55.4	20.7	52.8	18.0	55.2	8.4	54.9	83.2	24.4	80.3	11.5	82.7	7.9	80.9
DAN [30]	58.9	2.2	56.7	0.5	57.0	2.2	57.6	51.4	22.5	47.7	18.0	49.6	10.1	49.9	77.5	24.3	71.9	12.1	77.1	8.1	74.1
MT-EffNet [25]	57.1	2.2	57.1	0.6	57.1	2.0	57.5	52.0	16.2	49.9	18.1	51.8	6.9	51.5	72.6	24.9	68.1	12.7	71.6	8.3	70.4
MA-Net [37]	62.0	1.9	61.9	0.6	62.0	1.2	62.4	54.2	15.2	52.3	17.2	54.0	6.5	53.2	77.2	23.4	72.4	11.7	76.5	8.5	75.1
EAC [36]	62.4	2.1	61.6	0.6	62.4	1.2	62.7	55.5	14.8	53.5	17.2	55.3	6.5	54.0	81.0	23.3	76.5	12.3	80.1	8.9	79.1
DACL [7]	60.0	2.0	60.2	0.7	60.0	1.8	60.4	52.3	15.6	49.9	17.3	52.1	6.9	51.2	74.2	23.8	69.7	11.2	73.3	8.0	71.7
EmoGCN [2]	58.1	2.7	58.2	0.8	58.0	1.7	57.4	51.9	20	47.8	17.7	51.3	8.7	50.1	70.5	24.1	59.7	11.2	67.9	12.3	66.4
FUXI [34]	59.6	2.1	59.7	0.6	59.6	1.8	60.0	52.8	20.7	50.0	18.2	51.2	9.6	51.4	77.7	25.1	73.0	11.8	76.6	7.7	75.5
SITU [14]	59.2	2.0	59.2	0.5	59.2	1.3	59.6	53.8	15.9	51.9	17.7	53.6	7.7	52.6	73.6	24.8	70.1	10.4	72.8	7.5	73.2
CTC [33]	58.7	1.9	56.6	0.6	56.9	1.2	57.1	53.2	15.6	50.2	16.8	51.8	6.9	51.4	75.6	24.2	72.2	12.4	76.2	7.9	75.7

TABLE V

Performance comparison (in %) between various baseline and state-of-the-art works for AU detection; for F1 higher values are wanted; for DPD, a value of 0 is wanted, values in [0,10] indicate fair models

Model	**DISFA**							**EmotioNet**						
	Test	**Age**		**Gender**		**Race**		**Test**	**Age**		**Gender**		**Race**	
	F1	**DPD**	**F1**	**DPD**	**F1**	**DPD**	**F1**	**F1**	**DPD**	**F1**	**DPD**	**F1**	**DPD**	**F1**
ResNet18	42.7	39.0	36.2	8.5	40.8	39.9	35.3	57.3	15.3	54.7	5.4	57.6	4.2	55.2
ResNet50	39.2	38.5	36.3	3.7	40.4	21.9	36.5	56.0	13.3	52.8	5.0	55.9	2.9	55.2
ResNext50	38.8	39.2	34.9	5.9	38.1	32.7	35.5	55.2	13.7	52.4	4.8	54.9	3.6	52.1
DenseNet121	43.1	31.3	36.7	4.1	40.4	22.7	36.6	58.4	13.3	55.7	4.4	58.3	3.1	55.7
ViT_B_16	40.3	33.9	37.1	6.7	40.0	26.0	37.2	50.8	14.2	47.6	4.7	50.8	3.7	48.8
VGG16	43.2	30.5	43.6	9.4	42.4	27.1	38.3	59.5	15.2	57.5	5.0	58.9	3.4	55.9
EffNet_B0	40.5	38.0	39.0	8.4	39.8	33.9	37.0	56.9	13.7	54.5	4.9	57.2	3.0	56.5
EffNet_B7	41.4	31.1	35.7	5.4	36.9	26.3	32.5	56.1	12.3	53.3	5.1	56.3	3.7	54.1
Swin_B	44.6	36.6	41.5	8.9	42.8	28.8	36.9	56.8	14.7	53.3	5.2	56.2	3.4	53.5
Swin_V2_B	43.1	35.3	40.7	9.1	43.9	31.4	38.8	59.1	14.2	56.5	5.5	58.7	3.5	55.0
ConvNeXt_B	47.8	24.3	43.1	6.0	45.6	19.1	39.7	59.4	16.5	54.8	4.6	59.0	2.8	57.1
iResNet101	43.2	28.8	42.1	6.5	42.8	27.5	36.6	60.1	17.1	56.7	5.3	59.9	3.5	57.0
FUXI [34]	38.3	38.0	39.5	8.1	39.5	36.6	36.0	58.8	17.0	54.8	5.3	58.7	3.9	56.8
SITU [14]	40.6	37.9	39.3	9.8	41.2	35.5	38.4	57.8	14.1	54	4.5	57.7	3.3	57.4
ME-GraphAU [16]	46.1	23.9	46.0	6.7	47.1	19.9	42.9	61.5	13.1	57.6	5.0	61.7	3.3	59.6
AUNets [23]	53.3	27.2	48.5	5.1	52.2	24.0	46.0	62.8	14.3	66.5	5.1	67.3	3.3	64.9

