frontiers | Frontiers in Signal Processing

ORIGINAL RESEARCH
published: 27 April 2022
doi: 10.3389/frsip.2022.861641

Facial Expression Manipulation for Personalized Facial Action Estimation

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

[1]Fujitsu Research of America, Pittsburgh, PA, United States, [2]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands, [3]Department of Psychology, University of Pittsburgh, Pittsburgh, PA, United States, [4]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States

Limited sizes of annotated video databases of spontaneous facial expression, imbalanced action unit labels, and domain shift are three main obstacles in training models to detect facial actions and estimate their intensity. To address these problems, we propose an approach that incorporates facial expression generation for facial action unit intensity estimation. Our approach reconstructs the 3D shape of the face from each video frame, aligns the 3D mesh to a canonical view, and trains a GAN-based network to synthesize novel images with facial action units of interest. We leverage the synthetic images to achieve two goals: 1) generating AU-balanced databases, and 2) tackling domain shift with personalized networks. To generate a balanced database, we synthesize expressions with varying AU intensities and perform semantic resampling. Our experimental results on FERA17 show that networks trained on synthesized facial expressions outperform those trained on actual facial expressions and surpass current state-of-the-art approaches. To tackle domain shift, we propose personalizing pretrained networks. We generate synthetic expressions of each target subject with varying AU intensity labels and use the person-specific synthetic images to fine-tune pretrained networks. To evaluate performance of the personalized networks, we use DISFA and PAIN databases. Personalized networks, which require only a single image from each target subject to generate synthetic images, achieved significant improvement in generalizing to unseen domains.

Keywords: facial expression recognition, facial action unit intensity estimation, facial action unit detection, facial expression synthesis, generative adversarial network, 3D face registration, synthetic data augmentation, model personalization

OPEN ACCESS

Edited by:
Kidiyo Kpalma,
Institut National des Sciences Appliquées de Rennes, France

Reviewed by:
Alireza Sepas-Moghaddam,
Queen's University, Canada
Xianye Ben,
Shandong University, China

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

Specialty section:
This article was submitted to Image Processing, a section of the journal Frontiers in Signal Processing

Received: *24 January 2022*
Accepted: *05 April 2022*
Published: *27 April 2022*

Citation:
Niinuma K, Onal Ertugrul I, Cohn JF and Jeni LA (2022) Facial Expression Manipulation for Personalized Facial Action Estimation. Front. Sig. Proc. 2:861641. doi: 10.3389/frsip.2022.861641

1 INTRODUCTION

Facial expression conveys emotional state, behavioral intention, and physical state (Tian et al., 2001). In behavior sciences the gold-standard to decode facial expressions is the Facial Action Coding System (FACS) (Ekman et al., 2002). FACS decomposes facial expressions into anatomically based action units (AUs), which alone or in combinations can represent nearly all possible facial expressions. While much progress has been made in action unit detection, at least three significant problems impede further advances.

First, while abundant videos of spontaneous facial expression has been collected, only a fraction have been manually annotated. There are relatively few expert FACS annotators and the time required to comprehensively annotate action units slows the effort as well. AU annotation of a single minute of video typically requires one to three hours for a highly trained expert (Cohn and Ekman,

TABLE 1 | Comparison of AU intensity datasets.

	FERA[a]	DISFA	EmotioNet	UNBC Pain
# of AUs with intensity codes	7	12	12[b]	10
Continuous Video	✓	✓		✓
Manual Ground Truth	✓	✓	Semi automated[c]	✓
Social Context	✓			
Manual Coding	Good	Good	Unknown	Good

[a]*The FERA, 2017 dataset consists of BP4D (Zhang et al., 2014) and BP4D+(Zhang et al., 2016).*
[b]*In the EmotioNet Challenge 2020, 11 more AUs have been added for a total of 23 AUs. http://cbcsl.ece.ohio-state.edu/enc-2020/*
[c]*They manually FACS-coded 10% of this database.*

2005). Recently, Ben et al. (2021) released the initial work on AU annotation for micro-expression recognition by a micro-and-macro expression warehouse (MMEW), which is consonant with the needs of future cross-modal affective computing. Reaching human-like accuracy in a fully supervised way would require labeled datasets orders of magnitude larger than those available today.

Second, AU labels are highly skewed in spontaneous behaviors. Many AUs occur rarely and only a sparse subset of AU intensities occur at a time. As a consequence, rare classes do not contribute equally during classifier training, which hinders learning and undermines global performance. Although imbalanced learning for AU estimation has been well explored in the past, most approaches deal with a single majority and minority class and are not directly applicable to the multi-label AU domain.

And third, domain shift between source and target datasets degrades the performance of trained models. For these reasons, generalization abilities of trained AU occurrence and intensity models to unseen imaging conditions and new subjects are a critical research challenge. In research to date, some existing approaches have proposed personalized models to solve these problems (Wang and Wang, 2018; Yang et al., 2018; Lee et al., 2019; Wang et al., 2019; Cai et al., 2021). In these approaches, however, tasks are typically limited to basic facial expression recognition or at most AU occurrence detection. Because facial expression is a dynamic process, intensity estimation robust to domain shift and individual differences is needed.

To address these problems, we propose a generative semi-supervised method. Our approach tackles three obstacles within a common framework: the limited size of annotated data now available; the low frequencies of occurrence; and the domain shift between the source and target datasets.

Our approach makes use of a 3D facial expression generator trained on the labeled portion for two of the goals: 1) generating AU-balanced databases, and 2) tackling domain shift with personalized networks. The approach first reconstructs the 3D shape of the face from each video frame, aligns the reconstructed meshes to a canonical view to establish semantic correspondence across frames and subjects, and then trains a GAN-based network to synthesize novel images with facial action units of interest.

The intensity of facial actions may be one of the most important features in assessing a person's emotional state (McKeown et al., 2015). Low intensity actions are detectable through motion (Ambadar et al., 2005). **Table 1** compares AU intensity datasets. In order to be able to detect these fine scale changes, we selected FERA 2017 (Valstar et al., 2017), DISFA (Mavadati et al., 2013) and UNBC Pain(Lucey et al., 2011), which are video datasets that have manual fine-grained AU intensity annotations, to evaluate our approach. Note that **Table 1** does not include datasets without AU intensity annotation, such as Aff-Wild2 (Kollias and Zafeiriou, 2019).

Our novelties are threefold:

3D geometry based AU manipulation. Unlike previous work on facial AU manipulation that is limited to either 2D representations (Pumarola et al., 2019) or individual frames (Geng et al., 2019), our approach uses the 3D structure of the face to create semantic correspondence across video-frames and subjects.

Synthetic multi-label stratification of AUs. Many AUs occur infrequently, which undermines learning. To mitigate imbalanced learning, we increase the prevalence and variety of under-represented AUs by synthesizing new facial expression.

Personalization using person-specific synthetic images. To tackle domain shift, we personalize pre-trained models using person-specific synthesis images generated from each target subject. Our approach requires only a single image from each subject to generate synthetic images.

This study is an extended version of our conference paper (Niinuma et al., 2021a). In this study, we proposed a model personalization approach using person-specific synthesis images for AU intensity estimation, conducted experiments to show the effectiveness of our model personalization approach, and performed literature review for model personalization. We also performed additional experiments for models trained on combined datasets, and comparison of GAN architecture.

The rest of this paper is organized as follows: **Section 2** introduces related work, **Section 3** discusses the proposed methods, **Section 4** gives the experimental results and analysis, and **Section 5** provides conclusions.

2 RELATED WORK

Solutions to limited AU annotations: While a massive amount of facial expression data is available, high quality annotations of AU intensity labels are limited. To mitigate the problems in the AU annotations, weakly-supervised, semi-supervised, and self-supervised approaches have been proposed. Weakly-supervised approaches aim to exploit incomplete, inaccurate or inexact

annotations to provide supervision. Zhao et al. (2018) proposed a weakly supervised clustering approach utilizing a large set of web images with inaccurate annotations. The annotations were obtained from either pretrained models or query strings. Ruiz et al. (2015) proposed training AU detectors without any AU annotations by leveraging the expression labels and using prior knowledge on expression-dependent AU probabilities. Similarly, Zhang et al. (2018b) exploited expression-dependent and expression-independent joint AU probabilities as prior knowledge and learned to detect AUs without any AU annotation. In another study, Zhang et al. (2018a) used various types of domain knowledge including relative appearance similarity, temporal intensity ordering, facial symmetry, and contrastive appearance difference, to provide weak supervision for AU intensity estimation with extremely limited annotations. Peng and Wang (2018) proposed a method that learns AU classifiers from domain knowledge and expression-annotated facial images through adversarial training. Sun et al. (2021) proposed an AU detection framework leveraging enormous and diverse facial images labeled with six basic facial expressions without AU annotations. Li and Shan (2021) also proposed an approach utilizing expression-annotated images. Their approach automatically selected highly related facial expression samples by learning adaptative weights for the training facial expression samples in a meta learning manner. Song et al. (2021) focused on inferring temporal dynamics of facial actions when no explicit temporal information is available from still images, and presented a self-supervised approach to capture multiple scales of temporal dynamics for AU intensity estimation. Yin et al. (2021) proposed an AU detection architecture that can be jointly trained for self-supervised optical flow estimation, patch localization, supervised action unit detection, and adversarial domain adaptation.

Semi-supervised approaches deal with partially annotated data. They aim to leverage the unlabeled data with the assumption that unlabeled data follow continuity or form cluster with the labeled data (Zhao et al., 2018). Wu et al. (2015) used Restricted Boltzmann Machine to model the AU distribution using the annotated labels, which is used to train the AU classifiers with partially labeled data. Zeng et al. (2015) trained a quasi-semi-supervised (QSS) classifier with virtual labels provided by the confident positive and negative classifiers, which separate easily identified positive and negative samples from all else, respectively. Niu et al. (2019) proposed a semi-supervised co-training approach named as multi-label coregularization for AU recognition, which aims to improve AU recognition with abundant unlabeled face images and domain knowledge of AUs.

Recent work focuses on self-supervised approaches where the goal is to learn the discriminative representation from the massive amount of videos without annotations. Li et al. (2019) proposed a self-supervised learning framework named Twin-Cycle Autoencoder that disentangles the AU-related movements from the pose-related ones to learn AU representations from unlabeled videos. While the aforementioned approaches tackle the problems in the annotations, none of them aims to balance the distribution of AU intensities. For label balancing, upsampling approaches where the infrequent labels are selected multiple times (Li et al., 2017; Zhang Z. et al., 2018) or a multi-label minority oversampling majority undersampling approach (Chu et al., 2019) have been used. Since resampling is done within the dataset, such balancing methods do not contribute additional semantic information about the infrequent label.

