One-Frame Calibration with Siamese Network in Facial Action Unit Recognition

Shuangquan Feng Virginia R. de Sa
University of California San Diego, La Jolla, CA 92093
{s1feng,desa}@ucsd.edu

Abstract

Automatic facial action unit (AU) recognition is used widely in facial expression analysis. Most existing AU recognition systems aim for cross-participant non-calibrated generalization (NCG) to unseen faces without further calibration. However, due to the diversity of facial attributes across different identities, accurately inferring AU activation from single images of an unseen face is sometimes infeasible, even for human experts—it is crucial to first understand how the face appears in its neutral expression, or significant bias may be incurred. Therefore, we propose to perform one-frame calibration (OFC) in AU recognition: for each face, a single image of its neutral expression is used as the reference image for calibration. With this strategy, we develop a Calibrating Siamese Network (CSN) for AU recognition and demonstrate its remarkable effectiveness with a simple iResNet-50 (IR50) backbone. On the DISFA, DISFA+, and UNBC-McMaster datasets, we show that our OFC CSN-IR50 model (a) substantially improves the performance of IR50 by mitigating facial attribute biases (including biases due to wrinkles, eyebrow positions, facial hair, etc.), (b) substantially outperforms the naive OFC method of baseline subtraction as well as (c) a fine-tuned version of this naive OFC method, and (d) also outperforms state-of-the-art NCG models for both AU intensity estimation and AU detection.

1 Introduction

Facial expression analysis is important for understanding human emotions and behaviors across various fields, including human-computer interaction, psychology, and security. The Facial Action Coding System (FACS) is a comprehensive system for describing human facial movement developed by Ekman and Friesen [10], widely recognized and extensively used in facial expression analysis for its ability to describe facial movement objectively and systematically. It breaks down facial expressions into individual components of muscle movement, called action units (AUs). Table 1 introduces the primary AUs analyzed in this paper.

AU1	Inner Brow Raiser	AU12	Lip Corner Puller
AU2	Outer Brow Raiser	AU15	Lip Corner Depressor
AU4	Brow Lowerer	AU17	Chin Raiser
AU5	Upper Lid Raiser	AU20	Lip Stretcher
AU6	Cheek Raiser	AU25	Lips Part
AU9	Nose Wrinkler	AU26	Jaw Drop

Table 1. Names of the primary AUs analyzed in the paper.

As manual AU coding is expensive and time-consuming, automatic AU recognition is used widely in facial expression analysis. Most existing AU recognition systems aim for cross-participant non-calibrated generalization (NCG) to unseen faces [2, 13, 18, 29, 41, 46], where individual images/frames of the faces are fed into the trained model as input for AU recognition. However, due to the diversity of facial attributes across different identities, accurately inferring AU activation from single images of an unseen face is sometimes infeasible, even for human experts—it is crucial to first understand how the face appears in its neutral expression, or significant bias may be incurred. In the official FACS manual [11], the importance of taking the face's neutral appearance into account as the baseline is repeatedly emphasized for human scoring of various AUs. Firstly, without understanding the neutral appearance, permanent facial features (e.g. wrinkles, bulges, pouches) may be misidentified as evidence for AU activation. Secondly, scoring of many AUs is dependent upon the neutral appearance: for example, scoring of AU5 (upper lid raiser) is dependent upon whether the iris shows entirely in the neutral face or is partially covered; scoring of AU15 (lip corner depressor) is dependent upon whether the lip line is straight, slightly up, or slightly down in neutral.

Without face-specific calibration, automatic AU recognition systems would suffer from similar facial attribute biases for faces not seen in the training set. Therefore, we propose to perform one-frame calibration (OFC) in AU recognition: for each face, a single image of its neutral expression is used as the reference image for calibration. The

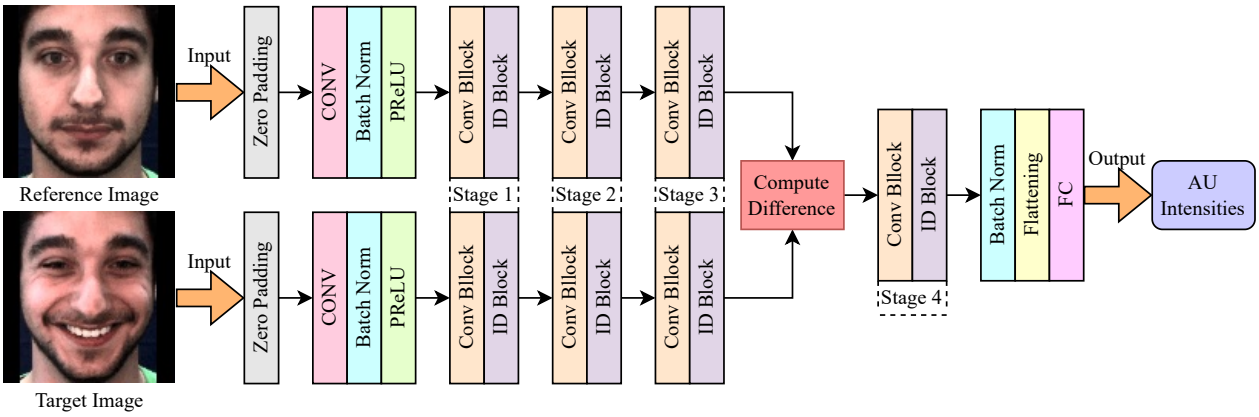


Figure 1. The CSN-IR50 network architecture. The example reference image and target image are from the DISFA dataset.

most intuitive OFC method would be to directly subtract the model's estimations of AUs in the neutral image from its AU estimations of all images of the same face. However, we show that the performance improvement from this method, if any, is limited. We propose a highly effective neural network architecture for OFC—Calibrating Siamese Network (CSN), where the reference image (the neutral face) and the target image are fed into two identical networks and joined in an intermediate stage of the network by computing the difference between their feature maps of that stage. In this paper, we demonstrate its remarkable effectiveness with a simple iResNet-50 (IR50) backbone. On the DISFA [23, 24], DISFA+ [22], and UNBC-McMaster [19] datasets, we show that our OFC CSN-IR50 model (a) significantly improves the performance of IR50 by mitigating facial attribute biases, (b) substantially outperforms the naive OFC method of baseline subtraction (BS) as well as (c) a fine-tuned version of this naive OFC method, and (d) also outperforms state-of-the-art (SOTA) NCG models.

2 Related Work

2.1 Bias in Facial Expression Recognition

Previous studies have shown that facial expression recognition systems can exhibit biases across groups based on gender [9], race [28], age [15], among others [7, 14], raising significant attention and concerns about the fairness of these systems. Due to the objectiveness of FACS by definition, some researchers believe AU recognition is less subject to bias and use it to investigate and mitigate bias in facial emotion annotation [6] and recognition [12, 33]. However, it was shown that AU recognition is also subject to bias, which researchers have developed methods to mitigate [7, 14].

Although group-based biases attract more attention, they are essentially specific manifestations of the broader issue of identity bias in facial expression recognition. Researchers have proposed two methods of addressing identity bias, either to develop identity-aware/personalized models for facial expression recognition [25, 36, 39, 42] or to apply adversarial training with respect to identities on the models to encourage them to disregard identity-related features [45].

2.2 Siamese Neural Network

The Siamese Neural Network was first introduced by Bromley et al. [3] and has been widely applied in facial identity-related tasks, such as face verification [34] and face recognition [40].

