frontiers in Computer Science

ORIGINAL RESEARCH
published: 29 April 2021
doi: 10.3389/fcomp.2021.636094

Systematic Evaluation of Design Choices for Deep Facial Action Coding Across Pose

Koichiro Niinuma[1]*, *Itir Onal Ertugrul*[2], *Jeffrey F. Cohn*[3] and *László A. Jeni*[4]

[1] Fujitsu Laboratories of America, Pittsburgh, PA, United States, [2] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, Netherlands, [3] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, United States, [4] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States

The performance of automated facial expression coding is improving steadily. Advances in deep learning techniques have been key to this success. While the advantage of modern deep learning techniques is clear, the contribution of critical design choices remains largely unknown, especially for facial action unit occurrence and intensity across pose. Using the The Facial Expression Recognition and Analysis 2017 (FERA 2017) database, which provides a common protocol to evaluate robustness to pose variation, we systematically evaluated design choices in pre-training, feature alignment, model size selection, and optimizer details. Informed by the findings, we developed an architecture that exceeds state-of-the-art on FERA 2017. The architecture achieved a 3.5% increase in F_1 score for occurrence detection and a 5.8% increase in Intraclass Correlation (ICC) for intensity estimation. To evaluate the generalizability of the architecture to unseen poses and new dataset domains, we performed experiments across pose in FERA 2017 and across domains in Denver Intensity of Spontaneous Facial Action (DISFA) and the UNBC Pain Archive.

Keywords: action unit, facial expression coding, design choice in deep learning, AU intensity estimation, AU occurrence detection, cross-pose evaluation, cross-domain evaluation

OPEN ACCESS

Edited by:
Jun Miura,
Toyohashi University of Technology, Japan

Reviewed by:
Nobutaka Shimada,
Ritsumeikan University, Japan
Prarinya Siritanawan,
Japan Advanced Institute of Science and Technology, Japan

***Correspondence:**
Koichiro Niinuma
kniinuma@fujitsu.com

Specialty section:
This article was submitted to Computer Vision, a section of the journal Frontiers in Computer Science

Received: *30 November 2020*
Accepted: *24 March 2021*
Published: *29 April 2021*

Citation:
Niinuma K, Onal Ertugrul I, Cohn JF and Jeni LA (2021) Systematic Evaluation of Design Choices for Deep Facial Action Coding Across Pose. Front. Comput. Sci. 3:636094. doi: 10.3389/fcomp.2021.636094

1. INTRODUCTION

Emotion recognition technologies play an important role in human computer interaction systems. Face-to-face interactions between social robots and people are but one example (McColl et al., 2016; Cavallo et al., 2018). To recognize human emotion, facial action units (AUs) (Ekman et al., 2002) have been widely used, which correspond to discrete muscle contractions. Individually, or in combination, they can represent nearly all possible facial expressions.

In the last-half decade, automated facial affect recognition (AFAR) systems have made major advances in detection of the occurrence and intensity of facial actions. Previous studies focused on relatively controlled laboratory settings. More recent studies emphasize on less-constrained and in-the-wild scenarios (Cohn and De la Torre, 2015; Li and Deng, 2018; Zhi et al., 2019). Because frontal face views occur commonly in less constrained settings, robustness to pose variation is essential. The Facial Expression Recognition and Analysis 2017 (FERA 2017) challenge provided the first common protocol to evaluate robustness to pose variation (Valstar et al., 2017). In FERA 2017, deep learning (DL)-based approaches achieved the best performance in sub-challenges (Tang et al., 2017) for occurrence detection (Zhou et al., 2017) and intensity estimation.

While the advantages of DL approaches are clear, little is known about critical design choices in crafting them. Most studies used *ad-hoc* or default parameters provided by the DL frameworks; however, they neglected to investigate the effect of different parameter settings on facial AU detection. Also, little is known about the relative contribution of different design choices in pre-training, feature alignment, model size, and optimizer details.

We are especially interested in design choices based on two scenarios. One is robustness to pose variation. Until recently, most systems were concerned with relatively frontal face views. With increased attention to less-constrained and in-the-wild contexts, it is critical for systems to be robust to pose variation in real-world settings where it is common. The other scenario is transfer to new dataset domains other than those in which they have been trained and tested. To meet the need for systems that are robust to new contexts, systems must perform well both in the domains from which they come and in the domains to which they may be applied. The evaluation of domain transfer in AU systems is relatively new (Cohn et al., 2019; Ertugrul et al., 2020).

To address questions in design choices, we systematically explored combinations of different components and their parameters in a DL pipeline. We investigated pre-training practices, image alignment for pre-processing, training set size, optimizer, and learning rate (LR). By utilizing the insights, we achieved state-of-the-art performance in both the occurrence detection and the intensity estimation sub-challenges of FERA 2017 (Valstar et al., 2017) and state-of-the art in cross-domain generalizability to the Denver Intensity of Spontaneous Facial Action (DISFA) dataset (Mavadati et al., 2013). We also are the first to report cross-domain generalizability to UNBC Pain Archive (Lucey et al., 2011). To reveal which facial regions our architecture responds to in detecting specific AUs at specific poses, we visualized occlusion sensitivity maps.

The study of Niinuma et al. (2019) was an earlier version of the current study. In the present study, we evaluated an additional DL architecture (ResNet50), performed cross-domain evaluation with an additional dataset (UNBC Pain), evaluated cross-pose generalizability, and visualized occlusion sensitivity maps.

2. RELATED WORK

Numerous approaches have been proposed for AU analysis (Cohn and De la Torre, 2015; Corneanu et al., 2016; Martinez et al., 2017; Li and Deng, 2018; Zhi et al., 2019). Most of these approaches had relatively frontal face orientation. Where moderate to large non-frontal pose has been considered (Kumano et al., 2009; Taheri et al., 2011; Jeni et al., 2012; Rudovic et al., 2013; Tősér et al., 2016), the lack of a common protocol has undermined comparisons.

The FERA 2017 challenge (Valstar et al., 2017) was the first to provide a common protocol to compare approaches for detection of AU occurrence and AU intensity robust to pose variation. FERA 2017 provided synthesized face images from BP4D (Zhang et al., 2014) with nine head poses, as shown in **Figure 1**. To generate the synthesized images, 3D models were rotated by −40, −20, and 0° pitch and −40, 0, and 40° yaw from frontal pose. The training set was based on the BP4D database (Zhang et al., 2014), which included digital videos of 41 participants. The development and test sets were derived from BP4D+ (Zhang et al., 2016) and included digital videos of 20 and 30 participants, respectively. FERA 2017 presented two sub-challenges: occurrence detection and intensity estimation, with 10 AUs labeled for the former and 7 AUs labeled for the latter.

For FERA 2017, the participants proposed a wide range of methods (Amirian et al., 2017; Batista et al., 2017; He et al., 2017; Li et al., 2017; Tang et al., 2017; Valstar et al., 2017; Zhou et al., 2017). **Table 1** compares them with each other and with two more recent studies from Ertugrul et al. (2018) and Li et al. (2018). F_1 score and Intraclass Correlation (ICC) were used to evaluate, performance for occurrence detection and intensity estimation, respectively.

Several comparisons are noteworthy. While detailed face alignment using facial landmarks was used for shallow approaches, simple face alignment using face position or resized images more often sufficed for DL approaches. As for architecture, DL performed better than shallow approaches, and DL approaches with pre-trained models performed better than ones without pre-trained models. For both sub-challenges, the methods showing the best performance (Tang et al., 2017) for occurrence detection and for intensity estimation (Zhou et al., 2017) used DL with a pre-trained model. As for training set size, each method used different numbers of training images. Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD) were popular choices for optimizer, and their LR varied between 10^{-3} and 10^{-4}.

According to the comparison of the existing methods, the effectiveness of DL approaches, especially the ones using pre-trained models, is indicated for this task, but every approach used a different fixed configuration, and the key parameters are unknown. The aim of this study is to investigate the key parameters for both AU occurrence detection and intensity estimation for this task and discover the optimal configuration.

3. METHODS

The main goal of this study is to investigate the effect of the different components and parameters and to provide best practices that researchers can use for training DL methods for automatic facial expression analysis. **Figure 1** shows an outline of our experimental design. We systematically varied parameters and design choices in this pipeline (key elements are denoted in blue color in **Figure 1**).