GAN-based facial expression transfer: Recently GANs have received attention to transfer facial expressions from a source subject to a target subject. Existing work on GAN-based facial expression transfer approaches focus on generating facial images with discrete emotions (Choi et al., 2018; Ding et al., 2018), or the specified facial action units (Pumarola et al., 2019; Liu et al., 2020). Some of the GAN-based approaches specifically aim to guide their models with the facial geometry information. Song et al. (2018) proposed a Geometry-Guided Generative Adversarial Network (G2-GAN) which employs fiducial points as a controllable condition to guide facial texture synthesis with specific expressions. Qiao et al. (2018) applied contrastive learning in GAN to embed geometry information onto a semantic manifold in the latent space for facial expression transfer. Geng et al. (2019) combined 3DMMs and deep generative techniques in a single framework for fine-grained face manipulation. Yet, in these studies transfer was limited to either 2D representation or to individual frames.

Synthetic data augmentation: Some of the recently proposed methods utilized synthetic data for facial expression analysis. Abbasnejad et al. (2017) pre-trained their model using synthetic face images and then fine-tuned it on real images. Zhu et al. (2018) proposed a data augmentation method using GAN to classify basic emotions. Kollias et al. (2020) proposed an approach using 3DMMs to synthesize facial affect: in terms of six basic emotions or in terms of valence and arousal. Unlike the existing methods, our approach is designed to generate a large AU-balanced dataset.

Model personalization: While there were some shallow based approaches (Rudovic et al., 2015; Chu et al., 2017), most recent methods utilized deep-learning based models. Lee et al. (2019) proposed an approach based on the notion of model-agnostic meta-learning for AU occurrence detection. Some other existing methods (Wang and Wang, 2018; Yang et al., 2018; Wang et al., 2019; Cai et al., 2021) utilized a GAN architecture to personalize models as with our proposed approach. Yang et al. (2018) utilized a GAN to generate a person-specific sub-space for each subject to perform six basic facial expression recognition. Wang et al. (2019) proposed an adversarial feature learning method addressing both identity and pose biases for basic expression recognition. Cai et al. (2021) proposed an approach synthesizing identity-free expressive faces to classify six basic expression. Wang and Wang (2018) proposed an architecture that trains a generator and a AU classifier at the same time for AU occurrence detection. Kim and Song (2021) presented a person independent facial emotion recognition by adversarily learning weak emotion samples based on strong emotion samples. While they showed promising results, their tasks were limited to basic facial expression recognition or AU occurrence detection. Unlike the existing methods, the target

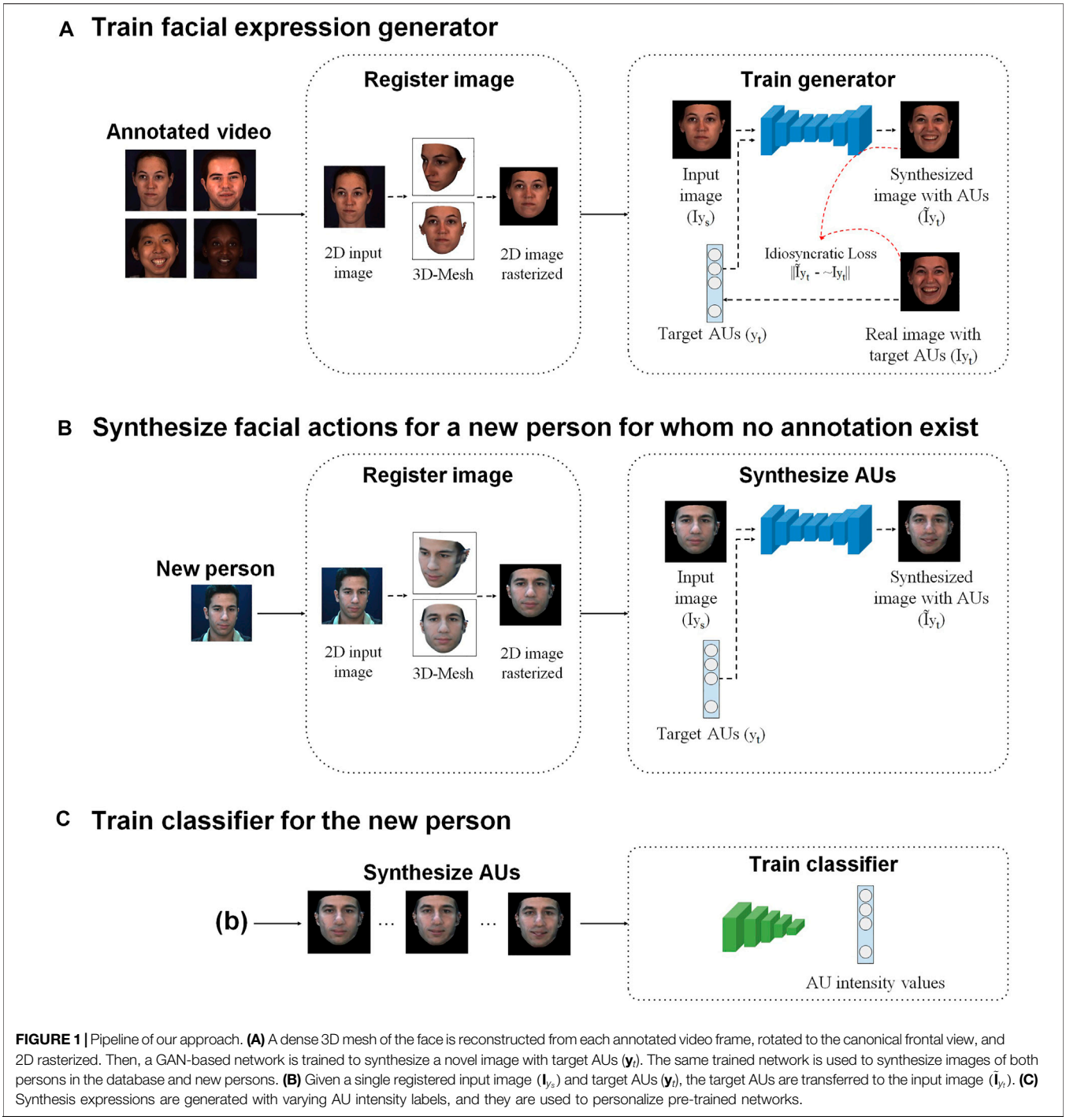

FIGURE 1 | Pipeline of our approach. **(A)** A dense 3D mesh of the face is reconstructed from each annotated video frame, rotated to the canonical frontal view, and 2D rasterized. Then, a GAN-based network is trained to synthesize a novel image with target AUs ($\mathbf{y}_t$). The same trained network is used to synthesize images of both persons in the database and new persons. **(B)** Given a single registered input image ($\mathbf{I}_{y_s}$) and target AUs ($\mathbf{y}_t$), the target AUs are transferred to the input image ($\tilde{\mathbf{I}}_{y_t}$). **(C)** Synthesis expressions are generated with varying AU intensity labels, and they are used to personalize pre-trained networks.

of our approach is AU intensity estimation that requires fine-grained facial expression synthesis.

3 PROPOSED METHOD

Figure 1 shows the pipeline of our approach. To generate the synthesis images, we first perform dense 3D registration from 2D images. Then, we train a GANimation-based architecture with idiosyncratic loss to synthesize new facial expressions. We synthesize images to achieve two objectives: 1) training generic models with a completely balanced AU database, and 2) personalizing models for each target subject. To obtain a balanced distribution of AU intensity labels, we generate synthesis images from facial images in the train dataset. The balanced dataset is used to train generic models. To obtain personalized models, we generate synthetic facial expressions of each target subject with varying AU intensity labels, and use the synthetic images to fine-tune the generic models.

3.1 3D Face Registration

We normalize videos using PRNet (Feng et al., 2018), a face alignment software that accomplishes dense 3D registration from 2D images without requiring person-specific training. PRNet uses an encoder-decoder architecture containing convolutional layers and residual blocks to jointly perform facial landmark alignment and 3D facial structure reconstruction. This architecture learns a mapping from an RGB image to UV position map (a 2D image representation of 3D coordinates in UV space keeping the position and semantic information). By learning the position map, it is possible to directly regress the complete 3D structure along with semantic meaning from a single image. Using PRNet, we obtain the dense 3D mesh of the face in a frontal view and texture information. Then, we map the texture to 3D mesh and rasterize it to 2D image of size 224×224.

3.2 Facial Expression Generation Architecture

We build upon GANimation (Pumarola et al., 2019) framework to synthesize novel facial expressions. First we map AU intensity labels (0 to E-level) to values in range [0,1]. Given a 3D registered source image $\mathbf{I}_{y_s}$ with the AU intensity values $\mathbf{y}_s = \{s_1, s_2, \ldots, s_n\}$, and target AU intensity values $\mathbf{y}_t = \{t_1, t_2, \ldots, t_n\}$, we synthesize $\tilde{\mathbf{I}}_{y_t}$. The integer n is the number of AUs. With our architecture, we aim to minimize these following terms:

Image Adversarial Loss: In order to obtain realistic synthesized images and ensure that the distribution of the generated images are similar to the distribution of the training images, we use image adversarial loss. Let $\mathcal{P}_s$ be the data distribution of the source image, $\mathcal{P}_{\tilde{I}}$ the random interpolation distribution, and λ_{ggp} the penalty loss. Then we can write the image adversarial loss $\mathcal{L}_{adv}(G, D_{adv}, \mathbf{I}_{y_s}, \mathbf{y}_t)$ as follows:

$$\mathbb{E}_{\mathbf{I}_{y_s}\sim\mathcal{P}_s}\left[D_{adv}\left(G\left(\mathbf{I}_{y_s}|\mathbf{y}_t\right)\right)\right] - \mathbb{E}_{\mathbf{I}_{y_s}\sim\mathcal{P}_s}\left[D_{adv}\left(\mathbf{I}_{y_s}\right)\right] + \lambda_{gp}\mathbb{E}_{\tilde{I}\sim\mathcal{P}_{\tilde{I}}}\left[\left(\|\nabla_{\tilde{I}} D_{adv}\left(\tilde{I}\right)\|_2 - 1\right)^2\right] \quad (1)$$

where G denotes generator and D_{adv} denotes adversarial discriminator.