Model	**GFT**							**RAF-AU**						
	Test	**Age**		**Gender**		**Race**		**Test**	**Age**		**Gender**		**Race**	
	F1	**DPD**	**F1**	**DPD**	**F1**	**DPD**	**F1**	**F1**	**DPD**	**F1**	**DPD**	**F1**	**DPD**	**F1**
ResNet18	41.0	43.9	39.3	13.4	41.1	31.7	39.1	68.2	24.6	62.8	2.2	67.8	13.8	66.4
ResNet50	41.5	44.8	39.0	12.9	41.3	30.0	39.4	66.8	25.4	62.5	2.2	66.6	11.3	65.0
ResNext50	39.8	39.2	36.5	9.7	39.6	28.4	36.9	63.7	24.7	56.7	2.6	63.8	13.0	61.3
DenseNet121	41.1	46.6	38.5	12.6	41.5	32.5	38.5	67.9	27.7	62.2	2.9	68.1	11.7	66.8
ViT_B_16	43.0	43.1	37.9	12.5	43.7	32.8	38.9	61.1	22.8	55.2	2.5	61.5	14.5	58.7
VGG16	38.1	34.0	33.9	12.0	37.2	20.9	34.4	55.8	19.2	52.4	1.6	55.9	8.7	54.4
EffNet_B0	42.7	40.4	38.8	9.8	42.3	30.2	37.7	63.6	27.6	58.9	1.9	63.6	10.8	61.7
EffNet_B7	44.8	44.0	41.5	13.8	43.7	35.0	41.3	68.4	26.6	63.3	2.5	68.2	12.4	67.3
Swin_B	42.7	42.1	40.8	14.7	43.4	26.5	40.1	70.8	25.4	66.7	3.2	70.8	11.3	69.8
Swin_V2_B	44.5	43.4	41.1	14.7	44.1	31.3	39.9	70.3	25.6	64.3	2.7	70.1	12.6	69.5
ConvNeXt_B	43.4	45.8	40.4	15.1	42.6	31.5	40.2	73.0	26.7	68.0	2.5	72.7	12.8	71.0
iResNet101	44.9	38.6	41.9	11.0	44.3	29.2	40.1	71.9	25.1	66.3	2.2	71.9	12.5	70.3
FUXI [34]	42.8	35.9	39.3	10.5	41.4	26.3	37.9	68.0	27.6	62.7	3.1	68.3	14.4	67.2
SITU [14]	45.1	48.1	39.0	12.4	43.5	28.0	39.3	72.3	28.1	67.5	2.5	72.4	12.9	71.8
ME-GraphAU [16]	48.6	42.8	43.5	14.3	48.1	26.5	42.7	73.5	25.4	67.1	2.9	73.2	12.5	71.9
AUNets [23]	50.5	41.0	42.9	14.0	49.7	25.5	43.7	72.4	26.3	65.8	2.1	70.9	13.5	67.7

TABLE VI

CCC Performance comparison (in %) between various baseline and state-of-the-art works for VA estimation, a higher value is desired

Model	**Test**	**Age**	**Gen.**	**Race**
ResNet18	70.0	69.0	69.3	69.3
ResNet50	69.3	68.3	69.0	68.6
resnext50	69.0	67.3	68.4	68.3
DenseNet121	69.5	68.0	68.8	68.8
ViT_B_16	70.5	69.0	70.1	69.8
VGG16	70.9	69.7	70.4	70.2
EffNet_B0	71.2	69.3	70.0	69.9
EffNet_B7	71.3	69.9	70.5	70.2
Swin_B	71.8	70.5	70.9	70.6
Swin_V2_B	71.6	70.5	71.1	71.0
ConvNeXt_B	72.5	70.7	71.0	71.0
iResNet101	71.1	69.8	70.5	70.3
FUXI [34]	74.3	72.8	73.3	73.3
SITU [14]	72.2	71.4	72.0	71.7
CTC [33]	72.9	71.2	72.3	71.5

III. EXPERIMENTAL RESULTS

We conducted experiments using datasets partitioned according to our proposed protocol to evaluate the performance and fairness of both models and datasets. These experiments incorporated prevalent baseline models (Base), top-ranking methods from the ABAW series challenges (ABAW), and state-of-the-art (SOTA) models.

