International Journal of Computer Vision
https://doi.org/10.1007/s11263-020-01304-3

Deep Neural Network Augmentation: Generating Faces for Affect Analysis

Dimitrios Kollias[1] · **Shiyang Cheng**[2] · **Evangelos Ververas**[1] · **Irene Kotsia**[3] · **Stefanos Zafeiriou**[1]

Received: 31 October 2018 / Accepted: 5 February 2020

Abstract

This paper presents a novel approach for synthesizing facial affect; either in terms of the six basic expressions (i.e., anger, disgust, fear, joy, sadness and surprise), or in terms of valence (i.e., how positive or negative is an emotion) and arousal (i.e., power of the emotion activation). The proposed approach accepts the following inputs:(i) a neutral 2D image of a person; (ii) a basic facial expression or a pair of valence-arousal (VA) emotional state descriptors to be generated, or a path of affect in the 2D VA space to be generated as an image sequence. In order to synthesize affect in terms of VA, for this person, 600,000 frames from the 4DFAB database were annotated. The affect synthesis is implemented by fitting a 3D Morphable Model on the neutral image, then deforming the reconstructed face and adding the inputted affect, and blending the new face with the given affect into the original image. Qualitative experiments illustrate the generation of realistic images, when the neutral image is sampled from fifteen well known lab-controlled or in-the-wild databases, including Aff-Wild, AffectNet, RAF-DB; comparisons with generative adversarial networks (GANs) show the higher quality achieved by the proposed approach. Then, quantitative experiments are conducted, in which the synthesized images are used for data augmentation in training deep neural networks to perform affect recognition over all databases; greatly improved performances are achieved when compared with state-of-the-art methods, as well as with GAN-based data augmentation, in all cases.

Keywords Dimensional · Categorical affect · Valence · Arousal · Basic emotions · Facial affect synthesis · 4DFAB · Blendshape models · 3DMM fitting · DNNs · StarGAN · GANimation · Data augmentation · Affect recognition · Facial expression transfer

1 Introduction

Rendering photorealistic facial expressions from single static faces while preserving the identity information is an open research topic which has significant impact on the area of affective computing. Generating faces of a specific person with different facial expressions can be used in various applications, including face recognition (Cao et al. 2018; Parkhi et al. 2015), face verification (Sun et al. 2014; Taigman et al. 2014), emotion prediction, expression database generation, facial expression augmentation and entertainment.

This paper describes a novel approach that uses an arbitrary face image with a neutral expression and synthesizes a new face image of the same person, but with a different expression, generated according to a categorical or dimensional emotion representation model. This problem cannot be tackled using small databases with labeled facial expressions, as it would be really difficult to disentangle facial expressions and identity information through them. Our approach is based

Communicated by Xavier Alameda-Pineda, Elisa Ricci, Albert Ali Salah, Nicu Sebe, Shuicheng Yan.

✉ Dimitrios Kollias
dimitrios.kollias15@imperial.ac.uk

Shiyang Cheng
shiyang.c@samsung.com

Evangelos Ververas
e.ververas16@imperial.ac.uk

Irene Kotsia
I.Kotsia@mdx.ac.uk

Stefanos Zafeiriou
s.zafeiriou@imperial.ac.uk

[1] Department of Computing, Imperial College London, Queen's Gate, London SW7 2AZ, UK

[2] Samsung AI Center, Cambridge, UK

[3] Department of Computer Science, Middlesex University of London, London NW4 4BT, UK

Published online: 22 February 2020

on the analysis of a large 4D facial database, the 4DFAB (Cheng et al. 2018), which we appropriately annotated and used for facial expression synthesis on a given subject's face.

At first, a dimensional emotion model, in terms of the continuous variables, valence (i.e., how positive or negative is an emotion) and arousal (i.e., power of the emotion activation) (Whissell 1989; Russell 1978), has been used to annotate a large amount of 600,000 facial images. This model can represent, not only primary, extreme expressions, but also subtle expressions which are met in everyday human to human, or human to machine interactions. According to the adopted dimensional view, all emotions can be discriminated by their position in the resulting coordinate system, the 2D Valence-Arousal Space.

The advantage of this model in comparison to the categorical approach (six basic expressions plus neutral state) is that this can lead to a very accurate assessment of the actual emotional state; valence and arousal are emotion-underlying dimensions and are therefore able to distinguish between different internal states. Also the categorical model has the disadvantage that a user can have other feelings than the specific ones, which then have to be mapped on the model's categories; this leads to some distortion of the actual impression. Thus there is poorer resolution of the categorical model in characterizing emotionally ambiguous examples. On the contrary, this is not the case in the dimensional model in which each affective state is represented.