Conditional Expression Loss: In order to enforce G to synthesize images containing the target expression $\mathbf{y}_t$, we use the following loss $\mathcal{L}_{exp}(G, D_{exp}, \mathbf{I}_{y_s}, \mathbf{y}_s, \mathbf{y}_t)$ to minimize the distance between AU intensities of the images and those predicted with D_{exp} for both source and synthesized images:

$$\mathbb{E}_{\mathbf{I}_{y_s}\sim\mathcal{P}_s}\left[\|D_{\exp}\left(G\left(\mathbf{I}_{y_s}|\mathbf{y}_t\right)\right) - \mathbf{y}_t\|_2^2\right] + \mathbb{E}_{\mathbf{I}_{y_s}\sim\mathcal{P}_s}\left[\|D_{exp}\left(\mathbf{I}_{y_s}\right) - \mathbf{y}_s\|_2^2\right] \quad (2)$$

where $D_{\exp}$ denotes discriminator for expression.

Identity Loss: We aim to guarantee that the face in both the input and output images belong to the same person. We use this cycle-consistency loss to penalize the difference between the original image $\mathbf{I}_{y_s}$ and its reconstruction $\mathbf{I}_{y_t}$.

$$\mathcal{L}_{idt}\left(G, \mathbf{I}_{y_s}, y_s, y_t\right) = \mathbb{E}_{\mathbf{I}_{y_s}\sim\mathcal{P}_s}\left[\|G\left(G\left(\mathbf{I}_{y_s}|y_t\right)|y_s\right) - \mathbf{I}_{y_s}\|_1\right] \quad (3)$$

Idiosyncratic Loss: With the GANimation architecture, we can transfer the AU intensity values $\mathbf{y}_t$ of a target image $\mathbf{I}_{y_t}$ to the source image $\mathbf{I}_{y_s}$ to synthesize $\tilde{\mathbf{I}}_{y_t}$. When the identity of source (I) and target (J) images are the same, we can minimize the difference between $\mathbf{I}_{y_t}$ and $\tilde{\mathbf{I}}_{y_t}$ to ensure that both expression and identity of the synthesized image are the same as the target image. Idiosyncratic loss can be defined as:

$$\mathcal{L}_{ids}\left(G, \mathbf{I}_{y_s}, y_s, y_t\right) = \mathbb{E}_{\mathbf{I}_{y_s}\sim\mathcal{P}_s}\left[\|\left(G\left(\mathbf{I}_{y_s}|y_t\right) - \mathbf{I}_{y_t}\|_1\right]\right] \quad (4)$$

Final Loss: We obtain our final loss by combining of the mentioned individual losses as follows:

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{exp}\mathcal{L}_{exp} + \lambda_{idt}\mathcal{L}_{idt} + \lambda_{ids}\mathcal{L}_{ids} \quad (5)$$

where λ_{adv}, $\lambda_{\exp}$, λ_{idt}, and λ_{ids} are the hyperparameters used to adjust the importance of different components.

3.3 Train AU Classifier Without Personalization

We perform semantic resampling for the imbalanced training dataset and create a training set having balanced AU intensity labels for each AU. Then we train convolutional neural networks (VGG16) using the balanced synthetic training set. We trained a separate model for each AU. During the test time, we obtain AU intensity outputs of each estimator.

3.4 Train AU Classifier With Personalization

We obtain a single image from each target subject, and generate synthetic expressions for each AU intensity. Then, we use the person-specific synthetic images to personalize the person-independent models (VGG16). Because AU intensity is categorized into 6 classes, we generate 6 synthesis images for each AU, and use them to fine-tune generic models.

4 EXPERIMENTS

4.1 Datasets

In all of our experiments, we used four facial expression datasets. For training the generator and evaluating within domain performance, we used the widely accepted 2017 Facial Expression Recognition Benchmark (FERA 2017) (Valstar et al., 2017). For generating out-of-domain samples, we used the high resolution images from MultiPIE (Gross et al., 2010). To evaluate the generalizability of our AU classifiers to another domain, we used the Denver Intensity of Spontaneous Facial Action Database (DISFA) (Mavadati et al., 2013) and the UNBC Pain dataset (Lucey et al., 2011).

FERA 2017: The FERA 2017 Challenge was the first to provide a common protocol with which to compare approaches to detection of AU occurrence and AU intensity robust to pose variation. FERA 2017 provided synthesized face images with 9 head poses as shown in **Figure 5**. The training set is based on the BP4D database (Zhang et al., 2014), which includes digital videos of 41 participants. The development set and test set are derived from BP4D+(Zhang et al., 2016) and

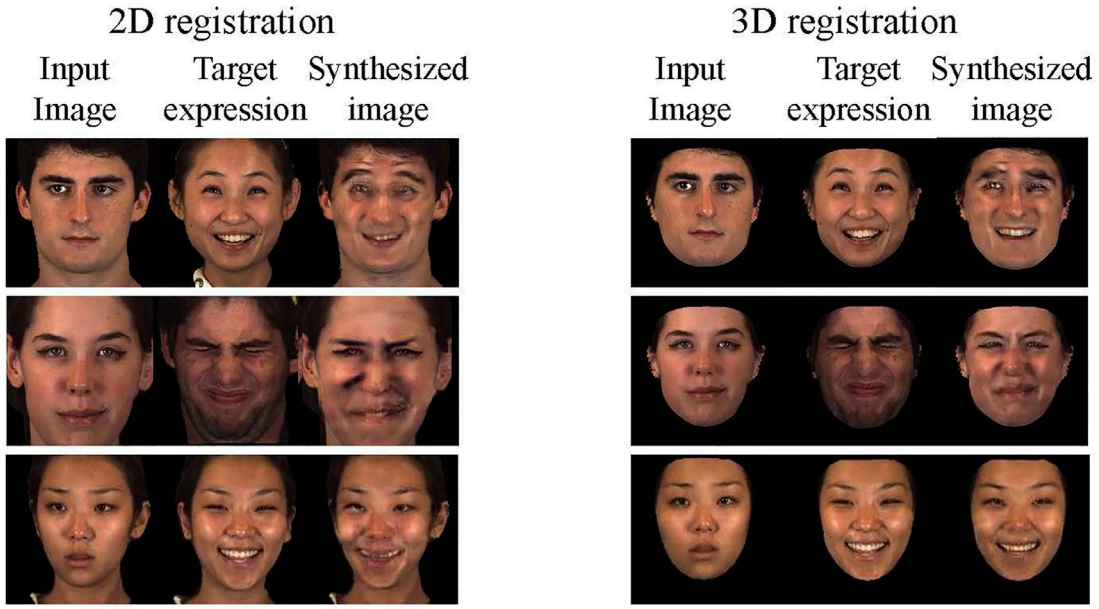

FIGURE 2 | Examples of 2D and 3D normalization. Note the higher image quality when 3D normalization was used.

TABLE 2 | Comparison of synthetic vs. real expressions under 2D vs. 3D alignment on FERA 2017 Test partition. Scores are Inter-rater reliability (ICC) of AU intensity level estimation. The Images row shows which is used to train classifiers: Real or Synthetic. All of the classifiers were tested on the real test dataset. The same Registration (2D or 3D) were applied to both train and test datasets.

Registration Images	2D Real	2D Synthetic	3D Real	3D Synthetic
AU1	**0.431**	0.336	0.343	0.381
AU4	0.223	0.116	**0.260**	0.219
AU6	0.796	0.790	0.751	**0.804**
AU10	0.777	**0.812**	0.785	0.773
AU12	0.801	0.792	**0.806**	0.795
AU14	0.118	0.238	0.084	**0.244**
AU17	0.395	0.374	0.391	**0.461**
Mean	0.506	0.494	0.489	**0.525**

The best results are shown in bold.

include digital videos of 20 and 30 participants, respectively. FERA 2017 presented two sub-challenges: occurrence detection and intensity estimation. For the former 10 AUs were labelled; for the latter, 7 AUs were labelled. For our experiments, 7 AUs for intensity estimation were used.

MultiPIE: The MultiPIE dataset contains images of 337 people recorded in up to four sessions over the span of five months. Subjects were imaged under 15 view points and 19 illumination conditions while displaying a range of facial expressions. In addition, high resolution frontal images were acquired as well. For synthesis, we used the high resolution images only.

DISFA: The DISFA dataset contains videos of 27 adult subjects (12 women, 15 men). It is manually annotated for AU intensity from 0 to E-level. Participants watched a video clip consisting of 9 segments intended to elicit a range of facial expressions of emotion.

UNBC Pain: The UNBC Pain dataset consists of videos of 25 adults, who had shoulder pain. It is also manually annotated for AU intensity from 0 to E-level. In the dataset, facial expressions are mostly associated with pain, and the correlation among AUs differs from that of FERA2017 and DISFA.

4.2 Experimental Setup

In this section we describe the experimental setup for the generator network and the classifier.

4.2.1 Train Generator

In our experiments, we used the FERA 2017 dataset (Valstar et al., 2017) to train facial expression generation models. The dataset consists of Train, Valid, and Test partitions. We used the Train partition only to train our model. All the images were resized to 224×224 pixels to match the receptive field of our AU estimation model (VGG16).

In all of our experiments, we used a GANimation (Pumarola et al., 2019) replicate implementation[1]. We modified the loss function with the idiosyncratic constraint as described in the previous section.

4.2.2 Train AU Classifier Without Personalization

For the baseline experiments, we used the training partition of the FERA 2017 dataset (Valstar et al., 2017) to train AU classifiers and used the test partition to test them. To create a balanced training set and compare methods, 5,000 frames

[1]https://github.com/donydchen/ganimation_replicate.