A. Expression Recognition

For the FER task, we conducted experiments using the AffectNet dataset (both for 7 and 8 expression categories) and the RAF-DB dataset. We employed the F1 score on the test set (F1 Test) to assess models' overall performance. In addition, we use carefully selected fairness evaluation metrics SP and average F1 across subgroups (F1) to evaluate the fairness of the models. SP measures equality in the probability of the model predicting positive outcomes across different demographic groups. The F1 aims to complement SP by evaluating the model's prediction accuracy across various demographic subgroups. The results are summarized in Table IV.

The F1 Test results demonstrate that SOTA models consistently exhibit superior recognition accuracy compared to baseline models. However, it is noteworthy that the ranking of models varies under our new partitioning protocol. For instance, in the original AffectNet dataset partition, models like MT-EffNet [25] and DAN [30] demonstrated a clear advantage; however, their performance diminished on the newly partitioned dataset, placing them at lower positions. Conversely, models such as EAC [36] and MA-Net[37], which previously ranked behind, exhibited improved performance on our revised dataset partition.

To elucidate the underlying reasons for this shift, we conducted analysis using fairness evaluation metrics. Model DAN and MT-EffNet exhibit significantly higher SP across Age, Gender and Race compared to EAC and MA-Net, indicating unfairness in model predictions across these demographic attributes. Upon closer examination of the data statistic in Table I, we attribute this discrepancy to the small size of the previous test set, which primarily consisted of approximately 4000 images concentrated within the 20-39 age group of the White demographic. In contrast, our new partition includes a more diverse range of samples presenting various races and ages in the test set. Consequently, the limitations of models in generalizing to more complex population distributions became apparent. In such intricate scenarios, the lack of fairness in models compromises their overall recognition performance, which emphasizes the importance of addressing fairness considerations alongside overall performance.

We also observed an intriguing phenomenon with the POSTER++ [17] model. Despite its subpar performance in fairness metrics, it still achieved relatively high overall recognition rates. This suggests that fairness and overall recognition performance may not exhibit a straightforward positive or negative correlation.

To explore the relationship between the two, we analyzed the performance of all models on our newly partitioned AffectNet-7 and AffectNet-8 datasets. In the AffectNet-7 dataset, the SP values of all models across Gender and Race ranged between [0-5], with slightly higher values for Age but still under 10. This indicates that the newly partitioned AffectNet-7 dataset inherently possesses fair features, as all models trained on this dataset exhibit good fairness. Conversely, AffectNet-8 demonstrates fairness only in Gender's SP, mainly ranging between [5-10], with poorer fairness performance in other attributes. This is primarily attributed to the addition of a small number of Contempt samples to the dataset, increasing the complexity of model classification from 7 to 8 classes, thereby making it more challenging for models to maintain fairness.

Furthermore, most models exhibiting superior fairness performance also demonstrate overall recognition proficiency (F1 Test) on both datasets. For instance, EAC [36] model achieves optimal performance in fairness metrics on both datasets, whilst obtaining the highest overall recognition performance. Despite POSTER++ displaying poor fairness metrics on AffectNet-8, it still achieves results comparable to the top-performing EAC model. However, its performance on the inherently fair dataset AffectNet-7 falls short compared to the fair model EAC. This indicates that while POSTER++ may excel in accuracy for specific subgroups, it fails to showcase its superiority on fair datasets. Therefore, we can conclude that while enhancing a model's performance for the majority of groups is undoubtedly effective, enhancing its fairness is even more crucial in intricate scenarios.

To investigate the inherent fairness character of the dataset, we conducted an analysis for model performance and dataset compositions on the newly partitioned RAF-DB and AffectNet-8 in comparison to AffectNet-7, leading to the following understanding: Firstly, as models tackle more complex tasks, achieving fairness becomes increasingly challenging. A comparison between AffectNet-7 and AffectNet-8 reveals that while both datasets share identical data samples for the seven base expressions, the addition of an extra class in AffectNet-8 makes it more difficult for models to achieve fairness. Secondly, smaller dataset scales and significant disparities in demographic attribute distributions can detrimentally impact model fairness. Despite both RAF-DB and AffectNet-7 handling seven-class classification tasks, differences in dataset scale and demographic attribute distributions, as shown in Table I, result in varied fairness performance. For instance, due to limitations in the original dataset's size, the newly partitioned RAF-DB training set contains only 366 samples for the '70+' age group and 865 samples for the '0-3' age group, as illustrated in the fairness evaluation metrics in Table IV, leading to noticeable unfairness in model predictions for the Age attribute.