3.1. FACS

The Facial Action Coding System (FACS) (Ekman et al., 2002) is an anatomically-based system annotating nearly all possible facial movements. FACS examines the shape and appearance changes produced by the muscles and soft tissues of the face. Each muscle movement constitutes an AU. We investigated both AU occurrence detection and AU intensity estimation. In the FERA 2017 dataset, 10 AUs (AU1, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, and AU23) were evaluated for occurrence

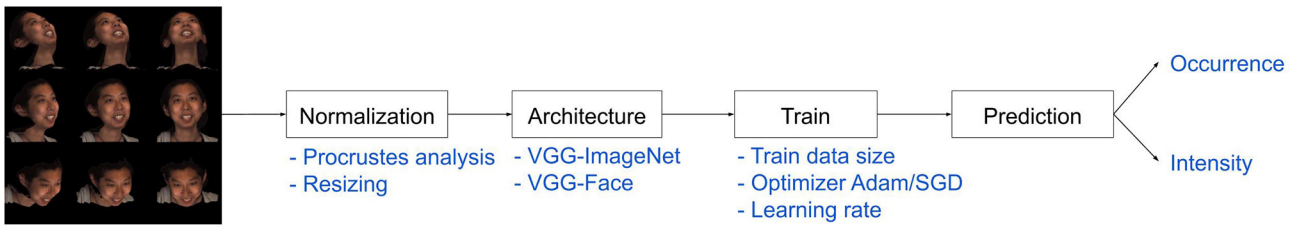

FIGURE 1 | An overview of the experimental design. Blue color denotes design choices and parameters for systematic evaluation.

TABLE 1 | An overview of the design choices from studies reporting performance on the FERA 2017 sub-challenges.

	Design choice						Performance	
	Normalization	**Architecture**	**Pre-training**	**Training set size per model**	**Optimizer**	**Learning rate**	**Occurrence performance (F_1 score)**	**Intensity performance (ICC)**
Valstar et al. (2017)	Facial landmarks	Shallow	n/a	n/r	n/a	n/a	0.452	0.217
Li et al. (2017)	Facial landmarks	Hybrid	VGG-Face[a]	26,582	n/a	n/a	0.498	n/a
Batista et al. (2017)	Face position	Deep	None[b]	1,321,472	Adam	10^{-3}	0.506	0.399
He et al. (2017)	Resizing[c]	Hybrid	None	146,847	n/r	n/r	0.507	n/a
Tang et al. (2017)	Face position[d]	Deep	VGG-Face	440,541 $+\alpha$[e]	SGD	10^{-3}	**0.574**	n/a
Ertugrul et al. (2018)	Face position	Deep	None	1,321,623	Adam	10^{-3}	0.525	n/a
Li et al. (2018)	Facial landmarks	Deep	ImageNet-VGG-VD19	260,000 $+\alpha$[f]	SGD	10^{-4}	n/a[g]	n/a
Amirian et al. (2017)	Facial landmarks	Shallow	n/a	n/r	n/a	n/a	n/a	0.295
Zhou et al. (2017)	Resizing	Deep	ImageNet-VGG-VD16	54,000	SGD	10^{-4}	n/a	**0.446**

For occurrence detection, F_1 scores are reported. For intensity detection, Intraclass Correlation coefficients (ICC) are reported. N/A denotes not applicable; N/R denotes not reported. Best scores are denoted in bold.

[a] *A VGG pre-trained model was used to extract features but was not used for classification.*

[b] *A VGG pre-trained model was used to detect faces but was not used for classification.*

[c] *Face detection was used for train and validation partition but was not for test partition.*

[d] *Face position was not directly used, but facial images were cropped by using morphology operations including binary segmentation, connected components labeling and region boundaries extraction.*

[e] *After down sampling to 440,541 images, Tang et al. increased the number of samples to balance positive and negative samples.*

[f] *Li et al. increased the number of samples to balance positive and negative samples.*

[g] *In their paper Li et al. reported F1 scores only on validation partition.*

detection, and 7 AUs (AU1, AU4, AU6, AU10, AU12, AU14, and AU17) were evaluated for intensity estimation. AU1, AU4, AU6, and AU7 are upper face AUs, and represent inner brow raiser, brow lowerer, cheek raiser, and lid tightener, respectively. AU10, AU12, AU14, AU15, AU17, and AU23 are lower face AUs, and represent upper lip raiser, lip corner puller, dimpler, lip corner depressor, chin raiser, and lip tightener, respectively (Cohn and De la Torre, 2015).

3.2. Architecture

Since the objective of this study is to investigate components that were commonly used by existing methods, we examined Visual Geometry Group (VGG) architectures. **Table 1** shows VGG pre-trained models that were widely used as architectures. To examine the impact of alternative DL architectures, we also conducted the experiments using the ResNet50 pre-trained model in section 4.11.

For VGG architectures, we selected two pre-trained models: VGG-ImageNet and VGG-Face. While VGG-ImageNet is a model that was trained on ImageNet for image classification (Simonyan and Zisserman, 2015), VGG-Face is a model that was trained on the face dataset for face recognition (Parkhi et al., 2015).

3.3. Baseline Configuration

In each experiment, we explored the effect of optimizer choice and parametric variation of key parameters. The experimental setup has five parameters (normalization, architecture, train set size, optimizer, and LR) and two tasks (occurrence detection and intensity estimation) in total. To vary all parameters would have resulted in 320 possible permutations. In consideration of computational cost and limits on how much could be visualized, we varied two parameters at a time and chose the top 50 permutations that we believed would be of most interest to developers of AFAR systems.

The baseline configuration used Procrustes analysis for face alignment and the VGG16 network trained on ImageNet. For optimizers, we compared Adam and SGD, with default learning rates of 5×10^{-5} and 5×10^{-3}, respectively. We fine-tuned the network from the third convolutional layer using 5,000 images for each pose and AU. The dropout rate was 0.5 throughout the experiments.

4. EXPERIMENTS

4.1. Normalization

We evaluated two methods for image normalization. In the first method, we applied Procrustes analysis (Gower, 1975) to the face shapes defined by the landmarks to estimate similarity normalized shapes. In the second method, we resized the images to the receptive field of the deep network.

Similarity normalization between source and template shapes using eye locations is a popular choice in the literature. One shortcoming of this approach is that the alignment error increases for landmarks farther away from the eye region. This artifact is more prominent under moderate-to-large head pose variations. To alleviate this problem, we used all 68 landmarks provided by the dlib face tracker (King, 2009) to calculate a Procrustes transformation between the predicted shape and a frontal looking template. We chose the size of the template to cover a bounding box of 224 × 224 pixels, which corresponds to the receptive field of the VGG network.

As for the second option, we resized each input image from the dataset to 224 × 224 pixel size to match the receptive field of the VGG network.

Figure 2A shows the F_1 scores and ICC averages for all nine poses for each AU. The left figures show results for Adam optimizer, and the right figures show results for SGD optimizer. The results indicate that the performance with Procrustes analysis is slightly better than the one with resizing, but the difference is small, only 1%. One possible explanation for this is that the network has enough capacity to learn all the nine different poses present in the training set. Another study indicates that a form of normalization is often helpful when classifiers are evaluated on poses different from the ones it was trained on (Ertugrul et al., 2018).

4.2. Pre-trained Architecture

Training deep models from scratch is time-consuming, and the amount of training data at hand may impede good performance. One popular solution is to select a model that was trained on large scale benchmark datasets (source domain) and fine-tune it on the data of our interest (target domain). Although this practice is effective, it is relatively neglected how the type of data in the source domain influence the performance of fine-tuning in the target domain.

To explore this question, we selected two models that were trained on very different domains: VGG-16 trained on ImageNet (Simonyan and Zisserman, 2015) and VGG-Face (Parkhi et al., 2015). We replaced the final layers of each networks with a 2-length one-hot representation for AU occurrence detection and with a 6-length one-hot representation for the intensity estimation task. In both cases, we trained separate models for each AU, resulting in 10 and 7 models for AU occurrence detection and AU intensity estimation, respectively. We fine-tuned the models for 10 epochs, validated their performance on the validation partition, and then reported their results on the subject-independent test partition. We used a PyTorch implementation for all of the models.