3 Methods

3.1 One-Frame Calibration

While most existing AU recognition systems aim for cross-participant non-calibrated generalization (NCG) to unseen faces, it is crucial to take the face's neutral appearance into account in AU coding. Thus, we propose one-frame calibration (OFC) for AU recognition: for each face, a single image of its neutral expression is used as the reference image for calibration.

In offline benchmarking, OFC primarily applies to video datasets. The selection of the reference image is achieved by manually selecting one image from all frames with zero activation of annotated AUs for each face. The aim of manual selection is to ensure that in the selected reference image, (a) the unannotated AUs are also not activated or only minimally activated, and (b) the face is at an appropriate angle and not partially occluded.

In real-life applications of AU recognition systems, the

ideal method of selecting the reference image for OFC is to directly ask the user to pose a neutral face before usage.

3.2 Calibrating Siamese Network for OFC

We propose the Calibrating Siamese Network (CSN) architecture for OFC. The input for CSN consists of the target image for AU recognition and the reference image. The two images are fed into two identical networks with shared weights and joined in an intermediate stage of the network by computing the difference between their feature maps.

This architecture design can be integrated with a variety of model designs for AU recognition. To demonstrate its effectiveness in a simple way, we use the classical iResNet-50 (IR50) as the backbone in this work and name it CSN-IR50.

3.2.1 CSN-IR50

Figure 1 illustrates the architecture of CSN-IR50. The reference image with a neutral expression and the target image for AU recognition are fed into two identical IR50 networks with shared weights; just before reaching stage 4 of the network, the difference between their feature maps is computed and then fed into the rest of the IR50 network until the AU intensities are outputted.

CSN-IR50 may be more precisely called CSN-IR50-Stage4, emphasizing stage 4 as the merge point, which is the main version of CSN-IR50 we primarily investigate in this work. Stage 4 is selected as the merge point because the feature maps in this stage capture high-level abstractions of the face features and still retain the fine-grained information. We will also compare it with other versions, including CSN-IR50-Stage1, CSN-IR50-Stage2, CSN-IR50-Stage3, CSN-IR50-FC, and CSN-IR50-Output.

3.3 Baseline Models

We compare the performance of our proposed method with those of various models. Firstly, we directly compare our method of OFC with CSN-IR50 with the vanilla NCG with IR50 to demonstrate the effectiveness of OFC. Secondly, we compare our model with the naive OFC method of baseline subtraction (BS) (IR50 OFC w/ BS) to demonstrate the superiority of our model as an OFC method. (Table 6 shows a comparison with CSN-IR50-Output, which is a fine-tuned version of this naive method (fine-tuned with same parameters as CSN-IR50)). Additionally, we also compare our model performance with that of other SOTA NCG models.

3.3.1 IR50 (NCG)

In IR50 (NCG), the IR50 is trained on individual images of the training set and directly applied to images of unseen faces during validation.

3.3.2 IR50 (OFC w/ BS)

In IR50 (OFC w/ BS), the IR50 is trained on individual images of the training set; however, during validation, its output on the reference image for each face is used as the baseline, which is subtracted from outputs on all images of the same face to produce final predictions.

4 Experiments

4.1 Datasets and Settings

DISFA [24] contains left-view and right-view facial video recordings of 27 participants with approximately 130,000 frames in total for each view. Each frame is annotated with intensities of 12 AUs on a scale of 0 to 5. Following previous studies, we perform participant-exclusive 3-fold cross-validation on DISFA.

DISFA+ [22] is an extension of the DISFA dataset. It contains facial video recordings of 9 participants' posed and spontaneous facial expression with each frame being annotated with the same 12 AUs on a scale of 0 to 5. We perform leave-one-participant-out cross-validation on DISFA+.

UNBC-McMaster [19] is a dataset originally collected for pain detection. Since it also contains frame-level AU intensity annotations of 10 AUs on a scale of 0 to 5 (except that the annotations for AU43 (eye closure) are binary), it is also appropriate for analyzing AU recognition methods. It contains facial video recordings of 25 participants with 48,398 frames in total. We perform participant-exclusive 5-fold cross-validation on UNBC-McMaster.

We did not include the widely used BP4D [43] dataset because it is not appropriate for OFC. Unlike the previous mentioned datasets, in BP4D, FACS coders selectively annotate a 20-second segment with the highest density of facial expression for each recording session, and only these segments are released in the dataset. Consequently, for most participants in BP4D, there is no appropriate "neutral face frame" to use as the reference image for OFC.

We evaluate our methods on both AU intensity estimation and AU detection. In AU intensity estimation, the model outputs estimations of intensities of the AUs (real values between 0 to 5). In AU detection, the model outputs predictions of whether each AU occurs in binary format (0 or 1). Following previous studies [30], for AU detection, we consider AU intensities greater or equal to 2 as occurrences, and we only include 8 of the 12 AUs for DISFA.

Metric	Method	AU												Average
		1	2	4	5	6	9	12	15	17	20	25	26	
ICC(3,1)↑	CCNN-IT [37]	.18	.15	.61	.07	.65	.55	.82	.44	.37	.28	.77	.54	.45
	2DC [17]	.70	.55	.69	.05	.59	.57	**.88**	.32	.10	.08	.90	.50	.50
	SCC-Heatmap [13]	.73	.44	.74	.06	.27	.51	.71	.04	.37	.04	.94	**.78**	.47
	iARL [30]	.13	.36	.68	.22	.56	.36	.86	**.52**	.37	.12	**.96**	.60	.48
	IR50 (NCG)	.53	.45	.75	.62	.55	.57	.84	.42	.47	.24	.93	.65	.59
	IR50 (OFC w/ BS)	.62	.51	.75	.55	.60	.59	.82	.39	.44	.20	.93	.68	.59
	CSN-IR50 (OFC)	**.75**	**.70**	**.80**	**.72**	**.67**	**.61**	.85	.33	**.52**	**.37**	.94	.77	**.67**
MAE↓	CCNN-IT [37]	87	.63	.86	.26	.73	.57	.55	.38	.57	.45	.81	.64	.61
	SCC-Heatmap [13]	**.16**	**.16**	**.27**	**.03**	**.25**	**.13**	.32	.15	**.20**	.09	.30	.32	**.20**
	iARL [30]	.30	.31	.52	.04	.36	.30	**.31**	**.05**	.33	**.08**	.29	**.26**	.26
	IR50 (NCG)	.37	.39	.44	.11	.35	.21	.34	.20	.39	.21	.32	.42	.31
	IR50 (OFC w/ BS)	.30	.36	.41	.14	.33	.20	.40	.18	.37	.31	.34	.38	.31
	CSN-IR50 (OFC)	.19	**.16**	.38	.08	.26	.19	**.31**	.17	.22	.13	**.27**	.27	.22

Table 2. The performance of different methods on AU intensity estimation on the DISFA dataset. For each metric, the best results in each column are shown in bold. The rows below the dashed lines in each section include the three methods we propose for comparison; the best results among them are underlined.

4.2 Implementation Details

Each frame is preprocessed with face detection [20], face alignment [5], and a combination of histogram equalization and linear mapping [16] for both training and validation.

We use the weights pre-trained on Glink360k [1, 8] for both IR50 and CSN-IR50, and the last layer of the network is modified to adapt to the output format for the AU recognition task.