B. AU Detection

For the AU detection task, we conducted experiments on our newly partitioned datasets: DISFA, EmotioNet, GFT, and RAF-AU. To assess the fairness of models, we carefully selected two fairness metrics: DPD to measure the disparity in positive class proportions across different attribute groups; and average F1 across subgroups (F1) to assess whether the model's performance is consistent across different demographic attributes. The experiment results are presented in Table V.

Analyzing the experimental results on the newly partitioned datasets revealed that all models demonstrated fairness for the Gender attribute across the EmotioNet, DISFA, GFT, and RAF-AU datasets, as evidenced by DPD values below 10. These results suggest that Gender is the attribute most conducive to achieving fairness. We attribute this mainly to the relatively balanced distribution of male and female samples within the datasets compared to other demographic attributes. While the GFT dataset displayed slight deviations from fairness in gender. Upon exploration of the dataset distributions (as shown in Table II), it was noted that the Gender ratios in EmotioNet and RAF-AU were close to 1:1, while GFT had ratios close to 1:1.4. For the Age attribute, F1 scores ranked the lowest across all demographic attributes, indicating considerable difficulty in achieving fairness for this attribute.

Additionally, for the EmotioNet dataset, fairness metrics across all demographic attributes outperformed the other three datasets. This can be attributed to the dataset's "in-the-wild" image nature, which provides richer attribute diversity compared to video datasets like DISFA and GFT. Although RAF-AU shares the "in-the-wild" nature, its limited data size may have hindered model performance. Notably, compared to the FER task, all models demonstrated generally poor fairness across all datasets, particularly with respect to attributes such as Age and Race. This underscores the inherent challenges in addressing fairness in AU tasks.

For the performance of different models across the four datasets, AUNets [23] and ME-GraphAU [16] demonstrated the best overall performance (Test F1) across all datasets. Notably, while ME-GraphAU exhibited slightly weaker performance compared to AUNets on the other datasets, the fairness metric results for ME-GraphAU surpassed those of AUNets across all datasets, establishing it as the best-performing model in terms of fairness among all models. The potential reason for this phenomenon could be that ME-GraphAU may have effectively learned to mitigate biases present in the dataset and make more equitable predictions across demographic subgroups.

This observation challenges the conventional notion that achieving fairness inevitably leads to a decrease in overall performance. On the contrary, it suggests that models can achieve both fairness and better overall performance concurrently. This indicates that there is no inherent trade-off between fairness and performance, debunking a common misconception in fair model design. This insight provides valuable guidance for future fair model development, emphasizing the importance of prioritizing fairness without compromising performance.

C. VA Estimation

For the VA task, experiments were conducted on the newly partitioned AffectNet-VA datasets. Given that VA involves continuous values, traditional fairness metrics such as SP and DPD are not applicable. Instead, the average CCC between valence and arousal across subgroups within different demographic attributes (Average CCC) was employed as a fairness evaluation. The experimental results, presented in Table VI, reveal that methods achieving excellent scores in the 5th ABAW challenge maintain favorable overall performance. Specifically, FUXI [34] and CTC [33] achieved the highest performance. Interestingly, these two models also demonstrated the best performance in fairness metrics, indicating that prioritizing fairness in VA predictions can significantly enhance overall model performance. This observation highlights the importance of incorporating fairness considerations into VA prediction models, as it not only promotes equitable outcomes but also contributes to overall performance enhancement.

In terms of fairness performance across all demographic attributes, Age exhibited the lowest fairness among the three demographic attributes, consistent with findings from previous tasks. This underscores the persistent challenge of achieving fairness for Age predictions.

IV. CONCLUSION

In conclusion, this research underscores the critical importance of addressing biases and promoting fairness in the deployment of machine learning algorithms, particularly within the realm of automatic affect analysis. The study conducted a thorough analysis of seven affective databases, annotating demographic attributes and proposing a standardized protocol for database partitioning with a focus on fairness in evaluations. Through extensive experiments with baseline and SOTA methods, the findings reveal the significant impact of these changes and highlight the inadequacy of prior assessments. The emphasis on considering demographic attributes and fairness in affect analysis research provides a solid foundation for enhancing the model performance and the development of more equitable methodologies.