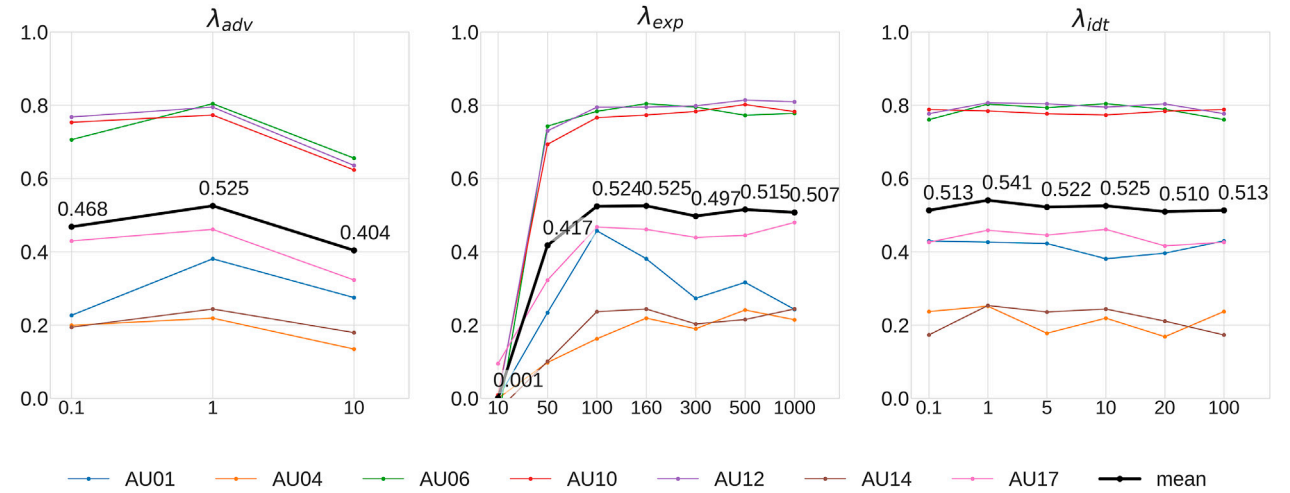

FIGURE 3 | Influence of parameter values on ICC for intensity estimation.

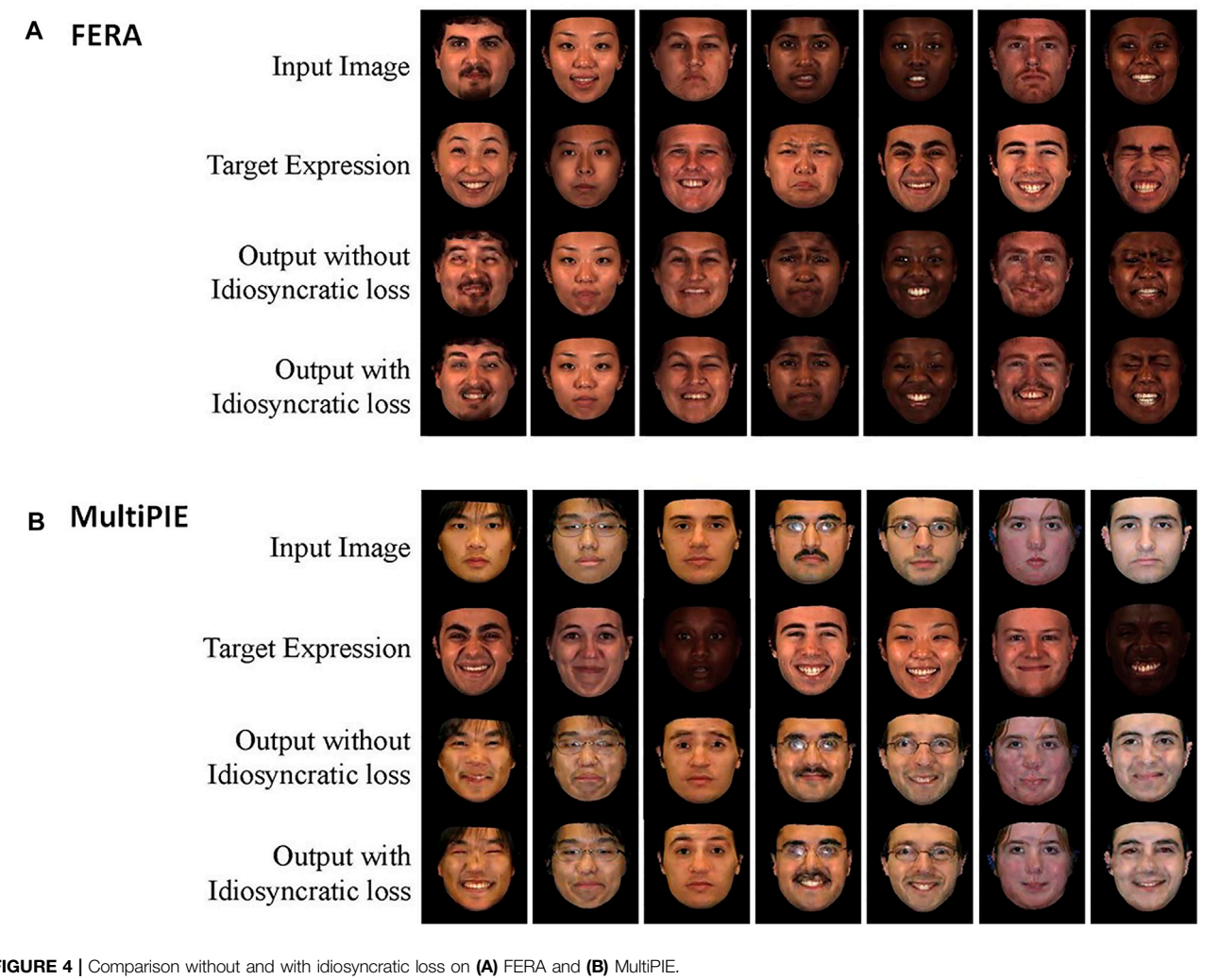

FIGURE 4 | Comparison without and with idiosyncratic loss on **(A)** FERA and **(B)** MultiPIE.

TABLE 3 | ICC for intensity estimation without and with idiosyncratic loss on FERA 2017 Test partition.

	Without idiosyncratic loss	With idiosyncratic loss
AU1	**0.381**	0.350
AU4	0.219	**0.302**
AU6	**0.804**	0.802
AU10	0.773	**0.780**
AU12	**0.795**	0.793
AU14	**0.244**	0.181
AU17	**0.461**	0.452
Mean	**0.525**	0.523

The best results are shown in bold.

were selected for each AU intensity. In the case of real images, we randomly selected 5,000 images from the six intensity classes (not present, and A to E levels). We down-sampled the majority classes and up-sampled the minority classes to reach this number. In the case of synthesized images, we first selected 5,000 pairs of input images and target AU labels for each intensity and each AU. Then we synthesized 5,000 images for these pairs. We selected 5,000 images according to experimental results by Niinuma et al. (2021b). They analyzed the influence of training set size on FERA17, and showed that the training set size have a minor influence on the performance: score peaked at 5,000 images, and after that performance plateaued. During the selection input images were selected only from frames that did not have the target AU of interest. For example, when we synthesized images for E level intensity of AU1, we randomly selected 5,000 input images that did not have AU1 present and 5,000 target AU labels having E level intensity for AU1. We employ this strategy for two main reasons. First, while the generator can add realistic facial deformations (like wrinkles and bulges) to neutral faces, it oftentimes fails in removing those features. Therefore, to acquire higher quality synthesized images, starting from neutral frame is preferable. Second, since the AU labels are sparse, there are many more frames in the dataset where the target AU is not present. This way we can obtain a variety of synthesized images for each facial expression.

We selected a VGG16 network pre-trained on ImageNet for the baseline architecture for AU estimation. Previous studies found this combination preferable for AU coding (Niinuma et al., 2021b). We replaced the final layer of the network with a 6-length one-hot representation, and fine-tuned VGG16 network from the third convolutional layer. The dropout rate was set to 0.5, and an Adam optimizer was used with $LR = 5 \times 10^{-5}$ as suggested in (Niinuma et al., 2021b). We fine-tuned the pre-trained model for 10 epochs, chose the model showing the best performance on the validation partition from the 10 models, and then reported the results on the subject-independent test partition.

4.2.3 Train AU Classifier With Personalization

In the experiments, the effect of our personalized AU classifier was evaluated. We first selected a single image from a target subject, generated person-specific synthesis images, and then personalized generic models by fine-tuning them with the synthesis images.

Any single image with any expression can be used to generate the synthesis images because our facial expression generator does not require AUs of the input image. In our experiments, we selected an image with a neutral expression from each target subject because a neutral frame is preferable to acquire higher quality synthesis images as described in **Section 4.2.2**. Since people often show a neutral or nearly neutral expression in most scenarios, we believe that it is realistic to assume that we can obtain a neutral frame. In our experiments, we trained a separate model for each AU. Because AU intensity is categorized into 6 ordinal classes (0 to 5), we generate 6 synthesis images for each AU and use them to fine-tune generic models.

We personalized the person-independent models trained on the AU balanced synthesis dataset described in **Section 4.2.2**. We used three test datasets to personalize the models: FERA test partition, DISFA, and UNBC Pain. The person-independent models were trained on the FERA train partition, but the subjects in the train partition do not overlap with the ones in the test partition. The Adam optimizer was used as with the classifier training in **Section 4.2.2**. However, the smaller learning rate ($LR = 5 \times 10^{-6}$) and larger epochs (30 epochs) were used for model personalization. Because we do not have a validation partition for the model personalization, the final model after 30 epochs was chosen.

4.3 Synthetic vs. Real Expressions Under 2D vs. 3D Alignment

In this set of experiments we studied how two main components affect the performance of the whole system.

First, we were interested in the effect of face alignment on the synthesis and AU recognition performance. We explored both 2D and 3D alignment. 2D normalization treats the face as a 2D object. That assumption is reasonable, as long as there is no head movement present. As soon as head orientation deviates from frontal, one expects the classifier's ability to measure expressions to degrade. On the other hand, 3D normalization should be able to preserve semantic correspondences of the different facial regions across poses, and result in higher performance. For 2D alignment, we applied Procrustes analysis between 68 landmarks provided by the dlib face tracker (King, 2009) on the frames and a frontal template. For 3D normalization we used the method described in **Sec. 3.1**. Note that there are not significant differences between shallow and deep approaches in terms of facial alignment (Jeni et al., 2016; Sagonas et al., 2016). In the 300-W Challenge (Sagonas et al., 2016), the two top methods are a cascade regressor and a CNN approach. The dlib face tracker is an implementation of a decision tree based cascade regressor (Kazemi and Sullivan, 2014).