For AU intensity estimation, we train the network to perform both regression and ordinal classification [26] on the AU intensities. The network outputs the estimation of the AUs in two formats: for estimating the intensity of the ith AU y_i, it outputs 1 value $\hat{y}_{i,\text{reg}}$ representing the numerical estimation of the AU intensity (in the format of regression) and 5 values $\sigma(\hat{y}_{i,\text{class}(1)})$, $\sigma(\hat{y}_{i,\text{class}(2)})$, $\sigma(\hat{y}_{i,\text{class}(3)})$, $\sigma(\hat{y}_{i,\text{class}(4)})$, and $\sigma(\hat{y}_{i,\text{class}(5)})$ respectively representing the estimated probability of the AU intensity being higher than or equal to 1, 2, 3, 4, and 5 (in the format of binary classifications). The loss function consists of three parts:

$$E_{\text{AUIE}} = E_{\text{reg,MSE}} + E_{\text{reg,cos}} + E_{\text{class}}, \tag{1}$$

where $E_{\text{reg,MSE}}$, $E_{\text{reg,cos}}$, and E_{class} respectively represent a mean squared error (MSE) loss for the numerical estimations

$$E_{\text{reg,MSE}} = \Sigma_{i=1}^{n} w_{i,y_i}(y_i - \hat{y}_{i,\text{reg}})^2, \tag{2}$$

a cosine similarity loss for the numerical estimations

$$E_{\text{reg,cos}} = 1 - \frac{\Sigma_{i=1}^{n} y_i \hat{y}_{i,\text{reg}}}{(\Sigma_{i=1}^{n} y_i^2)(\Sigma_{i=1}^{n} \hat{y}_{i,\text{reg}}^2)}, \tag{3}$$

and a cross entropy loss for the binary classification estimations

$$E_{\text{class}} = \Sigma_{i=1}^{n} \Sigma_{j=1}^{5} w_{i,j,\chi_{y_i \geq j}} CE(\chi_{y_i \geq j}, \sigma(\hat{y}_{i,\text{class(j)}})), \tag{4}$$

with the cross entropy function being

$$CE(y, \hat{y}) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]. \tag{5}$$

The weights for the MSE loss and those for the cross entropy loss are both inverse-frequency weighted and normalized within each AU for addressing class imbalance in the datasets (substantially higher number of occurrences for low AU intensities). However, since the number of occurrences of high intensities are too few for most AUs (resulting in too high weights for the MSE loss if used directly), we "bin" the intensities into 2 groups, and each group shares the same weight. Specifically, for the MSE loss, we apply one weight for occurrences of intensities of 0 and 1 and another weight for occurrences of intensities of 2, 3, 4, and 5, and these weights are computed based on the total number of occurrences within each intensity group.

Notably, although we train the network to learn both numerical estimations and binary classification estimations of the AU intensities, only the numerical estimations are used in model validation (and any further model inference).

For AU detection, we train the network to directly output the estimated probability of the occurrence of each AU $\sigma(\hat{y}_i)$, with an occurrence defined as $y_i \geq 2$. The loss function uses cross entropy loss:

$$E_{\text{AUD}} = \Sigma_{i=1}^{n} w_{i,\chi_{y_i \geq 2}} CE(\chi_{y_i \geq 2}, \sigma(\hat{y}_i)), \tag{6}$$

with similar inverse-frequency weights normalized within each AU. The detailed equations for weight computation in both AU intensity estimation and AU detection are provided in the technical appendix.

For model training, we employ the Adam optimizer with an initial learning rate of 10^{-4} for parameters of the last layer and an 10^{-5} for other parameters, a weight decay of 5×10^{-4}, and a batch size of 64. These hyperparameters were selected based on our prior work on other models for

AU recognition. For each fold of each dataset, we train each model for 3 epochs with the random seed 42 in a single run using PyTorch version 2.0.0 in Python version 3.9.7, as initial exploration showed that, with the pre-trained weights, performance does not significantly change after the first epoch. For all training/validation on DISFA, we used single NVIDIA GeForce GTX 1080Ti 11G GPUs and an Intel® Core™ i9-7900X CPU with 128 GB RAM; for all training/validation on DISFA+ and UNBC-McMaster, we used single NVIDIA RTX A6000 GPUs and dual AMD EPYC 7302 CPUs with a 512 GB 16-core 8-channel RAM.

4.3 Results of CSN-IR50

4.3.1 Main Results

Table 2 reports the performance of CSN-IR50 on AU intensity estimation on DISFA in comparison to other methods. Firstly, CSN substantially improves the performance of IR50 for NCG and also greatly outperforms the naive OFC method of IR50 with BS, with a difference of 0.08 in ICC(3,1) and a difference of 0.09 in Mean Absolute Error (MAE). Secondly, in comparison to other SOTA NCG methods, our CSN-IR50 demonstrates a substantially higher ICC(3,1) of 0.67 and a near-best MAE of 0.22 (the best being 0.20). ICC(3,1) measures the consistency between the model estimations and the human expert labels in the dataset. We believe ICC(3,1) is a better metric here because of the high imbalance of DISFA.

Table 3 reports the performance of CSN-IR50 on AU detection on DISFA in comparison to other models. Firstly, it again substantially improves the performance of IR50 for NCG and outperforms the naive OFC method of IR50 with BS[1] in both F1 score and accuracy. Secondly, in comparison to other SOTA NCG methods, our CSN-IR50 demonstrates a higher F1 score and equal-to-best accuracy of 94.1.

Note that the comparison between our CSN-IR50 and other SOTA models here is not an apples to apples comparison because CSN-IR50 is for OFC while the SOTA models are for NCG. However, also note that the effectiveness of our proposed CSN architecture is demonstrated only using the simple IR50 as the backbone for simplicity in our paper, and we believe the CSN architecture design has great potential to be integrated with more complicated, advanced backbone models for AU recognition to achieve higher performance.

[1]IR50 (OFC w/ BS) has substantially worse performance than other methods because baseline subtraction is intrinsically not appropriate for outputting AU occurrences in binary format. The IR50 network output for a neutral face is supposed to be very close to that for the reference image, either slightly higher and slightly lower. Thus, after baseline subtraction, the final output might be either a small positive or a small negative. When it is a small positive, it would be considered as high consistency in AU intensity estimation but would be considered as a false positive in AU detection.

As shown in Table 4, our CSN-IR50 demonstrates similar superiority over IR50 (NCG) and IR50 (OFC w/ BS) on the DISFA+ and UNBC-McMaster datasets. No results of other methods are shown because both of them have been hardly used for evaluating recent SOTA AU recognition models.

4.3.2 The Advantage of OFC with CSN

One interesting question is how/why OFC with CSN-IR50 outperforms the vanilla IR50. Two important observations provide insights into this question.

Firstly, Table 5 compares within-participant ICC(3,1) averaged over all participants and across-participant ICC(3,1) between different methods on AU intensity estimation on DISFA, DISFA+, and UNBC-McMaster. Within-participant ICC(3,1) only measures the consistency between the model estimations and human expert labels within individual participants; across-participant ICC(3,1) is the version used everywhere else in this paper, as it measures the consistency not only within but also across participants and thus more insightfully captures bias across different participants. As shown in Table 5, although our CSN-IR50 greatly improves the across-participant ICC(3,1) of IR50, its improvement in within-participant ICC(3,1) is much more modest. This suggests that the primary advantage of OFC with CSN-IR50 lies in its ability to calibrate for diverse facial attributes across different identities, which aligns with our original intent for CSN.

The comparison of precision and recall in Table 3 offers further insights into how the calibration is achieved: the CSN architecture generally increases precision while decreasing recall for most AUs. In other words, the CSN architecture substantially reduces the misidentification of non-activated AUs as activated (false positives), although this improvement comes with the cost of missing some actual AU activations (false negatives).