Secondly, a categorical emotion model, in terms of the six basic facial expressions (Anger, Disgust, Fear, Happiness, Sadness, Surprise), has been used, according to which 12,000 expressions from the 4DFAB were selected, including 2000 cases for each of the six basic expressions.

The proposed approach accepts:(i) a pair of valence-arousal values and synthesize the respective facial affect, (ii) a path of affect in the 2D VA space and synthesize a temporal sequence showing it, (iii) a value indicating the basic facial expression to be synthesized; a given neutral 2D image of a person is used in all cases to appropriately transfer the synthesized affect.

Section 2 refers to related work regarding facial expression synthesis, as well as data augmentation related methodologies. Section 3 presents materials and methods that are used in the current work. We describe the annotation and use of the 4DFAB database and the 3D Morphable Model that we utilize in our developments. Section 4 presents our approach, explaining in detail all steps used to synthesize affect on an image or image sequence. Section 5 mentions the categorical and dimensional databases, which are used by our approach.

An extensive experimental study is presented in Sect. 6. At first, a qualitative evaluation of the proposed approach is provided, also showing the achieved higher quality when compared to GAN-generated facial affect. Then, we use the synthesized facial images for data augmentation and train Deep Neural Networks over eight databases, annotated with either dimensional or categorical affect labels. We show that the achieved performance is much higher than(i) that obtained by the respective state-of-the-art methods, (ii) the performance of the same DNNs with data augmentation provided by the StarGAN and GANimation networks. A further comparison with GANs is performed, with the synthesized facial images being used, together with the original images, as DNN training and/or test data respectively; this also verifies the improved performance of our approach. An ablation study is also presented, illustrating the effect of data granularity and subjects' age on the performance of the proposed method. Finally, conclusions and future work are presented in Sect. 6.

The proposed approach includes many novel contributions. To the best of our knowledge, it is the first time that the dimensional model of affect is taken into account when synthesizing face images. As verified in the experimental study, the generated images are of high quality and realistic. All other methods produce synthesized faces according to the six basic, or a few more, expressions. We further show that the proposed approach can accurately synthesize the six basic expressions.

Moreover, it is the first time that a 4D face database is annotated in terms of valence and arousal and is then used for affect synthesis. The fact that this a temporal database ensures that successive video frames' annotations are adjacent in the VA space. Consequently, we generate temporal affect sequences on a given neutral face by using annotations that are adjacent in the VA space. Results are presented in the qualitative experimental study that illustrate this novel capability.

It should be also mentioned that the proposed approach works well, when presented with a neutral face image, obtained either in a controlled environment, or in-the-wild, e.g., irrespective of the head pose of the person appearing in the image.

An extensive experimental study is provided, over most significant databases with affect, showing that the developed DNNs based on the proposed facial affect synthesis approach outperform the existing state-of-the-art, as well the same DNNs based on facial affect synthesis produced by GAN architectures.

2 Related Work

Facial expression transfer is a research field for mapping and generating desired images of specified subject and facial expression. Many methods achieved significant results for high-resolution images and are applied to a wide range of applications, such as facial animation, facial editing, and facial expression recognition.

There are mainly two categories of methods for facial expression transfer from a single image: traditional graphic-based methods and emerging generative methods. In the first case, some methods directly warp the input face to create the targeted expression, by either 2D warps (Fried et al. 2016; Garrido et al. 2014), or 3D warps (Blanz et al. 2003; Cao et al. 2014; Liu et al. 2008). Other methods construct parametric global models. In Mohammed et al. (2009), a probabilistic model is learned, in which existing and generated images obey structural constraints. Averbuch-Elor et al. (2017) added fine-scale dynamic details, such as wrinkles and inner mouth, that are associated with facial expressions. Although these methods have achieved some positive results in high-resolution and one-to-many image synthesis, they are still limited due to their sophisticated design and expensive computation.

Thies et al. (2016) developed a real-time face-to-face expression transfer system, with an extra blending step for mouth. This 2D-to-3D approach shows promising results, but due to the nature of its formulation, it is unable to retrieve fine-details, and its applicability is limited to expressions lying in a linear shape subspace with known rank. The authors extended this system to human portrait video transfer (Thies et al. 2018). They captured facial expressions, eye gaze, rigid head pose, and motions of the upper body of a source actor and transferred them to a target actor in real time.