Second, we were interested how synthetic expressions would affect the classification performance. We compared multi-label minority oversampling and majority undersampling with the proposed, completely synthetic expression generation. Both methods balance the skewed distributions of AUs, but while

TABLE 4 | Comparison with state-of-the-art methods on FERA 2017 Test partition. Reported scores are ICC on frontal views only. The results for No norm is the same with the ones for 3D Synthetic in **Table 2**.

	Valstar et al. (2017)	Amirian et al. (2017)	Batista et al. (2017)	Zhou et al. (2017)	Niinuma et al. (2021b)	Ours		
						No norm	AU0 norm	Mean norm
AU1	0.025	0.270	0.311	0.286	0.433	0.381	0.539	**0.613**
AU4	0.003	0.074	0.098	0.130	0.281	0.219	0.361	**0.409**
AU6	0.616	0.644	0.721	0.625	0.786	**0.804**	0.764	0.779
AU10	0.662	0.733	0.741	0.739	0.768	0.773	**0.787**	0.757
AU12	0.709	0.745	0.754	**0.822**	0.812	0.795	0.792	0.794
AU14	0.066	0.030	0.127	0.075	0.153	**0.244**	0.114	0.170
AU17	0.015	0.271	0.252	0.342	0.382	0.461	0.288	**0.465**
Mean	0.299	0.395	0.429	0.431	0.516	0.525	0.521	**0.570**

The best results are shown in bold.

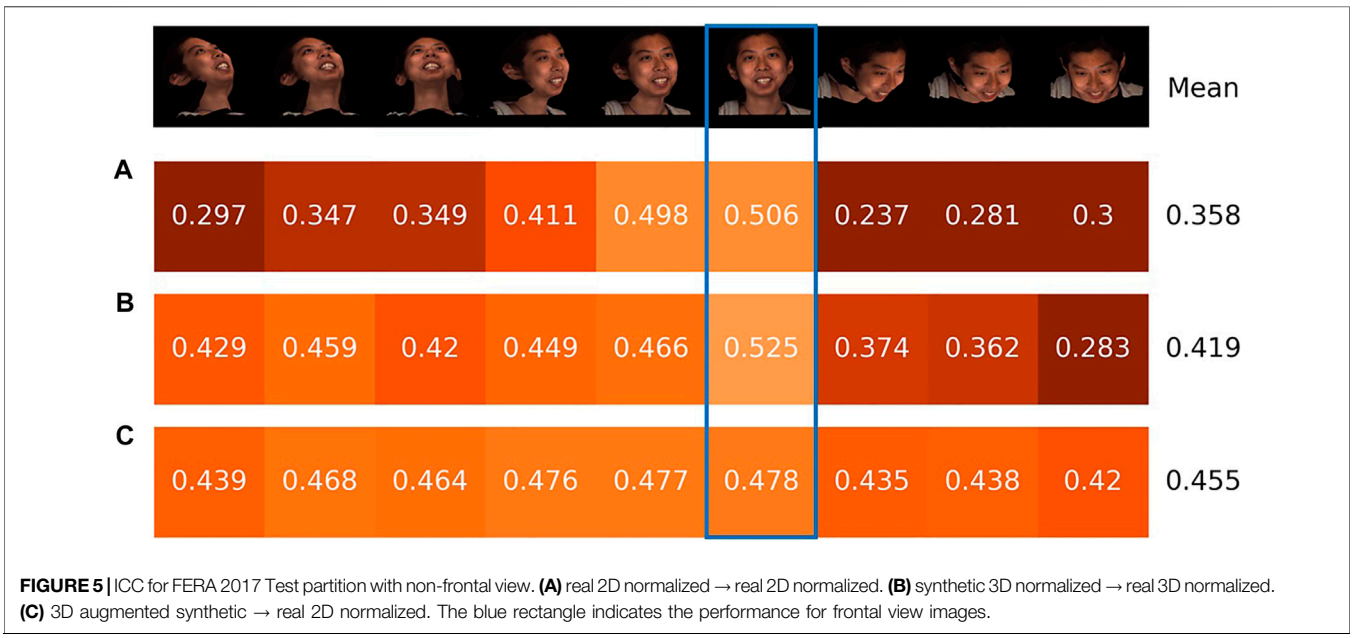

FIGURE 5 | ICC for FERA 2017 Test partition with non-frontal view. **(A)** real 2D normalized → real 2D normalized. **(B)** synthetic 3D normalized → real 3D normalized. **(C)** 3D augmented synthetic → real 2D normalized. The blue rectangle indicates the performance for frontal view images.

the first one cannot produce more varied minority samples, the latter one can. As mentioned in **Sec. 4.2.2**, we used 5,000 images for each intensity each AU for both real and synthesized image settings.

Figure 2 shows examples of synthetic expressions generated using 2D and 3D alignment. 3D alignment results in less ghosting and other texture artifacts and provides higher quality images. Inter-rater reliability (ICC) results of AU classification under these four different conditions are shown in **Table 2**. 3D registration with synthetic re-sampling outperformed the other three conditions.

In these experiments we conducted a parameter search to find the optimal values of the generator for classification. We varied parameters that control the contribution of the adversarial loss (λ_{adv}), conditional expression loss ($\lambda_{\exp}$), and the identity loss (λ_{idt}). In our baseline configuration, we used 1.0 for λ_{adv}, 160 for $\lambda_{\exp}$, and 10 for λ_{idt}. **Figure 3** shows the impact of parameter values on the intensity estimation for different AUs. λ_{idt} does not affect the classification significantly. For both $\lambda_{\exp}$ and λ_{adv}, we selected the global optimal values. Pumarola et al. introduced an attention mask and a color transformation term in the lost function to prevent the attention mask saturation (Pumarola et al., 2019). In our experiments we did not observe this saturation effect and removed these terms from the loss function.

4.4 The Effect of Idiosyncratic Loss

Although 3D alignment improves the performance, image quality is still low in some cases especially when target expression has a high AU intensity. To mitigate this problem, we introduced a new idiosyncratic term in the loss function. FERA 2017 datasets include many images with different facial expressions for each subject. Idiosyncratic loss utilizes this feature of the datasets. **Figure 4** shows some of the examples (λ_{ids} = 1). The ICC result with the new function is 0.523. To calculate the ICC result, 3D synthetic train images were used to train classifiers, and 3D real test images were used to test the classifiers. Compared with the ICC without it (0.525), it does not improve the ICC (See **Table 3** for details), but we confirmed that it improves the image quality.

TABLE 5 | Cross domain ICC performance. (Synthetic training set → Real test set).

	MultiPIE → FERA	FERA → DISFA	MultiPIE → DISFA	FERA → UNBC Pain	MultiPIE → UNBC Pain
AU01	0.311	0.314	0.418	-	-
AU04	0.202	0.400	0.541	0.130	0.149
AU06	0.786	0.573	0.524	0.496	0.434
AU10	0.726	-	-	0.034	0.038
AU12	0.792	0.748	0.698	0.402	0.367
AU14	0.168	-	-	-	-
AU17	0.365	0.373	0.290	-	-
Mean	0.479	0.482	0.494	0.266	0.247

TABLE 6 | ICC comparison of models trained on single vs. combined datasets for FERA 2017 Test partition and DISFA. Training size is the number of images per intensity per AU used to train models.

FERA 2017 Test partition

Train dataset	FERA real	FERA synthetic	FERA real + FERA synthetic	MultiPIE synthetic	FERA real + MultiPIE synthetic	FERA synthetic + MultiPIE synthetic
Train size	5,000	5,000	10,000	4,605	9,605	9,605
AU01	0.343	0.381	0.446	0.311	0.315	**0.457**
AU04	0.260	0.219	0.324	0.202	**0.342**	0.288
AU06	0.751	**0.804**	0.794	0.786	0.784	0.801
AU10	**0.785**	0.773	0.772	0.726	0.760	0.763
AU12	0.806	0.795	0.800	0.792	**0.807**	0.801
AU14	0.084	**0.244**	0.171	0.168	0.116	0.235
AU17	0.391	**0.461**	0.433	0.365	0.408	0.436
Mean	0.489	0.525	0.534	0.479	0.505	**0.540**

DISFA

Train dataset	FERA real	FERA synthetic	FERA real + FERA synthetic	MultiPIE synthetic	FERA real + MultiPIE synthetic	FERA synthetic + MultiPIE synthetic
Train size	5,000	5,000	10,000	4,605	9,605	9,605
AU01	0.394	0.314	0.365	0.418	**0.470**	0.346
AU04	**0.634**	0.400	0.571	0.541	0.544	0.402
AU06	0.404	0.573	0.507	0.524	0.496	**0.576**
AU12	0.750	0.748	0.723	0.698	0.749	**0.762**
AU17	0.293	0.373	0.296	0.290	0.377	**0.390**
Mean	0.495	0.482	0.493	0.494	**0.527**	0.495

The best results are shown in bold.

TABLE 7 | ICC comparison for GAN architectures.

	GANimation	StarGAN	GANimation internal classifier
AU1	**0.381**	0.367	0.380
AU4	0.219	**0.259**	0.065
AU6	**0.804**	0.793	0.712
AU10	0.773	**0.788**	0.743
AU12	0.795	**0.807**	0.793
AU14	**0.244**	0.199	0.123
AU17	**0.461**	0.451	0.364
Mean	**0.525**	0.523	0.454

The best results are shown in bold.

To evaluate the image quality, we performed three quantitative analyses: Frechet Inception Distance (FID), Structural Similarity Index (SSIM), and User Preference. The Frechet Inception Distance (FID) (Heusel et al., 2017) for the synthetic dataset with idiosyncratic loss (4.88) is better than the one without idiosyncratic loss (5.94). The Structural Similarity Index (SSIM) (Wang et al., 2004) for the one with idiosyncratic loss (0.835) is also better than the one without idiosyncratic loss (0.824). Note that a lower FID is better, and a higher SSIM is better. For User Preference, 20 human subjects were asked to pick an image with higher quality for randomly selected 20 pairs of images with and without idiosyncratic loss. The User Preference for images with idiosyncratic loss (52.5%) is higher than those without idiosyncratic loss (47.5%).