This reduction of false positives is achieved through mitigating facial attribute biases. More specifically, without face-specific calibration, some facial attributes are easily misidentified as AU activations, and our CSN addresses this issue. See Figure 2 for a variety of case examples. In Figure 2a, IR50 tends to overestimate AU1 (inner brow raiser) intensities due to the participant's wider eyebrow-to-eye distances, because AU1 produces wider distances between eyebrows and eyes; in Figure 2b, IR50 tends to overestimate AU4 (brow lowerer) intensities due to the participant's slight permanent wrinkle at the root of the nose, because AU4 may produce horizontal wrinkles at the root of the nose; in Figure 2c, IR50 tends to misidentify the bridge of eyeglasses as wrinkles caused by AU9 (nose wrinkler) activation, because AU9 produces wrinkles at the root of the nose; in Figure 2d; IR50 tends to misidentify the facial

Metric	Method	AU								Average
		1	2	4	6	9	12	25	26	
F1 score↑	ARL [30]	43.9	42.1	63.6	41.8	40.0	76.2	95.2	66.8	58.7
	UGN-B [32]	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
	JÂA-Net [31]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	**67.4**	63.5
	PIAP [35]	50.2	51.8	71.9	50.6	54.5	**79.7**	94.1	57.2	63.8
	ME-GraphAU [21]	54.6	47.1	72.9	**54.0**	**55.7**	76.7	91.1	53.0	63.1
	KDSRL [4]	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5
	CLEF [44]	64.3	61.8	68.4	49.0	55.2	72.9	89.9	57.0	64.8
	SACL [18]	62.0	**65.7**	74.5	53.2	43.1	76.9	**95.6**	53.1	65.5
	MDHR [38]	**65.4**	60.2	**75.2**	50.2	52.4	74.3	93.7	58.2	66.2
	IR50 (NCG)	35.4	33.0	64.4	48.6	51.8	77.1	91.9	58.1	57.5
	IR50 (OFC w/ BS)	16.3	17.4	36.1	21.1	11.5	39.1	61.8	26.6	28.7
	CSN-IR50 (OFC)	65.3	58.3	70.8	52.6	51.7	77.3	94.6	65.4	**67.0**
Accuracy↑	ARL [30]	92.1	92.7	88.5	91.6	95.9	**93.9**	97.3	94.3	93.3
	UGN-B [32]	95.1	93.2	88.5	**93.2**	**96.8**	93.4	94.8	93.8	93.4
	JÂA-Net [31]	**97.0**	**97.3**	88.0	92.1	95.6	92.3	94.9	**94.8**	94.0
	SACL [18]	96.1	96.9	**92.5**	91.7	95.0	93.7	**97.5**	89.1	94.1
	IR50 (NCG)	87.1	87.1	86.8	89.5	95.5	93.2	95.3	89.6	90.5
	IR50 (OFC w/ BS)	50.5	62.5	46.3	42.3	37.2	60.1	65.8	55.5	52.5
	CSN-IR50 (OFC)	96.9	96.9	90.4	91.7	94.6	93.5	96.9	92.8	**94.2**
Precision↑	IR50 (NCG)	23.6	21.2	54.8	39.7	46.8	67.9	89.0	45.0	48.5
	CSN-IR50 (OFC)	73.1	69.1	65.7	47.7	41.2	70.1	92.5	56.9	64.5
Recall↑	IR50 (NCG)	71.0	73.9	78.0	62.6	58.1	89.2	94.9	81.8	76.2
	CSN-IR50 (OFC)	59.0	50.4	76.8	58.7	69.5	86.3	96.8	76.8	71.8

Table 3. The performance of different methods on AU detection on the DISFA dataset. For each metric, the best results in each column are shown in bold. The rows below the dashed lines in each of the F1 score and accuracy sections include the three methods we propose for comparison; the best results among them are underlined. For precision and recall, the better results between IR50 (NCG) and CSN-IR50 (OFC) are also underlined.

Method	AU Intensity Estimation		AU Detection	
	ICC(3,1)↑	MAE↓	F1↑	Accuracy↑
DISFA+				
IR50 (NCG)	.81	.37	67.3	91.7
IR50 (OFC w/ BS)	.83	.32	28.3	47.4
CSN-IR50 (OFC)	.86	.23	78.6	96.2
UNBC-McMaster				
IR50 (NCG)	.30	.29	25.9	93.3
IR50 (OFC w/ BS)	.34	.23	9.2	36.1
CSN-IR50 (OFC)	.45	.20	34.2	95.9

Table 4. Performance of different methods on the DISFA+ and UNBC-McMaster datasets. The results shown here are all average values across all AUs. The best results in each column are underlined.

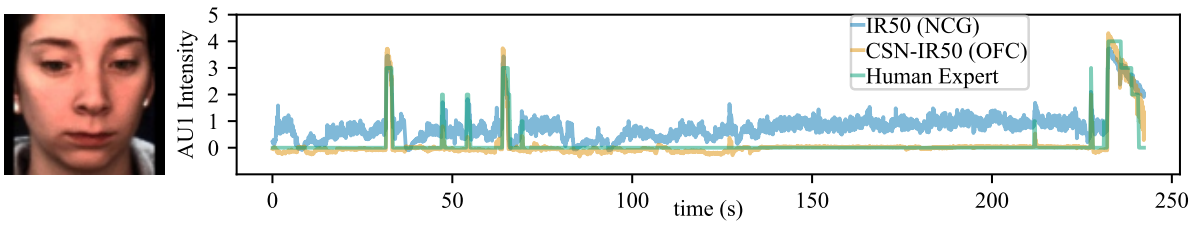
Method	Across-Participant ICC(3,1)↑	Within-Participant ICC(3,1)↑
DISFA		
IR50 (NCG)	.59	.51
CSN-IR50 (OFC)	.67	.53
DISFA+		
IR50 (NCG)	.81	.84
CSN-IR50 (OFC)	.86	.85
UNBC-McMaster		
IR50 (NCG)	.30	.22
CSN-IR50 (OFC)	.45	.26

Table 5. Comparison of within-participant ICC(3,1) averaged across all participants and across-participant ICC(3,1) between different methods on AU intensity estimation. (The across-participant ICC(3,1) is what we use everywhere else, as it more insightfully captures bias across different participants.) The results shown here are all average values across all AUs. The best results in each column are underlined.

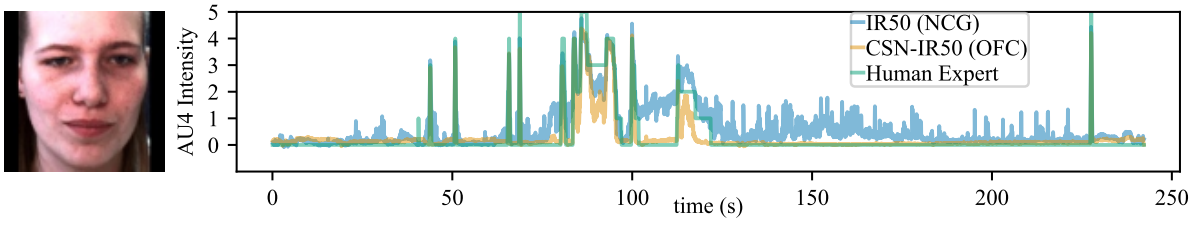
hair as wrinkles caused by AU17 (chin raiser) activation, because AU17 produces wrinkles on the chin boss. Interestingly, the first two examples are issues human FACS coders may also face without a neutral reference, while the latter two examples are facial attribute misidentification problems specific to machine learning models, partially due to insufficient training data, a limitation common in AU datasets. CSN-IR50, effectively addresses these issues by calibrating the AU estimations of different faces based on their neutral appearances.