Figure 2B shows that models pre-trained on ImageNet show better performance than the VGG-Face ones. VGG-Face was trained on face images for identification, while ImageNet includes many non-face images for image classification. One possible explanation is that VGG-Face learned to actively ignore facial expression in order to recognize the face. In this case, a generic image representation is more suitable for the task.

4.3. Training Set Size

Recently, multi-label stratified sampling was found advantageous over naive sampling strategies for AU detection (Chu et al., 2019). In this experiment, we employed this strategy and investigated the effect of different training set sizes on the performance. We down-sampled the majority class and up-sampled the minority class to build a stratified training set. We used this procedure for each pose and each AU. For example, in the case of AU occurrence detection, a 5,000 training set size indicate that 5,000 frames where the AU is present and 5,000 frames where the AU is not present were randomly selected for each pose and for each AU, resulting in 90,000 images in total (=5,000 images × 2 classes × 9 poses).

We repeated the same stratifying procedure with the six ordinal classes of the intensity sub-challenge. In this case, a 5,000 training set size means that 5,000 images were randomly selected from the six classes (not present, and A to E levels) for each pose and for each AU, resulting in 270,000 images in total (=5,000 images × 6 classes × 9 poses).

Figure 2C shows results as the function of different training set size. The training set size have minor influence on the performance: scores peaked at 5,000 images after that performance plateaued.

4.4. Optimizer and LR

In this experiment, we investigated the impact of different optimizers and LRs on the performance. We varied the LRs, but other optimizer parameters were set to the default values used in PyTorch: betas = (0.9, 0.999) without weight decay for Adam and no momentum, no dampening, no weight decay, and no Nesterov acceleration for SGD.

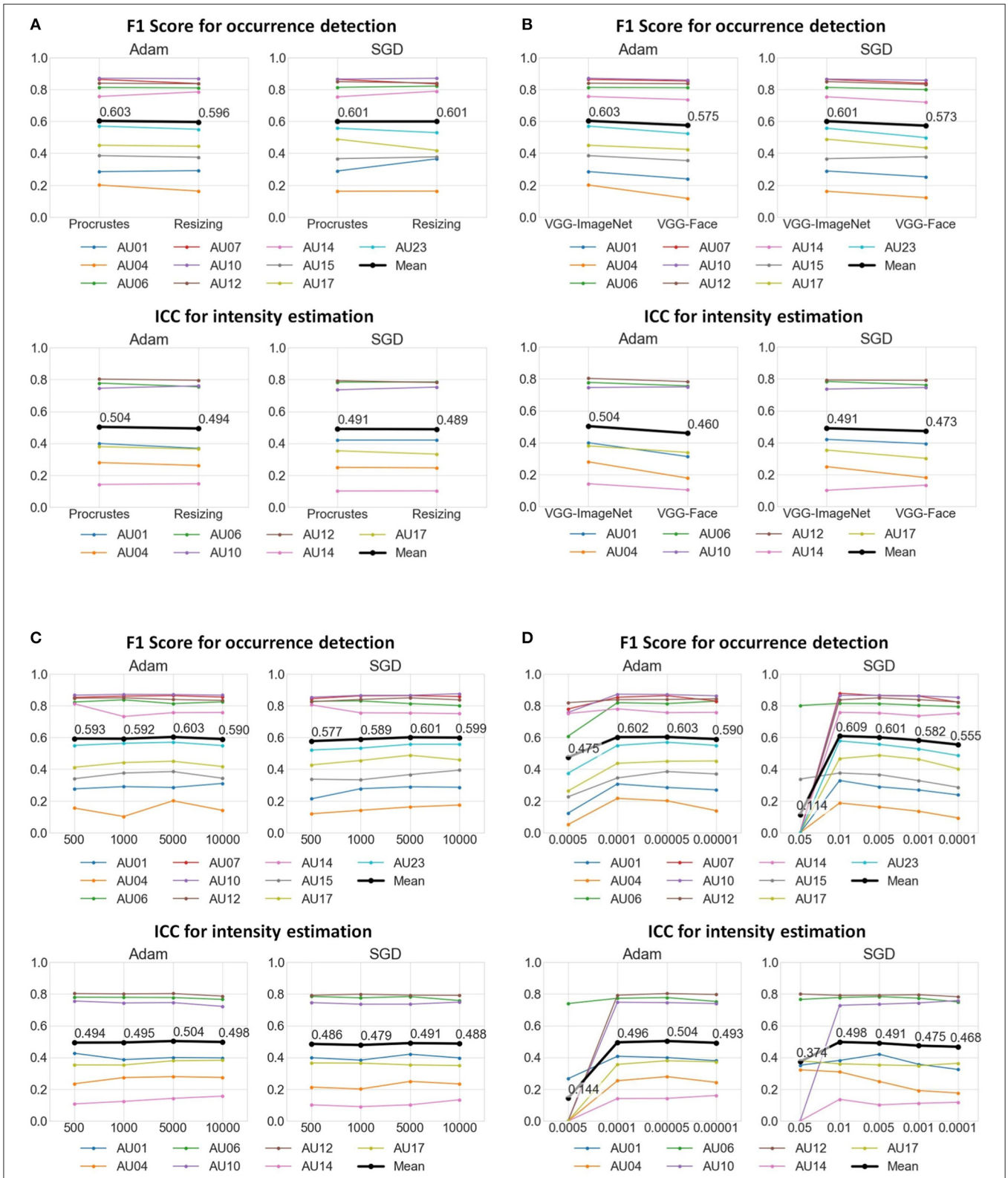

FIGURE 2 | Results on The Facial Expression Recognition and Analysis 2017 (FERA 2017). Test partition with **(A)** two normalization methods, **(B)** two pre-trained architecture, **(C)** different number of train set size, and **(D)** learning rates and choice of optimizers.

TABLE 2 | F1 scores for occurrence detection results on FERA 2017 Test partition.

	Valstar et al. (2017)	Li et al. (2017)	Batista et al. (2017)	He et al. (2017)	Ertugrul et al. (2018)	Tang et al. (2017)	Our model
AU01	0.147	0.215	0.219	0.198	0.196	0.263	**0.329**
AU04	0.044	0.044	0.056	0.043	0.067	0.118	**0.187**
AU06	0.630	0.755	0.785	0.747	0.766	0.776	**0.814**
AU07	0.755	0.805	0.816	0.784	0.791	0.808	**0.878**
AU10	0.758	0.810	0.838	0.816	0.840	**0.865**	**0.865**
AU12	0.687	0.753	0.780	0.809	0.819	**0.843**	0.837
AU14	0.668	0.750	0.747	0.691	**0.764**	0.757	0.758
AU15	0.220	0.208	0.145	0.208	0.247	0.362	**0.376**
AU17	0.274	0.286	0.388	0.398	0.349	0.424	**0.467**
AU23	0.342	0.356	0.286	0.374	0.413	0.519	**0.578**
Mean	0.452	0.498	0.506	0.507	0.525	0.574	**0.609**

The best results are shown in bold.

TABLE 3 | ICC for intensity estimation on FERA 2017 Test partition.

	Valstar et al. (2017)	Amirian et al. (2017)	Batista et al. (2017)	Zhou et al. (2017)	Our model
AU01	0.035	0.169	0.228	0.307	**0.400**
AU04	−0.004	0.021	0.057	0.147	**0.280**
AU06	0.461	0.509	0.702	0.671	**0.778**
AU10	0.451	0.590	0.710	0.735	**0.746**
AU12	0.518	0.615	0.732	0.793	**0.803**
AU14	0.037	−0.027	0.104	**0.147**	0.143
AU17	0.020	0.190	0.260	0.319	**0.380**
Mean	0.217	0.295	0.399	0.446	**0.504**

The best results are shown in bold.

Figure 2D shows that the optimal LR depends on the choice of optimizer. For Adam, LR $= 5 \times 10^{-5}$ gave the best results, and for SGD, LR $= 0.01$ reached the best performance for both occurrence detection and intensity estimation. In addition, we can see that the performance differences between Adam and SGD are negligible if one uses the optimal learning rates for each optimizer.