4.5 Temporal Normalization and Comparison With State of the Art

Results from the previous experiment suggest that precise spatial alignment improves the performance. As AUs are temporal, we decided to test the best method under different temporal normalization. We compared two methods that enhance the

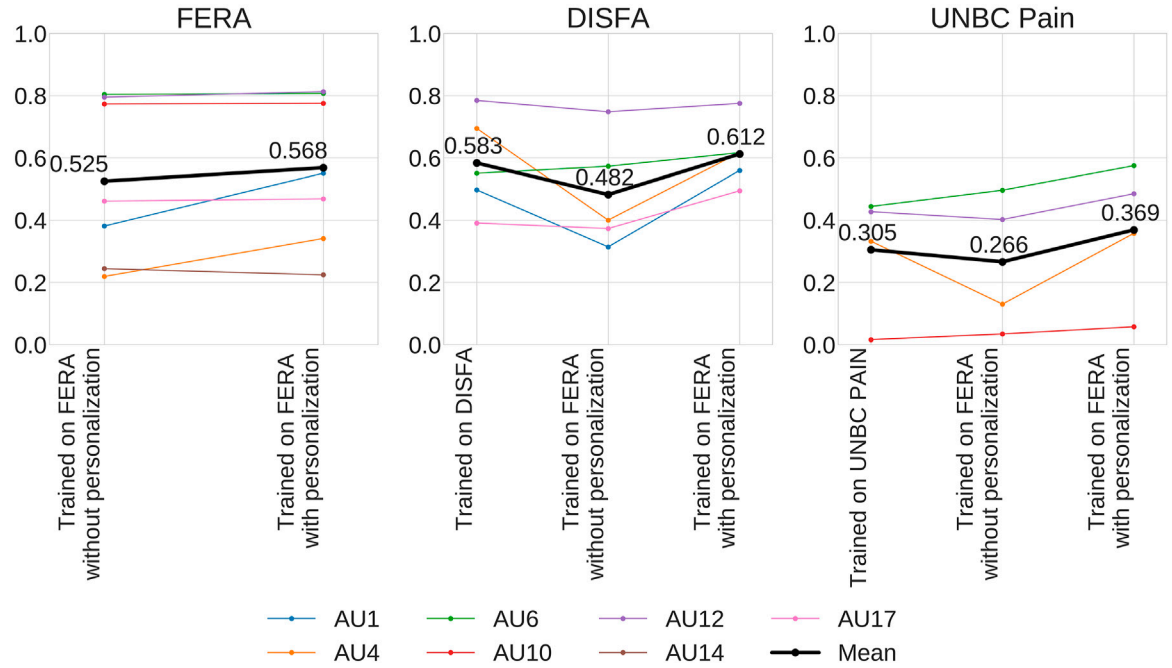

FIGURE 6 | Comparison of ICC between networks trained on the same-domain dataset (DISFA or UNBC Pain), networks trained on FERA without personalization, and networks trained on FERA with personalization. Personalized networks outperform the others.

temporal aspect of the AUs: AU0 normalization and mean texture normalization. AU0 normalization computes the appearance differences between the actual frame and a neutral frame. A neutral frame is not necessarily available in real life conditions, but we can assume that we have multiple frames from a single person. In personal mean texture normalization we calculate the average appearance of a person and then calculate the differences between each frame and the mean texture. This step minimizes individual differences in the appearance space. We trained and tested models using the appearance differences.

The last three columns of **Table 4** show the results. Personal mean texture normalization shows the best results. Although, on average, there is no gain with AU0 normalization, individual AU level differences are significant. While AU0 normalization shows better results for AU1 and AU4, it shows worse results for AU14 and AU17.

Table 4 shows the comparison with other state-of-the-art methods using only the frontal poses. We report Inter-rater reliability (ICC) that is the standard metric of the FERA 2017 benchmark. Our method (with or without temporal enhancement) outperforms all other methods. For a fair comparison, we compared our approach with existing methods' results on the frontal view of the test partition.

4.6 Experiment With Non-frontal Poses

Encouraged by the results of the previous experiment, we decided to evaluate the method's ability to generalize to unseen poses. FERA 2017 has nine different poses, and we report the performance on all of these using the Test partition. We investigated the following three scenarios:

1) real 2D normalized → real 2D normalized. For a baseline, we used real images with 2D alignment for training, and evaluated performance on 2D normalized real images from the testing set.
2) synthetic 3D normalized → real 3D normalized We trained on synthetic 3D normalized images, and tested on 3D normalized real images from the test set. We applied the 3D normalization procedure described in **Section 3.1** to each test image with a non-frontal pose. Self-occluded facial parts were filled with black color during the rasterization step.
3) 3D augmented synthetic → real 2D normalized We synthesized 3D meshes and rotated them into the nine standard orientations found in FERA 2017. We randomly selected 500 images for each intensity, AUs, poses. The total number of images for each intensity each AU is 4,500 while 5,000 images are selected for scenarios (a) and (b). We tested the system on 2D aligned test images.

Figure 5 shows the results. While (a) and (b) show low ICCs when face poses are largely different from frontal views, the performance drop for (c) is much smaller. The results show that our synthesized images with 3D registration are also effective for non-frontal views by recreating non-frontal view images from the synthesized images. Note that: 1) the approaches in **Figure 5** use frontal view images only to train models while the methods in FERA17 challenges used images with all 9 poses to train models and 2) the reason why the performance on frontal view for (c) is worse than (a) and (b) is that only 500 frontal view images for each intensity each AU are used to train.

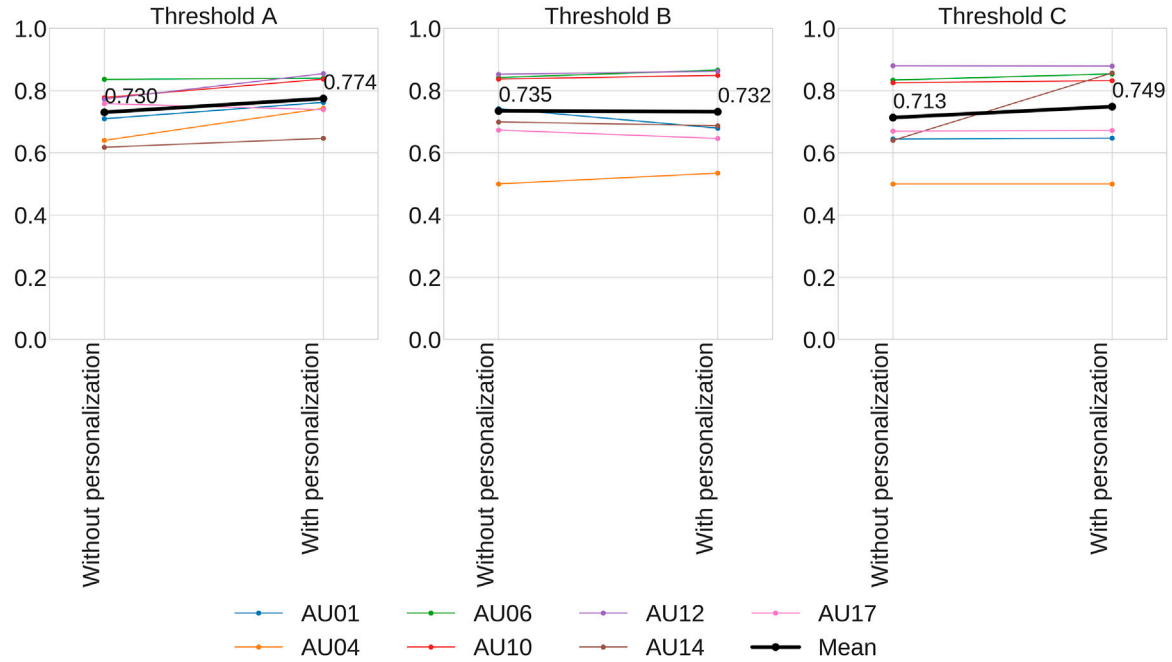

FIGURE 7 | Comparison of 2AFC for occurrence detection between with and without personalization on FERA 2017 Test partition. We evaluated 3 different thresholds. "Threshold A" means that the AU is present when the intensity is A-level or higher.

4.7 Cross-Domain Experiments

We have learned from the previous experiments that our AU classifiers can perform well when trained and tested within the same domain. To evaluate the generalizability of our approach to unseen domains (both in generating expressions and evaluating classifiers), we conducted two sets of experiments.

For the first experiment, we were interested in how generating out-of-domain samples would affect the performance on the FERA 2017 Test partition. In this case we trained the generator network on the FERA 2017 Train partition, but we synthesized new expressions using high resolution frontal images present in the MultiPIE dataset. We selected 921 MultiPIE images and generated 5 target expressions for each image, resulting in 4,605 images for each intensity and each AU. We trained the classifier on these images and tested the performance on FERA 2017 Test partition.

In the second cross-domain experiment, we evaluated out-of-domain classification. Here classifiers were trained either on synthetic expressions generated from FERA 2017 or synthetic expressions generated from MultiPIE, and they were tested on DISFA and UNBC Pain. The DISFA and UNBC Pain datasets differ in imaging condition and type of AU coding: context is not social in DISFA while in FERA subjects are interacting with the experimenter, and in DISFA, the base rates of most AUs is very low and limited to what occurs in a film-watching paradigm (Ertugrul et al., 2020). In the UNBC Pain dataset, facial expressions are mostly associated with pain, and the correlation among AUs differs from that of FERA and DISFA. In addition, the image size of UNBC Pain (320 × 240 or 352 × 240) is smaller than the other two datasets (FERA 2017: 1024 × 1024, and DISFA: 1024 × 768). These differences may cause the low ICC on UNBC Pain.

In all of these experiments, the FERA17 Train partition was used to train facial expression generation models, and 3D normalization was applied to each image. The whole dataset of DISFA or UNBC Pain was used to test the models.

Table 5 shows the results. Performance of models trained on synthesized MultiPIE expressions (0.479) is lower than the one trained on synthesized FERA 2017 expression (0.525), but there is only 1% difference with the one trained on real FERA17 expressions (0.489). The results for DISFA shows that the result trained on synthesized MultiPIE expressions (0.494) is slightly better than the one trained on synthesized FERA 2017 expressions (0.482) while the results for UNBC Pain shows that the result trained on synthesized FERA 2017 (0.266) is slightly better than the one trained on synthesized MultiPIE expressions (0.247).

4.8 Experiments for Models Trained on Combined Datasets

In this section, we report results for models trained on combined datasets. **Table 6** show the results for FERA Test partition and DISFA. The first, second and forth columns show the results for models trained on single datasets, and the third, fifth and sixth columns show results for models trained on combined datasets. As shown in the table, combined datasets show slightly better results.