4.3.3 Comparing Different Versions of CSN-IR50

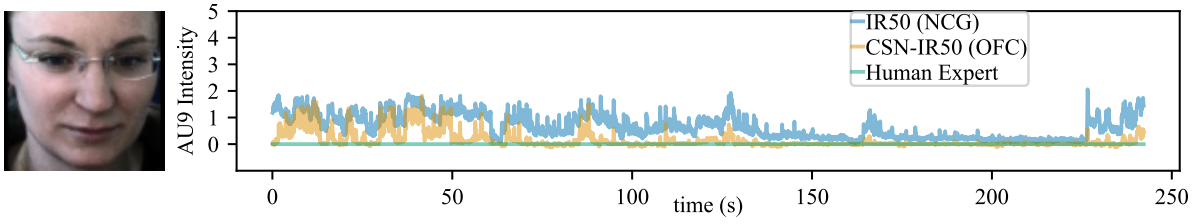
The CSN-IR50 we have presented so far is our main version, CSN-IR50-Stage4, in which the two networks for the reference image and the target image respectively merge just before stage 4 of IR50 (see Figure 1). Table 6 compares it with other versions of CSN-IR50 with different merge points on DISFA. (Each version is named after the first



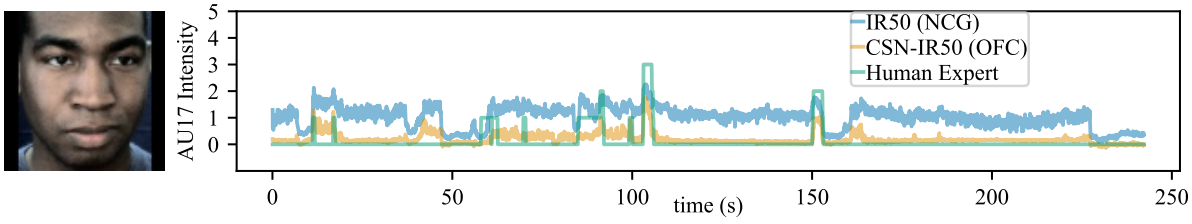
(a) CSN-IR50 (OFC) mitigates bias in AU1 (inner brow raiser) intensity estimation due to wider eyebrow-to-eye distances.



(b) CSN-IR50 (OFC) mitigates bias in AU4 (brow lowerer) intensity estimation due to the wrinkle.



(c) CSN-IR50 (OFC) mitigates bias in AU9 (nose wrinkler) intensity estimation due to the bridge of eyeglasses.



(d) CSN-IR50 (OFC) mitigates bias in AU17 (chin raiser) intensity estimation due to the facial hair.

Figure 2. Examples of CSN-IR50 mitigating facial attribute biases in AU intensity estimation. In each subfigure, the left panel is the (preprocessed) reference image of the specific participant, and the right panel is the comparison between AU intensities estimated by IR50 (NCG) and CSN-IR50 (OFC) and the human expert labeled intensities in the dataset. All examples are from the left-view videos of the DISFA dataset.

module after the merge point.) We can see that our selected merge point, stage 4, is the optimal one providing the best performance. Merging at earlier stages (CSN-IR50-Stage1, CSN-IR50-Stage2, and CSN-IR50-Stage3) provides somewhat suboptimal performances but still outperforms the IR50 baselines. We believe these versions suffer from insufficient processing of individual faces before the merge but still benefit from calibration with the neutral reference. On the other hand, merging later just before the fully connected layer (CSN-IR50-FC) or directly computing the difference between the output AU estimations of the two networks (CSN-IR50-Output) provides substantially worse performance possibly because the more fine-grained information is already lost at that stage. Note CSN-IR50-Output is the fine-tuned version of our naive baseline OFC method (IR50 (OFC w/ BS)). Fine-tuning seems to have a small effect on this method, slightly improving MAE but reducing ICC(3,1) on AU intensity estimation and slightly improving F1 and accuracy on AU detection on DISFA. Performance is much worse than our CSN-IR50-Stage4 model, which only differs in where the merging (difference computation) takes place.

4.3.4 Full Results

For the results presented in Tables 4 to 6, the full versions including individual values for each AU are in the supplementary material (see Tables S1 to S9).

Method	AU Intensity Estimation		AU Detection	
	ICC(3,1)↑	MAE↓	F1↑	Accuracy↑
CSN-IR50-Stage1	.61	.31	60.8	92.0
CSN-IR50-Stage2	.63	.29	65.2	93.3
CSN-IR50-Stage3	.66	.26	62.4	92.5
CSN-IR50-Stage4	.67	.22	67.0	94.2
CSN-IR50-FC	.55	.28	30.9	60.3
CSN-IR50-Output	.54	.28	29.9	58.5

Table 6. Performance of different versions of CSN-IR50 on AU intensity estimation and detection on the DISFA datasets. The suffix indicates where the two networks in CSN-IR50 merge; for example, CSN-IR50-Stage4 means that the two network merges (by computing their difference) just before stage 4 of IR50. The boxed CSN-IR50-Stage4 is the main version we use in the rest of the paper (and shown in Figure 1). The results shown here are all average values across all AUs. The best results in each column are underlined.

5 Discussion and Conclusion

In this paper, we propose to perform OFC in AU recognition for better generalizing the model to unseen faces and a CSN architecture design for OFC. For simplicity, we demonstrate the effectiveness of the CSN architecture with an IR50 backbone. On the DISFA, DISFA+, and UNBC-McMaster datasets, we show that our OFC CSN-IR50 model (a) substantially outperforms the performance of IR50 with NCG, (b) substantially outperforms IR50 with the naive OFC method of BS, as well as (c) the fine-tuned version of this method we call CSN-IR50-Output (note that it only differs from our model in where the merging takes place), and (d) also outperforms SOTA NCG models for both AU intensity estimation and AU detection. With further analysis, we show that the superiority of OFC with CSN-IR50 lies in its ability to calibrate for diverse facial attributes across different identities. Specifically, it substantially reduce false positives in AU recognition, albeit at the cost of increasing false negatives. With case examples, we demonstrate how the reduction of false positives is achieved through mitigating overestimation and misidentification of AUs due to various facial attribute biases, including eyebrow locations, wrinkles, eyeglasses, and facial hair.

One important note is that, while comparison with CSN-IR50-Output and IR50 (OFC w/ BS) is fair and shows large performance improvement, the comparison between our CSN-IR50 and other SOTA models is not an apples to apples comparison because CSN-IR50 is for OFC, enhanced with one important labeled frame of neutral expression from each participant in the validation set, while the SOTA models are for NCG. However, also note that the effectiveness of our proposed CSN architecture is demonstrated only using the simple IR50 as the backbone for simplicity in our paper. Additionally, CSN is not a replacement for existing NCG AU recognition models; rather, it is an augmentation that can be integrated with any existing model. Therefore, as an important future direction, we believe that our CSN architecture design has great potential to be integrated with more complicated, advanced backbone models for AU recognition to achieve higher performance.

Admittedly, OFC has limitations in real-life applications because of its reliance on a good reference image—a neutral face at an appropriate angle and not partially occluded. The ideal method of selecting the reference image is to directly asking the user to pose a neutral face before using the system. Despite its great accuracy and efficiency, this method only applies to scenarios where the user willingly uses the system with full awareness (which includes a wide range of applications, such as healthcare, education, and entertainment). One potential solution for other scenarios is to develop a method to automatically select a good reference image from real-time video streaming, which would also be an interesting future direction to explore.

In conclusion, we propose to perform OFC with a novel CSN architecture design for AU recognition and demonstrate its remarkable effectiveness with a simple IR50 backbone. We also believe it has great potential to be integrated with better backbone models to achieve higher performance.