It is worth noting that Zhou et al. (2017) used SGD with LR=10^{-4} for the AU intensity estimation task. The results indicate that using Adam optimizer or SGD optimizer with larger LR could have improved their performance. Tang et al. (2017) used SGD with LR = 10^{-3}, but they also applied momentum. Additional experiments revealed that, when momentum is used for SGD, smaller learning rate is preferable for optimal performance. More specifically, when we used the same parameters as Tang et al. (2017) reported for SGD (momentum $= 0.9$, weight decay $= 0.02$), F_1 score peaked at 0.596 using LR $= 10^{-4}$. Their LR was close to optimal, though SGD without momentum further improves F_1 score to 0.609 with LR $= 0.01$.

We note that, when the LR was set to a large value, some models did not converge and predicted the majority class for all samples. Under this rare condition, ICC converges to zeros, but this should not be interpreted as chance performance. As variation in predicted intensity values reduces, the ICC metric loses predictive power.

4.5. Comparison With Existing Methods

We compare our method with the state-of-the-art on both the AU occurrence detection (**Table 2**) and the AU intensity estimation (**Table 3**) sub-challenges from FERA 2017. The final parameters of the models are nearly identical for the two tasks: we used face alignment with Procrustes analysis as a pre-processing step, and we fine-tuned ImageNet pre-trained VGG16 model on stratified sets consisting of 5,000 samples per each class, pose, and AU. For AU occurrence detection, SGD with LR $= 0.01$ gave the best result ($F_1 = 0.609$), while for AU intensity estimation, Adam with LR $= 5 \times 10^{-5}$ reached the best performance (ICC $= 0.504$). These scores outperform other state-of-the-art methods.

We noted a few key differences that contributed to this achievement. The main difference with Tang et al. (2017) is that they used VGG-Face pre-trained model while we used ImageNet pre-trained model. Zhou et al. (2017) used SGD with small LR while the combination of our optimizer and learning rate is optimal. While Li et al. (2018) evaluated their method for AU occurrence detection using the FERA 2017 dataset, they reported performance only on the validation partition. Their best F_1 score (0.522) is 9% lower than ours (0.611) on the validation partition.

4.6. Effect of Head Pose on Performance

To understand the effect of head pose on classifier performance, we complied the performance scores into a tabular form, as shown in **Tables 4**, **5**.

TABLE 4 | F1 scores and Accuracy of our model for occurrence detection under nine facial poses on FERA 2017 Test partition.

Pose	1	2	3	4	5	6	7	8	9	Mean
						F1 score				
AU01	0.358	0.292	0.272	0.353	0.346	0.366	0.312	0.314	0.345	0.329
AU04	0.247	0.208	0.129	0.254	0.226	0.217	0.131	0.135	0.133	0.187
AU06	0.808	0.803	0.788	0.828	0.830	0.811	0.829	0.821	0.811	0.814
AU07	0.887	0.886	0.864	0.877	0.883	0.885	0.871	0.875	0.878	0.878
AU10	0.859	0.867	0.864	0.868	0.872	0.868	0.872	0.870	0.841	0.865
AU12	0.821	0.830	0.850	0.830	0.843	0.850	0.833	0.847	0.828	0.837
AU14	0.756	0.737	0.742	0.758	0.776	0.771	0.787	0.759	0.735	0.758
AU15	0.422	0.419	0.369	0.408	0.379	0.357	0.357	0.340	0.336	0.376
AU17	0.453	0.485	0.493	0.461	0.492	0.486	0.482	0.430	0.416	0.466
AU23	0.568	0.588	0.577	0.611	0.597	0.588	0.557	0.568	0.545	0.578
Mean	0.618	0.612	0.595	0.625	0.624	0.620	0.603	0.596	0.587	0.609
						Accuracy				
AU01	0.855	0.835	0.862	0.834	0.850	0.871	0.811	0.831	0.844	0.844
AU04	0.961	0.949	0.900	0.944	0.926	0.919	0.944	0.937	0.907	0.932
AU06	0.828	0.828	0.807	0.847	0.847	0.838	0.838	0.819	0.805	0.829
AU07	0.850	0.848	0.813	0.837	0.841	0.844	0.828	0.832	0.836	0.837
AU10	0.830	0.836	0.834	0.834	0.838	0.835	0.833	0.829	0.798	0.830
AU12	0.793	0.807	0.834	0.808	0.821	0.832	0.817	0.831	0.802	0.816
AU14	0.708	0.690	0.686	0.709	0.722	0.718	0.733	0.694	0.676	0.704
AU15	0.814	0.802	0.785	0.795	0.780	0.792	0.782	0.753	0.711	0.779
AU17	0.744	0.790	0.787	0.764	0.789	0.794	0.778	0.754	0.739	0.771
AU23	0.769	0.782	0.764	0.803	0.782	0.775	0.786	0.790	0.757	0.779
Mean	0.815	0.817	0.807	0.818	0.820	0.822	0.815	0.807	0.788	0.812

TABLE 5 | ICC of our model for intensity estimation under nine facial poses on FERA 2017 Test partition.

Pose	1	2	3	4	5	6	7	8	9	Mean
AU01	0.441	0.449	0.400	0.403	0.436	0.433	0.353	0.354	0.333	0.400
AU04	0.305	0.278	0.250	0.294	0.333	0.281	0.317	0.244	0.216	0.280
AU06	0.779	0.786	0.787	0.787	0.788	0.786	0.762	0.776	0.754	0.778
AU10	0.763	0.750	0.738	0.759	0.763	0.768	0.734	0.722	0.720	0.746
AU12	0.799	0.812	0.815	0.809	0.813	0.812	0.795	0.797	0.777	0.803
AU14	0.144	0.161	0.162	0.137	0.143	0.153	0.124	0.141	0.126	0.143
AU17	0.393	0.396	0.403	0.394	0.388	0.382	0.383	0.359	0.319	0.380
Mean	0.518	0.519	0.508	0.512	0.523	0.516	0.495	0.485	0.464	0.504

For each pose and AU, the tables show F_1 score and Accuracy for occurrence detection and ICC for intensity estimation. In the experiments, we used the same CNN models reported in section 4.5. We can see the effect of rotations in **Tables 4**, **5**. As for the pitch rotations, the performance with 0° pitch poses (Pose 4, 5, and 6) show better results than the others. As for yaw rotations, the performance scores are comparable for all poses.

4.7. Cross-Domain Evaluation

Differences in illumination, cameras, orientation of the face, quality, and diversity of the training data influence predictive performance between domains. To evaluate the generalizability of the method to unseen conditions, we reported performance on the DISFA (Mavadati et al., 2013) and UNBC McMaster Pain (Lucey et al., 2011) datasets.

These datasets were annotated with AU intensity labels. To create binary AU occurrences, we thresholded the 6-points intensity values at A-level (A-level or higher means the AU is present). We evaluated both occurrence detection and intensity estimation performance of our system. In these experiments, no fine tuning was performed on the target domain.

Figure 3A shows the F_1 scores with two normalization methods, Procrustes analysis and resizing. **Figure 3B** shows the F_1 scores with two pre-trained architectures, VGG-ImageNet and VGG-Face. In these experiments, we used the same configuration with Adam optimizer in sections 4.1 and 4.2, respectively. We used the built-in face detector in dlib (King, 2009) to detect

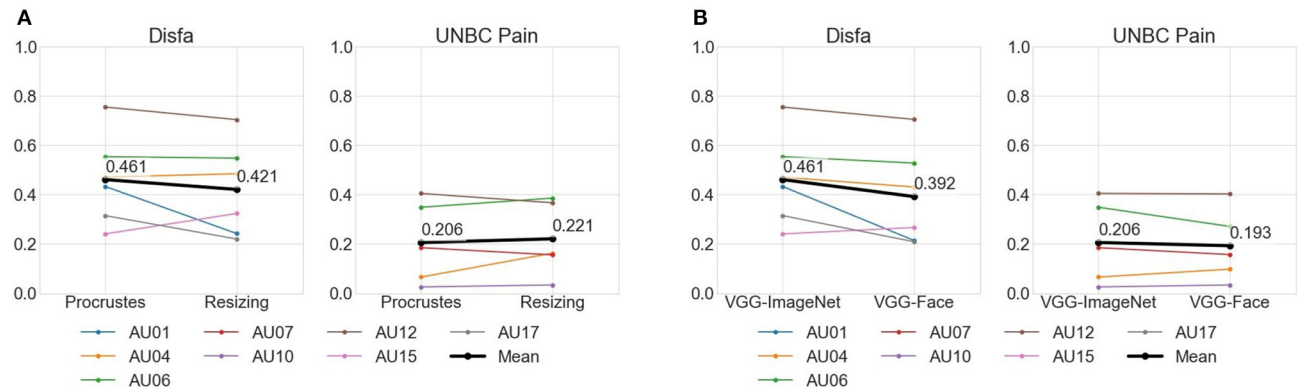

FIGURE 3 | F_1 scores for occurrence detection on the Denver intensity of spontaneous facial action (DISFA) and UNBC Pain with **(A)** two normalization methods and **(B)** two pre-trained architectures.

the face before applying Procrustes analysis. As for resizing, we extend the boxes of detected face positions by 30% to include whole faces and then crop and resize the boxes to 224 × 224 size images. For DISFA, we found that Procrustes analysis with VGG-ImageNet have better performance. For UNBC Pain Archive, the findings are in same direction but small.