References

[1] 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*, pages 1–21, 2024.

[2] P. Antoniadis, P. P. Filntisis, and P. Maragos. Exploiting emotional dependencies with graph convolutional networks for facial expression recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021.

[3] C. Benitez-Quiroz, R. Srinivasan, and A. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16)*, Las Vegas, NV, USA, June 2016.

[4] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo. Vitfer: facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4):80, 2022.

[5] Z. Cui, C. Kuang, T. Gao, K. Talamadupula, and Q. Ji. Biomechanics-guided facial action unit detection through force modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8694–8703, 2023.

[6] P. Ekman, W. Friesen, and J. Hager. Facial action coding system (facs). a human face 2002, 2002.

[7] A. H. Farzaneh and X. Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021.

[8] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 581–588. IEEE, 2017.

[9] S. Handrich, L. Dinges, A. Al-Hamadi, P. Werner, and Z. Al Aghbari. Simultaneous prediction of valence/arousal and emotions on affectnet, aff-wild and afew-va. *Procedia Computer Science*, 170:634–641, 2020.

[10] K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[11] J. Kossaifi, A. Toisoul, A. Bulat, Y. Panagakis, T. M. Hospedales, and M. Pantic. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6060–6069, 2020.

[12] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2852–2861, 2017.

[13] Y. Li, J. Zeng, S. Shan, and X. Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer vision and pattern recognition*, pages 10924–10933, 2019.

[14] C. Liu, X. Zhang, X. Liu, T. Zhang, L. Meng, Y. Liu, Y. Deng, and W. Jiang. Facial expression recognition based on multi-modal features for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5878, 2023.

[15] Y. Liu, W. Wang, Y. Zhan, S. Feng, K. Liu, and Z. Chen. Pose-disentangled contrastive learning for self-supervised facial representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9717–9728, 2023.

[16] 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. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 1239–1246, 2022.

[17] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang. Poster v2: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*, 2023.

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

[19] A. Mitenkova, J. Kossaifi, Y. Panagakis, and M. Pantic. Valence and arousal estimation in-the-wild with tensor methods. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.

[20] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

[21] R. Parameshwara, I. Radwan, A. Asthana, I. Abbasnejad, R. Subramanian, and R. Goecke. Efficient labelling of affective video datasets via few-shot & multi-task contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6161–6170, 2023.

[22] R. Parameshwara, I. Radwan, R. Subramanian, and R. Goecke. Examining subject-dependent and subject-independent human affect inference from limited video data. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2023.

[23] A. Romero, J. León, and P. Arbeláez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, 2018.

[24] J. A. Russell. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152, 1978.

[25] A. V. Savchenko, L. V. Savchenko, and I. Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022.

[26] R. T. Schaefer. *Encyclopedia of race, ethnicity, and society*, volume 1. Sage, 2008.

[27] Z. Shao, Y. Zhou, F. Li, H. Zhu, and B. Liu. Joint facial action unit recognition and self-supervised optical flow estimation. *Pattern Recognition Letters*, 2024.

[28] V. Suresh, G. Yeo, and D. C Ong. Critically examining the domain generalizability of facial expression recognition models. *arXiv preprint arXiv:2106.15453*, 2023.

[29] P. Wang, Z. Wang, Z. Ji, X. Liu, S. Yang, and Z. Wu. Tal emotionet challenge 2020 rethinking the model chosen problem in multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 412–413, 2020.

[30] Z. Wen, W. Lin, T. Wang, and G. Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, 8(2):199, 2023.

[31] P. Werner, F. Saxen, and A. Al-Hamadi. Facial action unit recognition in the wild with multi-task cnn self-training for the emotionet challenge. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 410–411, 2020.

[32] W.-J. Yan, S. Li, C. Que, J. Pei, and W. Deng. Raf-au database: in-the-wild facial expressions with subjective emotion judgement and objective au annotations. In *Proceedings of the Asian conference on computer vision*, 2020.

[33] J. Yu, R. Li, Z. Cai, G. Zhao, G. Xie, J. Zhu, W. Zhu, Q. Ling, L. Wang, C. Wang, et al. Local region perception and relationship learning combined with feature fusion for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5784–5791, 2023.

[34] W. Zhang, B. Ma, F. Qiu, and Y. Ding. Multi-modal facial affective analysis based on masked autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2023.

[35] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6077–6086, 2024.

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

[37] Z. Zhao, Q. Liu, and S. Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021.

[38] C. Zheng, M. Mendieta, and C. Chen. Poster: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3146–3155, 2023.