6 Acknowledgements

We thank Xiaojing Xu and Yuan Tang for helpful prior work. We are grateful for support from NSF IIS 1817226 and IIS 2208362 and seed funding from UC San Diego Social Sciences and the Sanford Institute for Empathy and Compassion as well as hardware funding from NVIDIA, Adobe, and Sony.

References

[1] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition CNNs by partial FC. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4051, 2022. 4

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016. 1

[3] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993. 2

[4] Yanan Chang and Shangfei Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 20417–20426, 2022. 6

[5] Cunjian Chen. PyTorch Face Landmark: A fast and accurate facial landmark detector, 2021. Open-source software available at https://github.com/cunjian/pytorch_face_landmark. 4

[6] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021. 2

[7] Nikhil Churamani, Ozgur Kara, and Hatice Gunes. Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition. *IEEE Transactions on Affective Computing*, 14(4):3191–3206, 2022. 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 4

[9] Artem Domnich and Gholamreza Anbarjafari. Responsible ai: Gender bias assessment in emotion recognition. *arXiv preprint arXiv:2103.11436*, 2021. 2

[10] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1

[11] Paul Ekman, Wallace V Friesen, and Joseph C Hager. *Facial Action Coding System: The Manual*. A Human Face, 2002. 1

[12] Sarah Fabi, Xiaojing Xu, and Virginia R. de Sa. Exploring the racial bias in pain detection with a computer vision model. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, pages 358–365, 2022. 2

[13] Yingruo Fan, Jacqueline Lam, and Victor Li. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12701–12708, 2020. 1, 4

[14] Ozgur Kara, Nikhil Churamani, and Hatice Gunes. Towards fair affective robotics: continual learning for mitigating bias in facial expression and action unit recognition. *arXiv preprint arXiv:2103.09233*, 2021. 2

[15] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 638–644, 2021. 2

[16] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis. A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2121–2129, 2018. 4

[17] Dieu Linh Tran, Robert Walecki, Stefanos Eleftheriadis, Bjorn Schuller, Maja Pantic, et al. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3190–3199, 2017. 4

[18] Xin Liu, Kaishen Yuan, Xuesong Niu, Jingang Shi, Zitong Yu, Huanjing Yue, and Jingyu Yang. Multi-scale promoted self-adjusting correlation learning for facial action unit detection. *arXiv preprint arXiv:2308.07770*, 2023. 1, 6

[19] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 57–64. IEEE, 2011. 2, 3

[20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 4

[21] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022. 6

[22] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. Extended DISFA dataset: Investigating posed and spontaneous facial expressions. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–8, 2016. 2, 3

[23] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, and Philip Trinh. Automatic detection of non-posed facial action units. In *2012 19th IEEE International Conference on Image Processing*, pages 1817–1820. IEEE, 2012. 2

[24] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. DISFA: a spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2, 3

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

[26] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 4

[27] Melinda Ozel. FACS LITE – STILL & FAST, 2024. A visual reference guide for FACS at https://melindaozel.com/facs-lite-still-fast (last accessed on 03/13/2024). 2

[28] Abdallah Hussein Sham, Kadir Aktas, Davit Rizhinashvili, Danila Kuklianov, Fatih Alisinanoglu, Ikechukwu Ofodile, Cagri Ozcinar, and Gholamreza Anbarjafari. Ethical ai in facial expression analysis: racial bias. *Signal, Image and Video Processing*, 17(2):399–406, 2023. 2

[29] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018. 1

[30] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 13(3):1274–1289, 2019. 3, 4, 6

[31] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. JÂA-Net: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321–340, 2021. 6

[32] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5993–6001, 2021. 6

[33] Varsha Suresh and Desmond C Ong. Using positive matching contrastive loss with facial action units to mitigate bias in facial expression recognition. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2022. 2

[34] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 2

[35] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. PIAP-DF: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12899–12908, 2021. 6

[36] Cheng-Hao Tu, Chih-Yuan Yang, and Jane Yung-jen Hsu. Idennet: Identity-aware facial action unit detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2

[37] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017. 4

[38] Zihan Wang, Siyang Song, Cheng Luo, Songhe Deng, Weicheng Xie, and Linlin Shen. Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1270–1280, 2024. 6

[39] Xiaojing Xu and Virginia R de Sa. Personalized pain detection in facial video with uncertainty estimation. In *International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 4163–4168. IEEE, 2021. 2

[40] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4362–4371, 2017. 2

[41] Kaishen Yuan, Zitong Yu, Xin Liu, Weicheng Xie, Huanjing Yue, and Jingyu Yang. AUformer: Vision transformers are parameter-efficient facial action unit detectors. *arXiv preprint arXiv:2403.04697*, 2024. 1

[42] Haifeng Zhang, Wen Su, Jun Yu, and Zengfu Wang. Identity–expression dual branch network for facial expression recognition. *IEEE transactions on cognitive and developmental systems*, 13(4):898–911, 2020. 2

[43] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. BP4D-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 3

[44] Xiang Zhang, Taoyue Wang, Xiaotian Li, Huiyuan Yang, and Lijun Yin. Weakly-supervised text-driven contrastive learning for facial behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20751–20762, 2023. 6

[45] Zheng Zhang, Shuangfei Zhai, Lijun Yin, et al. Identity-based adversarial training of deep cnns for facial action unit recognition. In *BMVC*, page 226. Newcastle, 2018. 2

[46] Kaili Zhao, Wen-Sheng Chu, and Honggang 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. 1

One-Frame Calibration with Siamese Network in Facial Action Unit Recognition

Supplementary Material

A Visual Reference Guide for the Primary AUs

See Figure S1 for a visual reference guide for the primary AUs analyzed in this paper (the AUs included in the DISFA and DISFA+ datasets).

B Equations for Loss Weight Computation

We illustrated the method of computing the weights for the losses in our main paper. The exact equations for computing the weights are included in this section.

In AU intensity estimation, the weights for the MSE loss are defined as

$$w_{i,j} = \begin{cases} \frac{2 \cdot \frac{1}{\Sigma_{j'=0}^{1} n_{i,j'}}}{2 \cdot \frac{1}{\Sigma_{j'=0}^{1} n_{i,j'}} + 4 \cdot \frac{1}{\Sigma_{j'=2}^{5} n_{i,j'}}}, & \text{for } j = 0, 1 \\ \frac{4 \cdot \frac{1}{\Sigma_{j'=2}^{5} n_{i,j'}}}{2 \cdot \frac{1}{\Sigma_{j'=0}^{1} n_{i,j'}} + 4 \cdot \frac{1}{\Sigma_{j'=2}^{5} n_{i,j'}}}, & \text{for } j = 2, 3, 4, 5 \end{cases} \tag{7}$$

while the weights for the cross entropy loss are defined as

$$\begin{cases} w_{i,j,1} = \frac{\frac{1}{\Sigma_{j'=j}^{5} n_{i,j'}}}{\Sigma_{j''=1}^{5} (\frac{1}{\Sigma_{j'=0}^{j''-1} n_{i,j'}} + \frac{1}{\Sigma_{j'=j''}^{5} n_{i,j'}})} \\ w_{i,j,0} = \frac{\frac{1}{\Sigma_{j'=0}^{j-1} n_{i,j'}}}{\Sigma_{j''=1}^{5} (\frac{1}{\Sigma_{j'=0}^{j''-1} n_{i,j'}} + \frac{1}{\Sigma_{j'=j''}^{5} n_{i,j'}})}, \end{cases} \tag{8}$$

where $n_{i,j}$ represents the number of occurrences of the ith AU with an intensity of j.