Tables 6, **7** show the results from our model on both tasks. In these experiments, we used two types of models: (1) All poses: the previously trained CNN models reported in section 4.5, and (2) Pose 6 only: models trained on images only with Pose 6, which is equivalent to BP4D. **Table 6** includes the comparison with cross-domain methods for occurrence detection on DISFA. Both Baltrušaitis et al. (2015) and Ghosh et al. (2015) used BP4D to train their model and thresholded AU intensity values at A-level to create binary events. Our models were also trained using BP4D because the train set for FERA 2017 is synthesized from BP4D. Pose 6 in FERA 2017 is the same as the pose shown in BP4D. To train the models for Pose 6 only, the same number of images as All poses are used. More specifically, 45,000 frames were randomly selected per class per AU, resulting in 90,000 images in total for each AU. As we discussed in section 4.3, we down-sampled the majority class and up-sampled the minority class. We also report Accuracy and 2AFC scores that Ghosh et al. (2015) used.

When All poses model and Pose 6 only model were compared, both models showed that almost the same performance for Accuracy, F1 score, and AUC though All poses shows slightly better results than Pose 6 only for 2AFC. The results look reasonable because most images in DISFA is frontal or near frontal. In comparison with Ghosh et al. (2015), our models outperform their method in both metrics. Baltrušaitis et al. (2015) report cross-domain scores only for two AUs (AU12 and AU17). Our models show better performance except for AU12 on Pose 6 only. These results show the robustness of our model for cross-domain situation. To the best our knowledge, there are no other methods that perform cross-domain evaluation on these datasets. **Table 7** depicts the results of our methods.

It is worth mentioning that some differences on UNBC Pain may cause the low F_1 scores. The base rates on UNBC Pain is small (DISFA: 13.3%, and UNBC Pain: 7.2%) and the image size of UNBC Pain (320 × 240 or 352 × 240) is also smaller than the other two datasets (FERA2017: 1,024 × 1,024, and DISFA: 1,024 × 768). In addition, in UNBC Pain, facial expressions are mostly associated with pain, and the correlation among AUs differs from that of FERA2017 and DISFA. **Tables 6**, **7** also show AUC.

4.8. Cross-Pose Evaluation

We also performed cross-pose experiments to evaluate the generalization of our method to unseen poses. We report the results of two types of experiments: (1) We trained the architecture using eight of the nine poses of training set and tested it with the remaining pose of test set (**Figure 4**), (2) We trained the architecture using one pose of training set and tested it with nine poses of test set (**Figure 5**). The baseline configuration with Adam optimizer is used for cross-pose experiments.

Figure 4 shows that the differences between the models trained with eight poses and those trained with nine poses. The horizontal axis represents the pose that was excluded from train set and used as test set. The value is zero if the performance between two models are the same, and the value is >0 if the performance with eight poses is better than the one obtained with nine poses. By training the models with all of the nine poses, the best performance since the model learns information about all poses is expected. With eight-pose experiments, we can see that, even if the test pose is excluded from the training set, our model performs similarly to the one in which the test pose is included in the training set. The results indicate that our model performs reasonably well on the unseen poses.

As for **Figure 4**, we provide a more detailed analysis. Accuracy for AU4 is higher for poses 1 and 2. No difference, however, is found for AU4 intensity. Given that the occlusion sensitivity maps for AU4 appear similar across poses, the difference for poses 1 and 2 in occurrence may be due to noise. AU15, on the other hand, showed the decreased accuracy for poses 7, 8, and 9. This effect would be expected. AU 15 results in localized, small movement, and appearance change below the lip corners.

TABLE 6 | Comparison of cross-domain performance to DISFA dataset for occurrence detection.

	Accuracy			2AFC		
	Our (all poses)	**Our (pose 6 only)**	**Ghosh et al. (2015)**	**Our (all poses)**	**Our (pose 6 only)**	**Ghosh et al. (2015)**
AU01	**0.932**	0.922	0.838	0.714	**0.720**	0.660
AU04	0.806	**0.874**	0.833	0.723	**0.759**	0.740
AU06	**0.860**	0.811	0.703	0.758	0.744	**0.870**
AU12	**0.859**	0.812	0.624	0.859	0.823	**0.873**
AU15	0.823	**0.943**	0.752	**0.671**	0.618	0.617
AU17	**0.738**	0.657	0.689	**0.742**	0.720	0.585
Mean	**0.836**	**0.836**	0.740	**0.745**	0.730	0.724

	F1			AUC	
	Our (all poses)	**Our (pose 6 only)**	**Baltrušaitis et al. (2015)**	**Our (all poses)**	**Our (pose 6 only)**
AU01	**0.475**	0.456	–	0.787	**0.819**
AU04	0.531	**0.629**	–	0.809	**0.886**
AU06	**0.567**	0.506	–	**0.867**	0.837
AU12	**0.742**	0.679	0.700	**0.934**	0.902
AU15	0.253	**0.345**	–	0.761	**0.795**
AU17	**0.361**	0.316	0.260	**0.823**	0.784
Mean	**0.488**	**0.488**	–	0.830	**0.837**

The best results are shown in bold.

TABLE 7 | Cross-domain performance to DISFA dataset for intensity estimation and UNBC Pain dataset for occurrence detection and intensity estimation.

DISFA		UNBC pain				
Intensity	**Estimation**	**Occurrence**	**Detection**		**Intensity**	**Estimation**
	ICC		**F1**	**AUC**		**ICC**
AU01	0.533	AU01	–	–	AU01	–
AU04	0.560	AU04	0.195	0.863	AU04	0.152
AU06	0.451	AU06	0.249	0.720	AU06	0.262
AU07	–	AU07	0.188	0.784	AU07	–
AU10	–	AU10	0.028	0.743	AU10	0.018
AU12	0.747	AU12	0.405	0.785	AU12	0.388
AU17	0.319	AU17	–	–	AU17	–
Mean	0.522	Mean	0.213	0.779	Mean	0.205

When the face is viewed from above (poses 7, 8, and 9), the target region is occluded. As for AU23, there was decreased accuracy for pose 9. Lip-corner tightening may be more difficult to perceive when viewed from above, but that was not found for two of the three extreme poses. Thus, variation for this AU occurrence is difficult to interpret. Unfortunately, AU intensity is not available for comparison.

Figure 5 shows the results of the second experiment. Each cell of a 3 × 3 matrix shows the performance of each pose. Performance at a cell of the grid corresponds to the pose at the same cell given in **Figure 1** The blue rectangle is a pose that was used to train a model. For example, for a model trained with Pose 1, F_1 score is 0.604 when we test it with Pose 1 of the test set and 0.446 when we test it with Pose 9 of the test set. The figure shows that maximum results are obtained with within-pose. Smaller decreases are observed in the performance when the models are tested with the poses in the neighboring cells. The performance is largely decreased when we test a model with largely different poses.

4.9. Occlusion Sensitivity Maps

To discover key features for the classifier, we generated occlusion sensitivity maps (Zeiler and Fergus, 2014) for each pose and each AU. We used an occlusion patch of a 45 × 45 size with Gaussian random noise. We slid the patch over the original image of 224 × 224 size with a stride 15. For each AU each pose, we selected 100 images that contained the specific AU and 100 images that did not contain it. We tested the 200 images for each AU and each pose and obtained accuracy values. **Figure 6** shows the maps, where the darker red colors represent lower accuracy values. Significant regions are the ones colored with red because their occlusion causes the largest decrease in the accuracy.