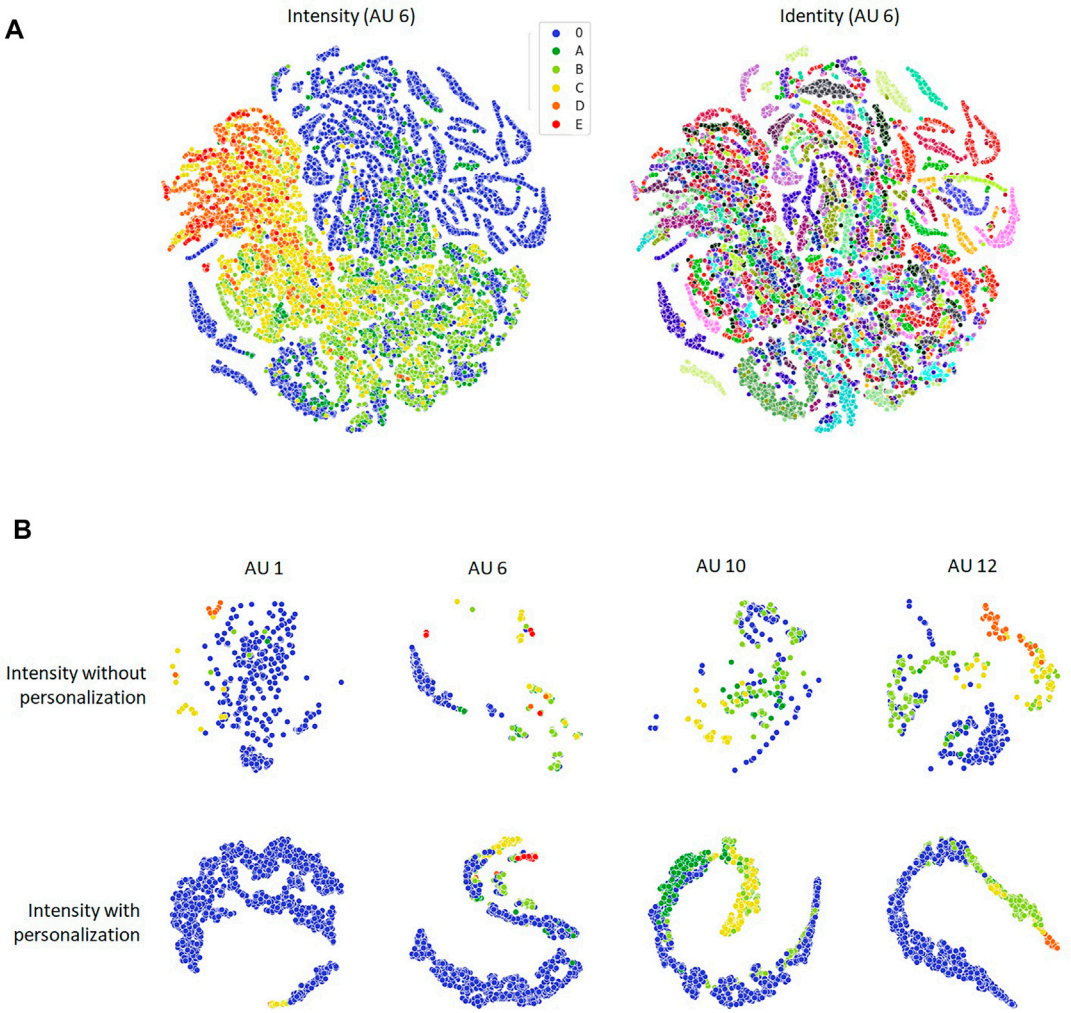

FIGURE 8 | t-SNE visualization. **(A)** Separation of our model trained without personalization for AU 6 in terms of AU intensity and identity. **(B)** Intensity separation for AU 1, 6, 10 and 12 between models without and with personalization for the same subject.

4.9 Comparison of GAN Architectures

To examine the influence of GAN architectures, we conducted experiments for GANimation, StarGAN and GANimation internal classifier D_{exp}. **Table 7** indicates that GANimation and StarGAN show almost the same performance, but the GANimation internal classifier performs worse than the others.

4.10 The Effect of Personalized Networks

We performed two sets of experiments to evaluate the effect of our personalized networks.

In the first set of experiments, we evaluated ICC on the FERA 2017, DISFA, and UNBC Pain datasets. **Figure 6** shows results for three types of networks: 1) networks trained on the same domain (DISFA or UNBC Pain), 2) networks trained on FERA without personalization, and 3) networks trained on FERA with personalization. To evaluate the networks trained on the same-domain dataset (DISFA or UNBC Pain), we performed three-fold cross-validation. To create the train datasets for each validation, we randomly selected 5,000 images from the six intensity classes, and generated AU balanced train datasets as described in **Section 4.2.2**. **Figure 6** does not include results of networks trained on the same-domain dataset for FERA because the other networks were also trained on FERA. Networks trained on FERA without personalization are the same AU classifiers reported in **Section 4.5**, and networks trained on FERA with personalization were trained by fine-tuning the person-independent networks according to the steps described in **Section 4.2.3**.

Figure 6 shows that the personalized networks outperform not only the networks trained on FERA without personalization but also the networks trained on the same-domain datasets with the test datasets. The differences between with and without personalization is especially large on the cross-domain situation (DISFA and UNBC Pain): 13% on DISFA, and 10.3% on UNBC Pain. The results indicate that our personalized

networks can greatly reduce the domain-shift problem even though only a single image is required from each target subject to synthesize person-specific images. **Figure 6** also shows that the effect of personalized networks varies depending on the AU. We can see that the personalized networks are especially effective for AU 1 and AU 4.

In the second set of experiments, we examined sensitivities of the models by evaluating occurrence detection for different thresholds. **Figure 7** shows AU occurrence detection performance of models trained with and without personalization on FERA 2017 Test partition. Three different thresholds are evaluated. In the figure, "Threshold A" means that the AU is present when the intensity is A-level or higher. In this experiments, we used 2AFC because the metric is robust to imbalanced data while F1 is not (Jeni et al., 2013). The 2AFC is a good approximation of the area under the receiver operator characteristic curve (AUC) (Valstar et al., 2017). **Figure 7** demonstrates that the performance of the models trained with personalization outperforms the one without personalization.

4.11 t-SNE Visualization

To visualize the separability of the proposed method, we performed t-SNE visualization. In this experiment, we extracted the features of the second to last layer (4,096 dimensions) from each image in FERA 2017 Test partition, and perform t-SNE on them. **Figure 8A** shows the separation of our model trained without personalization for AU 6 in terms of AU intensity and identity. The figure shows that the separation between intensities is clear while the one between identities is not, which implies that our model is correctly trained to estimate AU intensities. **Figure 8B** compare the AU intensity separation between models without and with personalization for the same subject. The figure shows that the output of the models with personalization is more structured and more separable than the one without personalization.

5 CONCLUSION

We have proposed a generative approach that achieves 3D geometry based AU manipulation to synthesize facial expressions. Generating expressions using the 3D registered facial images gives better AU intensity estimation performance compared to using 2D registered ones. Moreover, our proposed idiosyncratic loss has improved the visual quality of the outputs. The synthesized images are used for two goals: semantic resampling and network personalization.

With the semantic resampling, our approach provides a balanced distribution of AU intensity labels, which is crucial to train AU intensity estimators. We have shown that using the balanced synthetic set for training performs better than using the real training dataset on the same test set. Cross-pose and cross-domain results reveal that classifiers trained on our synthesized images are also effective on non-frontal views and on unseen domains. Network personalization is performed to tackle domain shift. We generate synthetic expressions from a single image of each target subject, and use the synthesized images to personalize the networks. Our personalized networks outperform not only the person-independent networks but also networks trained on the same-domain datasets with the target datasets.

Future improvements will include tackling training facial images with a non-frontal view. In our current approach, face images in Train datasets need to be near frontal for accurate 3D normalization. We also plan to expand the data augmentation capacity of the proposed method to handle other challenges, such as occlusion and illumination.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by CMU Institutional Review Board. 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 design and writing. LJ contributed to 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, NIH awards NS100549 and MH096951, and NSF award CNS-1629716.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frsip.2022.861641/full#supplementary-material

REFERENCES

Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., and Lucey, S. (2017). "Using Synthetic Data to Improve Facial Expression Analysis with 3d Convolutional Networks," in International Conference on Computer Vision Workshops (ICCVW). doi:10.1109/iccvw.2017.189

Ambadar, Z., Schooler, J. W., and Cohn, J. F. (2005). Deciphering the Enigmatic Face: The Importance of Facial Dynamics in Interpreting Subtle Facial Expressions. *Psychol. Sci.* 16 (5), 403–410. doi:10.1111/j.0956-7976.2005.01548.x

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 International Conference on Automatic Face & Gesture Recognition (FG), 854–859. doi:10.1109/fg.2017.109

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 International Conference on Automatic Face & Gesture Recognition (FG), 868–871. doi:10.1109/fg.2017.111

Ben, X., Ren, Y., Zhang, J., Wang, S.-J., Kpalma, K., Meng, W., et al. (2021). "Video-based Facial Micro-expression Analysis: A Survey of Datasets, Features and Algorithms," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/tpami.2021.3067464

Cai, J., Meng, Z., Khan, A. S., Li, Z., O'Reilly, J., Han, S., et al. (2021). "Identity-free Facial Expression Recognition Using Conditional Generative Adversarial Network," in IEEE International Conference on Image Processing (ICIP). doi:10.1109/icip42928.2021.9506593

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). "Stargan: Unified Generative Adversarial Networks for Multi-Domain Image-To-Image Translation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2018.00916

Chu, W.-S., De la Torre, F., and Cohn, J. F. (2017). Selective Transfer Machine for Personalized Facial Expression Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 529–545. doi:10.1109/tpami.2016.2547397

Chu, W.-S., Torre, F. D. l., and Cohn, J. F. (2019). Learning Facial Action Units with Spatiotemporal Cues and Multi-Label Sampling. *Image Vis. Comput.* 81, 1–14. doi:10.1016/j.imavis.2018.10.002

Cohn, J. F., and Ekman, P. (2005). "Measuring Facial Action," in *The New Handbook of Methods in Nonverbal Behavior Research*, 9–64.

Ding, H., Sricharan, K., and Chellappa, R. (2018). "Exprgan: Facial Expression Editing with Controllable Expression Intensity," in *AAAI*.