In AU detection, the weights for the cross entropy loss are defined as

$$\begin{cases} w_{i,1} = \frac{\frac{1}{n_{y_i \geq 2}}}{\frac{1}{n_{y_i \geq 2}} + \frac{1}{n_{y_i < 2}}} \\ w_{i,0} = \frac{\frac{1}{n_{y_i < 2}}}{\frac{1}{n_{y_i \geq 2}} + \frac{1}{n_{y_i < 2}}}, \end{cases} \tag{9}$$

where $n_{y_i \geq 2}$ and $n_{y_i < 2}$ represent the number of occurrences for $y_i \geq 2$ and $y_i < 2$ respectively.

C Full Results

Tables S1 to S4 present the breakdown of the results in Table 4 for each AU. Tables S5 to S7 present the breakdown of the results in Table 5 for each AU. Tables S8 and S9 present the breakdown of the results in Table 6 for each AU.

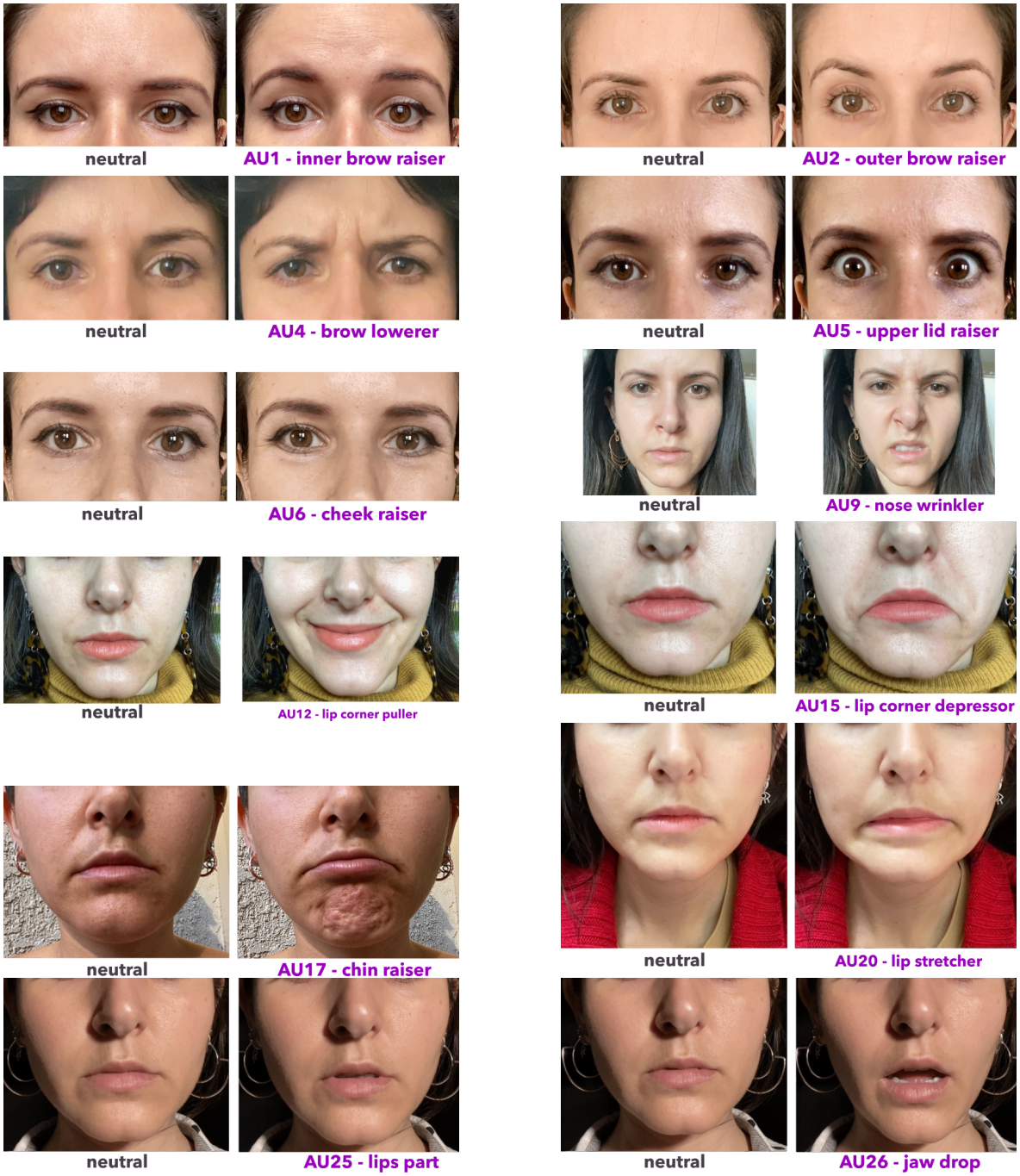


Figure S1. A visual reference guide for the primary AUs extracted from [27].

Metric	Method	AU												Average
		1	2	4	5	6	9	12	15	17	20	25	26	
ICC(3,1)↑	IR50 (NCG)	.81	.76	.88	.79	.87	.87	.90	.80	.73	.60	.91	.82	.81
	IR50 (OFC w/ BS)	.83	.85	.87	.83	.88	.89	.90	.81	.75	.60	.94	.85	.83
	CSN-IR50 (OFC)	.90	.92	.88	.89	.90	.91	.92	.81	.80	.57	.94	.88	.86
MAE↓	IR50 (NCG)	.48	.51	.40	.44	.29	.24	.32	.24	.42	.30	.40	.39	.37
	IR50 (OFC w/ BS)	.49	.42	.35	.42	.26	.21	.29	.21	.30	.26	.29	.29	.32
	CSN-IR50 (OFC)	.26	.22	.29	.29	.23	.18	.24	.17	.21	.21	.24	.24	.23

Table S1. The performance of different methods on AU intensity estimation on the DISFA+ dataset. For each metric, the best results in each column are underlined.

Metric	Method	AU												Average
		1	2	4	5	6	9	12	15	17	20	25	26	
F1 score↑	IR50 (NCG)	71.0	63.1	84.0	63.7	85.9	51.0	84.5	64.7	58.8	49.6	74.3	56.9	67.3
	IR50 (OFC w/ BS)	37.4	33.4	38.7	35.3	29.6	12.9	37.3	13.2	18.9	13.7	41.3	27.9	28.3
	CSN-IR50 (OFC)	86.8	84.0	87.0	79.5	86.0	86.4	89.2	70.1	73.3	35.5	95.0	70.9	78.6
Accuracy↑	IR50 (NCG)	90.3	88.6	94.3	87.8	96.0	91.5	95.7	95.5	90.9	93.1	89.1	87.4	91.7
	IR50 (OFC w/ BS)	58.1	58.1	49.1	55.6	37.9	28.9	57.0	44.4	39.4	34.4	54.2	51.8	47.4
	CSN-IR50 (OFC)	96.5	96.3	95.8	94.6	96.1	98.6	97.1	97.4	96.3	93.9	98.4	93.8	96.2

Table S2. The performance of different methods on AU detection on the DISFA+ dataset. For each metric, the best results in each column are underlined.

Metric	Method	AU									Average
		4	6	7	9	10	12	20	25	26	
ICC(3,1)↑	IR50 (NCG)	.19	.51	.30	.32	.29	.55	.17	.28	.09	.30
	IR50 (OFC w/ BS)	.24	.59	.37	.32	.33	.62	.17	.31	.15	.34
	CSN-IR50 (OFC)	.41	.67	.49	.47	.49	.70	.14	.39	.29	.45
MAE↓	IR50 (NCG)	.21	.48	.34	.12	.11	.48	.19	.33	.34	.29
	IR50 (OFC w/ BS)	.16	.37	.28	.12	.10	.40	.13	.26	.28	.23
	CSN-IR50 (OFC)	.12	.28	.27	.08	.08	.37	.11	.26	.22	.20

Table S3. The performance of different methods on AU intensity estimation on the UNBC-McMaster dataset. For each metric, the best results in each column are underlined.