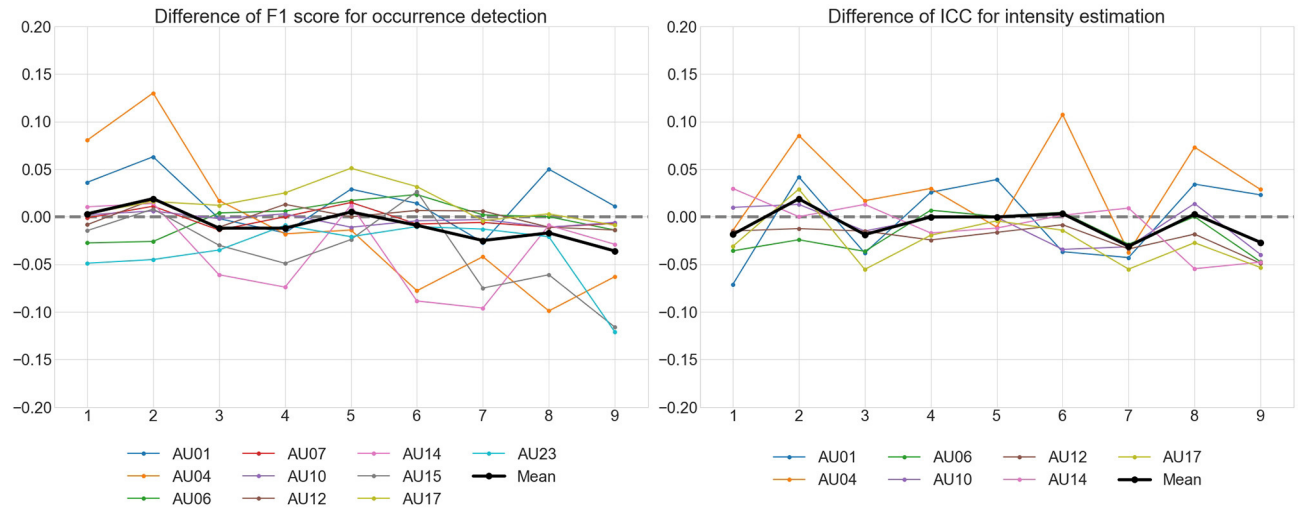

FIGURE 4 | Performance difference between models trained with eight poses and with nine poses. Horizontal axis shows each pose.

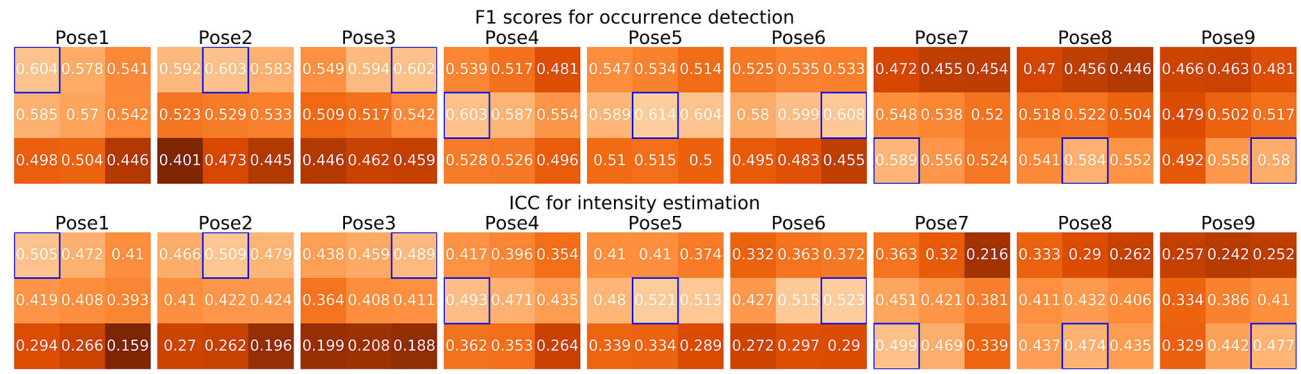

FIGURE 5 | F_1 scores and Intraclass Correlation coefficients (ICC) for models using one pose of training set and with nine poses of test set. Only mean values are reported.

As can be seen in **Figure 6**, for most of the AUs, the significant regions are localized at the regions where each AU is observed (e.g., around eyes, eyebrows, and forehead for AU1 and AU4, and around the mouth and chin for AU15 and AU17). The results indicate that the models learn of where to look at on the input to detect the specific AU correctly. Note that significant regions in **Figure 6** are off to the left side even for frontal faces. This seems to be reasonable because the pitch and yaw rotations of images in FERA2017 datasets is in one direction, as shown in **Figure 1**. If any occlusion does not cause large decrease in accuracy, the map does not include dark red colors. For example, we see weak activation on the heatmap for the AU12 frontal face. The map indicates that, even if a large part of mouth is occluded, our model can detect AU12 by using the other part of the face.

4.10. Saliency Maps

We generated saliency maps using basic backpropagation (Simonyan et al., 2014) to compare the learned features. For each AU each pose, we selected 100 images that contained the specific AU and 100 images that did not contain it. We then obtained a mean image of saliency maps from the 200 images. **Figure 7** shows the results of this experiment. Brighter areas are more important for the classifier to detect the related AUs.

The important regions are expected to be localized at the regions where each action unit is observed. Like the occlusion sensitivity map, the saliency map aims to find important regions to detect, but there are differences in their methodology and in the way they define what is important. The occlusion sensitivity map follows a perturbation-based (forward propagation) approach. Perturbed (occluded) inputs are forwarded through the network, and its effect on the output prediction is investigated. Contrary to the occlusion sensitivity map, saliency map is a gradient-based (back propagation) approach. The idea behind saliency map is to compute the gradient of the output category with respect to the input image pixels. This shows the amount of change in the output when a pixel value is slightly changed. **Figure 7** shows that the important regions are well-localized for both VGG-ImageNet and VGG-Face. However, in comparison with VGG-ImageNet, the regions