Ekman, P., Friesen, W., and Hager, J. (2002). *Facial Action Coding System: Research Nexus*, 1. Network Research Information.

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," in *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*. doi:10.1109/tbiom.2020.2977225

Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018). "Joint 3d Face Reconstruction and Dense Alignment with Position Map Regression Network," in European Conference on Computer Vision (ECCV). doi:10.1007/978-3-030-01264-9_33

Geng, Z., Cao, C., and Tulyakov, S. (2019). "3d Guided fine-grained Face Manipulation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 9821–9830. doi:10.1109/cvpr.2019.01005

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image Vis. Comput.* 28, 807–813. doi:10.1016/j.imavis.2009.08.002

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). "Gans Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Neural Information Processing Systems (NIPS)*.

Jeni, L. A., Cohn, J. F., and Torre, F. D. L. (2013). "Facing Imbalanced Data Recommendations for the Use of Performance Metrics," in International Conference on Affective Computing and Intelligent Interaction (ACII). doi:10.1109/acii.2013.47

Jeni, L. A., Tulyakov, S., Yin, L., Sebe, N., and Cohn, J. F. (2016). "The First 3d Face Alignment in the Wild (3dfaw) challenge," in European Conference on Computer Vision Workshops (ECCVW), 511–520. doi:10.1007/978-3-319-48881-3_35

Kazemi, V., and Sullivan, J. (2014). "One Millisecond Face Alignment with an Ensemble of Regression Trees," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2014.241

Kim, D., and Song, B. C. (2021). "Contrastive Adversarial Learning for Person Independent Facial Emotion Recognition," in *AAAI*.

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *J. Machine Learn. Res.* 10, 1755–1758.

Kollias, D., Cheng, S., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). "Deep Neural Network Augmentation: Generating Faces for Affect Analysis," in International Journal of Computer Vision (IJCV). doi:10.1007/s11263-020-01304-3

Kollias, D., and Zafeiriou, S. (2019). "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and Arcface," in British Machine Vision Conference (BMVC).

Lee, M., Rudovic, O., Pavlovic, V., and Pantic, M. (2019). "Fast Adaptation of Deep Models for Facial Action Unit Detection Using Model-Agnostic Meta-Learning," in IJCAI Workshop on Artificial Intelligence in Affective Computing.

Li, W., Abtahi, F., Zhu, Z., and Yin, L. (2017). "Eac-net: A Region-Based Deep Enhancing and Cropping Approach for Facial Action Unit Detection," in International Conference on Automatic Face & Gesture Recognition (FG), 103–110. doi:10.1109/fg.2017.136

Li, Y., and Shan, S. (2021). "Meta Auxiliary Learning for Facial Action Unit Detection," in IEEE Transactions on Affective Computing. doi:10.1109/taffc.2021.3135516

Li, Y., Zeng, J., Shan, S., and Chen, X. (2019). "Self-supervised Representation Learning from Videos for Facial Action Unit Detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2019.01118

Liu, Z., Liu, D., and Wu, Y. (2020). "Region Based Adversarial Synthesis of Facial Action Units," in International Conference on Multimedia Modeling, 514–526. doi:10.1007/978-3-030-37734-2_42

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 International Conference on Automatic Face & Gesture Recognition (FG), 57–64. doi:10.1109/fg.2011.5771462

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Trans. Affective Comput.* 4, 151–160. doi:10.1109/t-affc.2013.4

McKeown, G., Sneddon, I., and Curran, W. (2015). Gender Differences in the Perceptions of Genuine and Simulated Laughter and Amused Facial Expressions. *Emot. Rev.* 7, 30–38. doi:10.1177/1754073914544475

Niinuma, K., Ertugrul, I. O., Cohn, J. F., and Jeni, L. A. (2021a). "Synthetic Expressions Are Better Than Real for Learning to Detect Facial Actions," in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 1248–1257. doi:10.1109/wacv48630.2021.00129

Niinuma, K., Ertugrul, I. O., Cohn, J. F., and Jeni, L. A. (2021b). Systematic Evaluation of Design Choices for Deep Facial Action Coding across Pose. *Front. Comp. Sci.* doi:10.3389/fcomp.2021.636094

Niu, X., Han, H., Shan, S., and Chen, X. (2019). "Multi-label Co-regularization for Semi-supervised Facial Action Unit Recognition," in *Advances in Neural Information Processing Systems*, 907–917.

Peng, G., and Wang, S. (2018). "Weakly Supervised Facial Action Unit Recognition through Adversarial Training," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2188–2196. doi:10.1109/cvpr.2018.00233

Pumarola, A., Agudo, A., Martinez, A., Sanfeliu, A., and Moreno-Noguer, F. (2019). "Ganimation: One-Shot Anatomically Consistent Facial Animation," in *International Journal of Computer Vision (IJCV)*. doi:10.1007/s11263-019-01210-3

Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., and Wang, H. (2018). *Geometry-contrastive gan for Facial Expression Transfer. arXiv preprint arXiv:1802.01822*.

Rudovic, O., Pavlovic, V., and Pantic, M. (2015). Context-sensitive Dynamic Ordinal Regression for Intensity Estimation of Facial Action Units. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 944–958. doi:10.1109/tpami.2014.2356192

Ruiz, A., Van de Weijer, J., and Binefa, X. (2015). "From Emotions to Action Units with Hidden and Semi-hidden-task Learning," in IEEE International Conference on Computer Vision (ICCV), 3703–3711. doi:10.1109/iccv.2015.422

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 Faces In-The-Wild challenge: Database and Results. *Image Vis. Comput.* 47, 3–18. doi:10.1016/j.imavis.2016.01.002

Song, L., Lu, Z., He, R., Sun, Z., and Tan, T. (2018). "Geometry Guided Adversarial Facial Expression Synthesis," in ACM international conference on Multimedia, 627–635. doi:10.1145/3240508.3240612

Song, S., Sanchez, E., Shen, L., and Valstar, M. (2021). "Self-supervised Learning of Dynamic Representations for Static Images," in International Conference on Pattern Recognition (ICPR). doi:10.1109/icpr48806.2021.9412942

Sun, X., Zeng, J., and Shan, S. (2021). "Emotion-aware Contrastive Learning for Facial Action Unit Detection," in Automatic Face & Gesture Recognition (FG). doi:10.1109/fg52635.2021.9666945

Tian, Y.-I., Kanade, T., and Cohn, J. F. (2001). Recognizing Action Units for Facial Expression Analysis. *IEEE Trans. Pattern Anal. Machine Intell.* 23, 97–115. doi:10.1109/34.908962

Valstar, M. F., Sánchez-Lozano, E., Cohn, J. F., Jeni, L. A., Girard, J. M., Zhang, Z., et al. (2017). "FERA 2017 - Addressing Head Pose in the Third Facial Expression Recognition and Analysis challenge," in International Conference on Automatic Face & Gesture Recognition (International Conference on), 839–847. doi:10.1109/fg.2017.107

Wang, C., Wang, S., and Liang, G. (2019). "Identity- and Pose-Robust Facial Expression Recognition through Adversarial Feature Learning," in ACM International Conference on Multimedia, 238–246. doi:10.1145/3343031.3350872

Wang, C., and Wang, S. (2018). "Personalized Multiple Facial Action Unit Recognition through Generative Adversarial Recognition Network," in ACM international conference on Multimedia, 302–310. doi:10.1145/3240508.3240613

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* 13, 600–612. doi:10.1109/tip.2003.819861

Wu, B., Lyu, S., Hu, B.-G., and Ji, Q. (2015). Multi-label Learning with Missing Labels for Image Annotation and Facial Action Unit Recognition. *Pattern Recognition* 48, 2279–2289. doi:10.1016/j.patcog.2015.01.022

Yang, H., Zhang, Z., and Yin, L. (2018). "Identity-adaptive Facial Expression Recognition through Expression Regeneration Using Conditional Generative Adversarial Networks," in International Conference on Automatic Face & Gesture Recognition (FG). doi:10.1109/fg.2018.00050

Yin, Y., Lu, L., Wu, Y., and Soleymani, M. (2021). "Self-supervised Patch Localization for Cross-Domain Facial Action Unit Detection," in International Conference on Automatic Face and Gesture Recognition (FG). doi:10.1109/fg52635.2021.9667048

Zeng, J., Chu, W.-S., De la Torre, F., Cohn, J. F., and Xiong, Z. (2015). Confidence Preserving Machine for Facial Action Unit Detection. in IEEE International Conference on Computer Vision (ICCV), 3622–3630. doi:10.1109/iccv.2015.413

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., and Horowitz, A. (2014). BP4D-spontaneous: a High-Resolution Spontaneous 3D Dynamic Facial Expression Database. *Image Vis. Comput.* 32, 692–706. doi:10.1016/j.imavis.2014.06.002

Zhang, Y., Dong, W., Hu, B.-G., and Ji, Q. (2018b). "Classifier Learning with Prior Probabilities for Facial Action Unit Recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5108–5116. doi:10.1109/cvpr.2018.00536

Zhang, Y., Dong, W., Hu, B., and Ji, Q. (2018a). "Weakly-supervised Deep Convolutional Neural Network Learning for Facial Action Unit Intensity Estimation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2314–2323. doi:10.1109/cvpr.2018.00246

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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3438–3446. doi:10.1109/cvpr.2016.374

Zhang, Z., Zhai, S., and Yin, L. (2018c). "Identity-based Adversarial Training of Deep Cnns for Facial Action Unit Recognition," in British Machine Vision Conference (BMVC), 1–13.

Zhao, K., Chu, W., and Martinez, A. M. (2018). "Learning Facial Action Units from Web Images with Scalable Weakly Supervised Clustering," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2090–2099. doi:10.1109/cvpr.2018.00223

Zhou, Y., Pi, J., and Shi, B. E. (2017). "Pose-independent Facial Action Unit Intensity Regression Based on Multi-Task Deep Transfer Learning," in International Conference on Automatic Face & Gesture Recognition (FG), 872–877. doi:10.1109/fg.2017.112

Zhu, X., Liu, Y., Li, J., Wan, T., and Qin, Z. (2018). Emotion Classification with Data Augmentation Using Generative Adversarial Networks. Pacific-Asia Conference on Knowledge Discovery and Data Mining , 349–360. doi:10.1007/978-3-319-93040-4_28

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.