Metric	Method	AU										Average
		4	6	7	9	10	12	20	25	26	43	
F1 score↑	IR50 (NCG)	17.2	47.1	39.8	19.2	15.4	48.6	2.9	20.0	16.3	32.3	25.9
	IR50 (OFC w/ BS)	4.0	20.6	11.4	1.9	2.1	25.8	2.8	9.3	8.7	5.1	9.2
	CSN-IR50 (OFC)	36.4	56.1	39.3	22.9	24.5	62.1	5.5	39.1	26.9	29.0	34.2
Accuracy↑	IR50 (NCG)	96.5	90.1	95.8	95.9	95.4	88.3	94.8	86.7	92.7	96.5	93.3
	IR50 (OFC w/ BS)	29.9	40.8	35.8	29.6	31.6	46.3	39.7	38.9	35.0	33.4	36.1
	CSN-IR50 (OFC)	98.1	92.9	95.5	97.3	97.8	92.4	97.0	94.7	96.1	96.7	95.9

Table S4. The performance of different methods on AU detection on the UNBC-McMaster dataset. For each metric, the best results in each column are underlined.

Metric	Method	AU												Average
		1	2	4	5	6	9	12	15	17	20	25	26	
Across-Participant ICC(3,1)↑	IR50 (NCG)	.53	.45	.75	.62	.55	.57	.84	.42	.47	.24	.93	.65	.59
	CSN-IR50 (OFC)	.75	.70	.80	.72	.67	.61	.85	.33	.52	.37	.94	.77	.67
Within-Participant ICC(3,1)↓	IR50 (NCG)	.40	.35	.66	.38	.54	.46	.83	.27	.45	.26	.93	.57	.51
	CSN-IR50 (OFC)	.46	.43	.70	.39	.57	.48	.85	.23	.44	.27	.93	.66	.53

Table S5. Comparison of within-participant ICC(3,1) averaged across all participants and across-participant ICC(3,1) between different methods on AU intensity estimation on the DISFA dataset. For each metric, the better results in each column are underlined.

Metric	Method	AU												Average
		1	2	4	5	6	9	12	15	17	20	25	26	
Across-Participant ICC(3,1)↑	IR50 (NCG)	.81	.76	.88	.79	.87	.87	.90	.80	.73	.60	.91	.82	.81
	CSN-IR50 (OFC)	.90	.92	.88	.89	.90	.91	.92	.81	.80	.57	.94	.88	.86
Within-Participant ICC(3,1)↓	IR50 (NCG)	.84	.85	.89	.80	.87	.89	.90	.79	.80	.61	.94	.85	.84
	CSN-IR50 (OFC)	.89	.91	.88	.86	.90	.91	.92	.76	.82	.57	.94	.88	.85

Table S6. Comparison of within-participant ICC(3,1) averaged across all participants and across-participant ICC(3,1) between different methods on AU intensity estimation on the DISFA+ dataset. For each metric, the better results in each column are underlined.

Metric	Method	AU									Average
		4	6	7	9	10	12	20	25	26	
Across-Participant ICC(3,1)↑	IR50 (NCG)	.19	.51	.30	.32	.29	.55	.17	.28	.09	.30
	CSN-IR50 (OFC)	.41	.67	.49	.47	.49	.70	.14	.39	.29	.45
Within-Participant ICC(3,1)↓	IR50 (NCG)	.11	.53	.24	.18	.05	.50	.07	.22	.11	.22
	CSN-IR50 (OFC)	.18	.56	.26	.20	.08	.56	.07	.27	.13	.26

Table S7. Comparison of within-participant ICC(3,1) averaged across all participants and across-participant ICC(3,1) between different methods on AU intensity estimation on the UNBC-McMaster dataset. For each metric, the better results in each column are underlined.

Metric	Method	AU												Average
		1	2	4	5	6	9	12	15	17	20	25	26	
ICC(3,1)↑	CSN-IR50-Stage1	.61	.57	.76	.60	.66	.61	.84	.26	.48	.32	.92	.71	.61
	CSN-IR50-Stage2	.66	.69	.77	.59	.60	.58	.83	.32	.51	.32	.94	.77	.63
	CSN-IR50-Stage3	.68	.68	.81	.67	.64	.59	.85	.36	.55	.38	.95	.74	.66
	CSN-IR50-Stage4	.75	.70	.80	.72	.67	.61	.85	.33	.52	.37	.94	.77	.67
	CSN-IR50-FC	.57	.50	.71	.51	.57	.55	.79	.29	.40	.19	.88	.61	.55
	CSN-IR50-Output	.58	.49	.70	.48	.55	.54	.79	.35	.41	.21	.87	.56	.54
MAE↓	CSN-IR50-Stage1	.34	.31	.48	.13	.31	.23	.36	.24	.35	.23	.33	.41	.31
	CSN-IR50-Stage2	.32	.27	.45	.14	.32	.22	.35	.21	.34	.20	.28	.34	.29
	CSN-IR50-Stage3	.28	.24	.42	.11	.32	.22	.33	.19	.26	.18	.25	.32	.26
	CSN-IR50-Stage4	.19	.16	.38	.08	.26	.19	.31	.17	.22	.13	.27	.27	.22
	CSN-IR50-FC	.28	.30	.44	.13	.31	.20	.34	.17	.28	.17	.41	.33	.28
	CSN-IR50-Output	.28	.29	.43	.12	.30	.20	.37	.16	.29	.19	.42	.34	.28

Table S8. The performance of different versions of CSN-IR50 on AU intensity estimation on the DISFA dataset.

Metric	Method	AU								Average
		1	2	4	6	9	12	25	26	
F1 score↑	CSN-IR50-Stage1	50.8	43.7	65.9	53.3	46.7	75.1	90.6	60.6	60.8
	CSN-IR50-Stage2	54.0	54.7	70.5	52.3	48.4	77.2	93.4	71.5	65.2
	CSN-IR50-Stage3	50.1	44.0	73.5	54.7	41.0	75.3	93.4	67.3	62.4
	CSN-IR50-Stage4	65.3	58.3	70.8	52.6	51.7	77.3	94.6	65.4	67.0
	CSN-IR50-FC	21.7	18.6	38.8	23.2	14.0	42.0	61.9	27.1	30.9
	CSN-IR50-Output	18.8	17.8	39.1	23.0	13.5	40.7	60.5	25.8	29.9
Accuracy↑	CSN-IR50-Stage1	93.2	92.7	88.2	90.4	93.1	92.3	94.8	91.5	92.0
	CSN-IR50-Stage2	93.4	95.0	89.9	90.5	94.1	93.3	96.3	94.4	93.3
	CSN-IR50-Stage3	92.3	93.7	90.7	90.2	90.6	92.4	96.3	93.4	92.5
	CSN-IR50-Stage4	96.9	96.9	90.4	91.7	94.6	93.5	96.9	92.8	94.2
	CSN-IR50-FC	69.1	75.3	52.4	49.0	50.3	64.7	65.9	55.5	60.3
	CSN-IR50-Output	62.6	74.4	53.3	48.4	49.0	62.8	63.9	53.6	58.5

Table S9. The performance of different versions of CSN-IR50 on AU detection on the DISFA dataset.