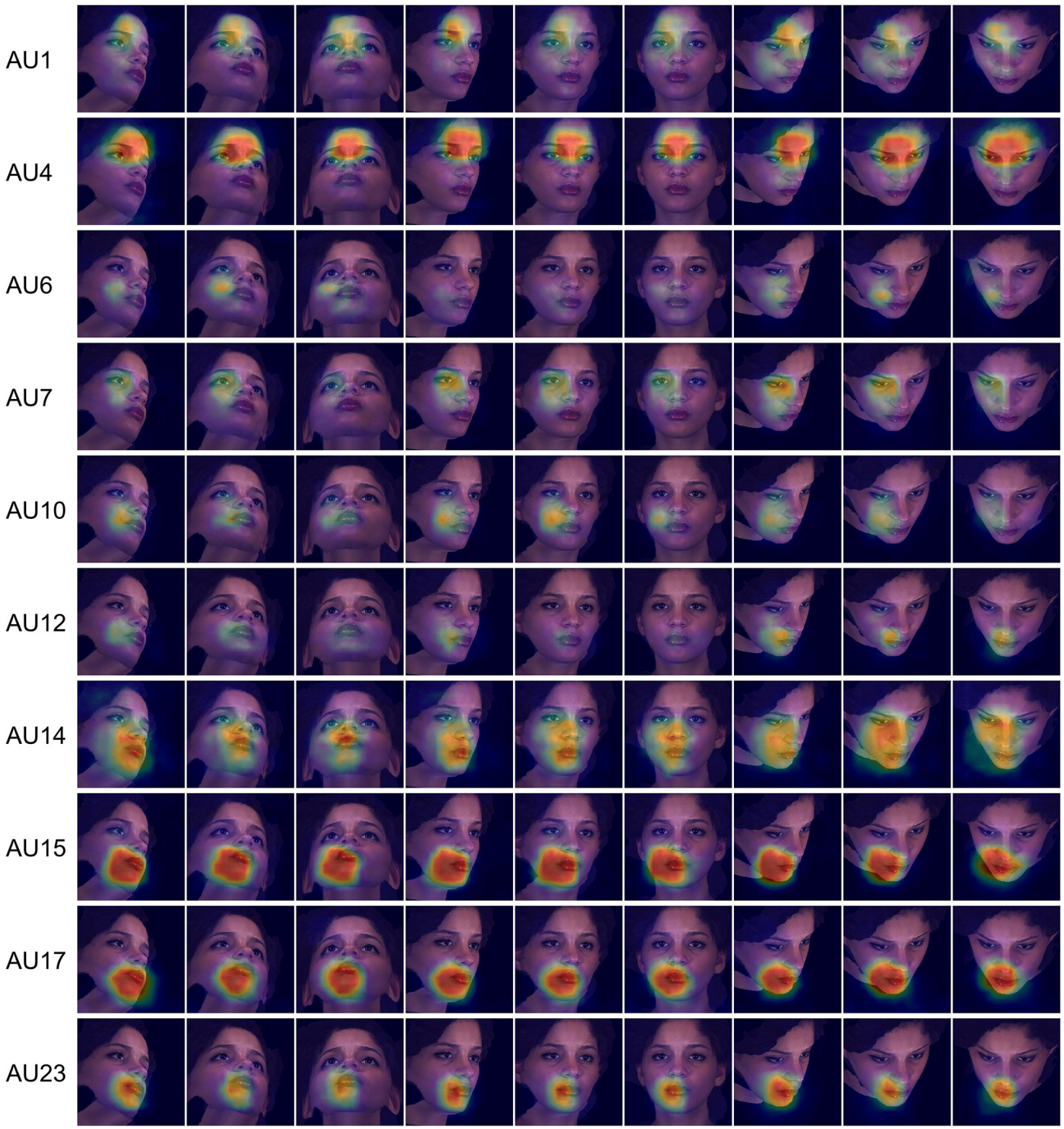

FIGURE 6 | Occlusion sensitivity maps for each pose each AU. Models trained with our baseline configuration are used.

of VGG-Face are wider and include more areas that are not related to each AU. The results indicate that the important regions are better localized for the VGG-ImageNet compared to VGG-Face. This is consistent with the experimental results that show that VGG-ImageNet pre-trained models outperform VGG-Face pre-trained models. Note that the important regions are off to the left side, like occlusion sensitivity maps, as we discussed in section 4.9.

4.11. ResNet

To examine the impact of different DL architectures, we conducted the experiments using ResNet50 pre-trained on ImageNet. In this experiment, we fine-tuned the network from the first layer. For the other parameters, our baseline configuration was used. **Figure 8** shows the results of the experiment. ResNet50 (0.516) shows better performance than VGG16 (0.504) for intensity estimation, while VGG16

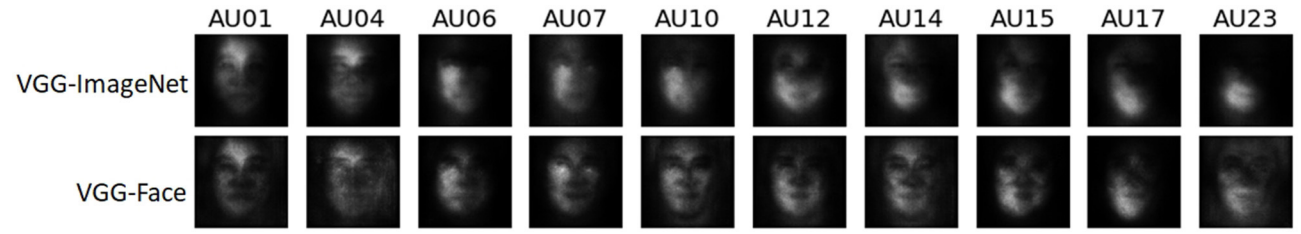

FIGURE 7 | Saliency maps extracted using basic backpropagation.

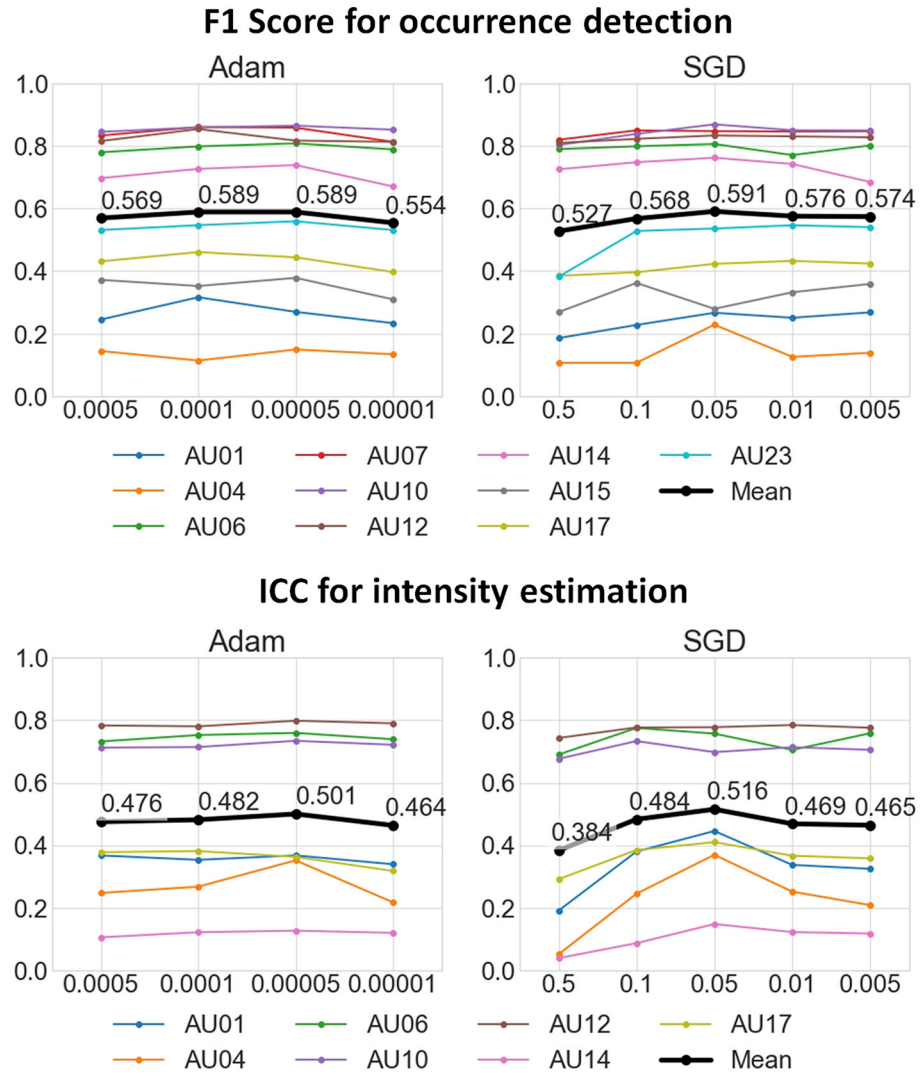

FIGURE 8 | Effect of learning rates and choice of optimizers for ResNet50 on the FERA2017 Test partition.

(0.609) shows better performance than ResNet50 (0.591) for occurrence detection.

5. CONCLUSIONS

By evaluating combinations of different components and their parameters, we addressed how design choices in DL systems influence performance in facial AU coding and several findings standout. The source domain in which pre-training was performed influenced the performance of fine-tuning in the target domain. Generic pre-training proved better than a face-specific one. Face-specific pre-training indicates the training to learn identity but ignore the facial expression. Another important factor contributing to performance is the choice

of different learning rates for different optimizers. For Adam optimizer, small LR was optimal. For SGD optimizer, large LR was optimal for expression coding. Best parameters of the optimizers were similar for both AU occurrence detection and AU intensity estimation, while varying the training set size and the type of image normalization had little effect on performance.

We also evaluated cross-pose and cross-domain generalizability of the proposed method and presented occlusion sensitivity maps and saliency maps to reveal key features for each facial AU. Our models outperformed other state-of-the-art approaches in the cross-domain experiments. Cross-pose evaluation showed that our models performed well for unseen poses.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html, http://mohammadmahoor.com/disfa/, http://www.jeffcohn.net/Resources/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the IRB committee of Carnegie Mellon University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

KN implemented the architecture, ran the experiments, and wrote the manuscript with support from the other authors. IO implemented the visualization modules. JC contributed to the design and writing. LJ contributed to the conceptualization, design, and writing and supervised the project. All authors discussed the results and contributed to the final manuscript.

FUNDING

This research was supported in part by Fujitsu Laboratories of America, NIH awards NS100549 and MH096951, and NSF award CNS-1629716.

REFERENCES

Amirian, M., Kächele, M., Palm, G., and Schwenker, F. (2017). "Support vector regression of sparse dictionary-based features for view-independent action unit intensity estimation," in *Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC), 854–859. doi: 10.1109/FG.2017.109

Baltrušaitis, T., Mahmoud, M., and Robinson, P. (2015). "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Automatic Face & Gesture Recognition and Workshops (FG 2015)* (Ljubljana). doi: 10.1109/FG.2015.7284869

Batista, J. C., Albiero, V., Bellon, O. R. P., and Silva, L. (2017). "AUMPNet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network," in *Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC), 868–871. doi: 10.1109/FG.2017.111

Cavallo, F., Semeraro, F., Fiorini, L., Magyar, G., Sinčák, P., and Dario, P. (2018). Emotion modelling for social robotics applications: a review. *J. Bionic Eng.* 15, 185–203. doi: 10.1007/s42235-018-0015-y

Chu, W.-S., De la Torre, F., and Cohn, J. F. (2019). Learning facial action units with spatiotemporal cues and multi-label sampling. *Image Vision Comput.* 81, 1–14. doi: 10.1016/j.imavis.2018.10.002

Cohn, J. F., and De la Torre, F. (2015). "Automated face analysis for affective computing," in *Handbook of Affective Computing*, eds R. A. Calvo, S. K. D'Mello, J. Gratch, and A. Kappas (New York, NY: Oxford), 131–150.

Cohn, J. F., Ertugrul, I. O., Chu, W. S., Girard, J. M., Jeni, L. A., and Hammal, Z. (2019). "Chapter 19–affective facial computing: generalizability across domains," in *Multimodal Behavior Analysis in the Wild, Computer Vision and Pattern Recognition* (Academic Press), 407–441. doi: 10.1016/B978-0-12-814601-9.00026-2

Corneanu, C. A., Simón, M. O., Cohn, J. F., and Guerrero, S. E. (2016). Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 1548–1568. doi: 10.1109/TPAMI.2016.2515606

Ekman, P., Friesen,W., and Hager, J. (2002). *Facial Action Coding System: Research Nexus Network Research Information*. Salt Lake City, UT: Paul Ekman Group.

Ertugrul, I. O., Cohn, J. F., Jeni, L. A., Zhang, Z., Yin, L., and Ji, Q. (2020). Crossing domains for AU coding: perspectives, approaches, and measures. *IEEE Trans. Biometr. Behav. Identity Sci.* 2, 158–171. doi: 10.1109/TBIOM.2020.2977225

Ertugrul, I. O., Jeni, L. A., and Cohn, J. F. (2018). "FACSCaps: pose independent facial action coding with capsules," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (Salt Lake City, UT), 2130–2139. doi: 10.1109/CVPRW.2018.00287

Ghosh, S., Laksana, E., Scherer, S., and Morency, L. P. (2015). "A multi-label convolutional neural network approach to cross-domain action unit detection," in *International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xian). doi: 10.1109/ACII.2015.7344632

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika* 40, 33–51. doi: 10.1007/BF02291478

He, J., Li, D., Yang, B., Cao, S., Sun, B., and Yu, L. (2017). "Multi view facial action unit detection based on CNN and BLSTM-RNN," in *Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC), 848–853. doi: 10.1109/FG.2017.108

Jeni, L. A., Lorincz, A., Nagy, T., Palotai, Z., Sebok, J., Szabo, Z., et al. (2012). 3D shape estimation in video sequences provides high precision evaluation of facial expressions. *Image Vision Comput.* 30, 785–795. doi: 10.1016/j.imavis.2012.02.003

King, D. E. (2009). Dlib-ml: a machine learning toolkit. *J. Mach. Learn. Res.* 10, 1755–1758. doi: 10.5555/1577069.1755843

Kumano, S., Otsuka, K., Yamato, J., Maeda, E., and Sato, Y. (2009). Pose-invariant facial expression recognition using variable-intensity templates. *Int. J. Comput. Vision* 83, 178–194. doi: 10.1007/s11263-008-0185-x

Li, S., and Deng, W. (2018). Deep facial expression recognition: a survey. *arXiv* 1804.08348. doi: 10.1109/TAFFC.2020.2981446

Li, W., Abtahi, F., Zhu, Z., and Yin, L. (2018). EAC-Net: deep nets with enhancing and cropping for facial action unit detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 2583–2596. doi: 10.1109/TPAMI.2018.2791608

Li, X., Chen, S., and Jin, Q. (2017). "Facial action units detection with multi-features and -AUs fusion," in *Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC), 860–865. doi: 10.1109/FG.2017.110

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., and Matthews, I. (2011). "Painful data: the Unbc-Mcmaster shoulder pain expression archive database," in *Face & Gesture Recognition and Workshops (FG 2011)* (Santa Barbara, CA), 57–64. doi: 10.1109/FG.2011.5771462

Martinez, B., Valster, M. F., Jiang, B., and Pantic, M. (2017). Automatic analysis of facial actions: a survey. *IEEE Trans. Affect. Comput*. 10, 325–347. doi: 10.1109/TAFFC.2017.2731763

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). DISFA: a spontaneous facial action intensity database. *IEEE Trans. Affect. Comput*. 4, 151–160. doi: 10.1109/T-AFFC.2013.4

McColl, D., Hong, A., Hatakeyama, N., Nejat, G., and Benhabib, B. (2016). A survey of autonomous human affect detection methods for social robots engaged in natural HRI. *J. Intell. Robot. Syst*. 82, 101–133. doi: 10.1007/s10846-015-0259-2

Niinuma, K., Jeni, L. A., Ertugrul, I. O., and Cohn, J. F. (2019). "Unmasking the devil in the details: what works for deep facial action coding?" in *British Machine Vision Conference (BMVC)* (Cardiff).

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognition," in *British Machine Vision Conference* (Swansea). doi: 10.5244/C.29.41

Rudovic, O., Pantic, M., and Patras, I. (2013). Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Trans. Pattern Anal. Mach. Intell*. 35, 1357–1369. doi: 10.1109/TPAMI.2012.233

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). "Deep inside convolutional networks: visualising image classification models and saliency maps," in *International Conference on Learning Representations (ICLR) Workshop* (Banff, AB).

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)* (Vancouver, BC).

Tősér, Z., Jeni, L. A., Lőrincz, A., and Cohn, J. F. (2016). "Deep learning for facial action unit detection under large head poses," in *Computer Vision–ECCV 2016 Workshops* (Amsterdam), 359–371. doi: 10.1007/978-3-319-49409-8_29

Taheri, S., Turaga, P., and Chellappa, R. (2011). "Towards view-invariant expression analysis using analytic shape manifolds," in *Face and Gesture 2011* (Santa Barbara, CA), 306–313. doi: 10.1109/FG.2011.5771415

Tang, C., Zheng, W., Yan, J., Li, Q., Li, Y., Zhang, T., et al. (2017). "View-independent facial action unit detection," in *Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC), 878–882. doi: 10.1109/FG.2017.113

Valstar, M. F., Sánchez-Lozano, E., Cohn, J. F., Jeni, L. A., Girard, J. M., Zhang, Z., Yin, L., and Pantic, M. (2017). "FERA 2017–addressing head pose in the third facial expression recognition and analysis challenge," in *Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC), 839–847. doi: 10.1109/FG.2017.107

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision* (Zurich), 818–833. doi: 10.1007/978-3-319-10590-1_53

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., and Girard, J. M. (2014). BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image Vision Comput*. 32, 692–706. doi: 10.1016/j.imavis.2014.06.002

Zhang, Z., Girard, J. M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., et al. (2016). "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 3438–3446. doi: 10.1109/CVPR.2016.374

Zhi, R., Liu, M., and Zhang, D. (2019). A comprehensive survey on automatic facial action unit analysis. *Visual Comput*. 36, 1067–1093. doi: 10.1007/s00371-019-01707-5

Zhou, Y., Pi, J., and Shi, B. E. (2017). "Pose-independent facial action unit intensity regression based on multi-task deep transfer learning," in *Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC), 872–877. doi: 10.1109/FG.2017.112

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.