The second category of methods is based on data-driven generative models. At the beginning, some generative models, such as deep belief nets (DBN) (Susskind et al. 2008) and higher-order Boltzmann machines (Reed et al. 2014), had been applied to facial expression synthesis. However, these models faced problems such as blurry generated images, incapability of fine control of facial expression and low-resolution outputs.

With the recent development of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), these networks have been applied to facial expression transfer; due to the fact that the generated images are of high-quality, these provided positive results. A generative model is trained according to a dataset, including all information about identity, expression, viewing angle, etc, while performing facial expression transfer. Generative modeling methods reduce the complicated design of the connection between facial textures and emotional states and encode intuitionistic facial features into parameters of data distribution. However, the main drawback of GANs is the training instability and the trade-off between visual quality and image diversity.

Since the original GAN could not generate facial images with a specific facial expression referring to a specific person, some methods conditioned on expression categories have been proposed. Conditional GANs (cGANs) (Mirza and Osindero 2014) (and conditional variational autoencoders (cVAEs) Sohn et al. 2015) can generate samples conditioned on attribute information, when this is available. Those networks require large training databases so that identity information can be properly disambiguated. Otherwise, when presented with an unseen face, the networks tend to generate faces which look like the "closest" subject in the training datasets. During training, those networks require the knowledge of the attribute labels; it is not clear how to adapt them to new attributes without retraining from scratch. Finally, these networks suffer from mode-collapse (e.g., the generator only outputs samples from a single mode, or with extremely low variety) and blurriness.

The conditional difference adversarial autoencoder (CDAAE) (Zhou and Shi 2017) aims at synthesizing specific expressions for unseen persons with a targeted emotion or facial action unit label. However, such GAN-based methods are still limited to discrete facial expression synthesis, i.e., they cannot generate a face sequence showing a smooth transition from an emotion to another. Ding et al. (2018) proposed an Expression Generative Adversarial Network (ExprGAN) in which the expression intensity could be controlled in a continuous manner from weak to strong. The identity and expression representation learning were disentangled and there was no rigid requirement of paired samples for training. The authors developed a three-stage incremental learning algorithm to train the model on small datasets.

Pham et al. (2018) proposed a weakly supervised adversarial learning framework for automatic facial expression synthesis based on continuous action unit coefficients. In Pumarola et al. (2018), the GANimation was proposed that additionally controlled the generated expression by AU labels, and allowed a continuous expression transformation. The authors introduced an attention-based generator to promote the robustness of their model for distracting backgrounds and illuminations.

There are some differences between continuous expression synthesis based on AUs and VA. Firstly, AUs are related to some facial muscles, with only a small number of them being mapped to facial expression modelling. On the contrary, the VA model covers the whole spectrum of emotions. Moreover, mapping AUs to emotions is not straightforward (different psychological studies provide different results). GANimation is solely based on automatic annotation of AUs, whilst the proposed methodology is based on manual, i.e., more robust and trusted, VA annotation of the 4DFAB database. Finally, it can be mentioned that annotation of AUs needs experienced FACS coders; especially in in-the-wild datasets. That is why, there exists only one in-the-wild database annotated for AUs (existence and not intensity information), the EmotioNet, which only contains 50,000 annotations, in terms of 12 AUs.

Recently, Song et al. (2018) utilized landmarks and proposed the geometry-guided GAN (G2GAN) to generate smooth image sequences of facial expressions. G2GAN uses

geometry information based on dual adversarial networks to express face changes and synthesizes facial images. Through manipulating landmarks, smoothly changed images can also be generated. However, this method demands a neutral face of the targeted person as the intermediate of facial expression transfer. Although the expression removal network could generate a neutral expression of a specific person, this procedure brings additional artifacts and degrades the performance of expression transition.

Qiao et al. (2018) used geometry (facial landmarks) to control the expression synthesis with a facial geometry embedding network and proposed a Geometry-Contrastive Generative Adversarial Network (GC-GAN) to transfer continuous emotions across different subjects, even if there existed big difference in shapes. Wu et al. (2018) proposed a boundary latent space and boundary transformer. They mapped the source face into the boundary latent space, and transformed the source face's boundary to the target's boundary, which was the medium to capture facial geometric variances during expression transfer.

In Ma and Deng (2019), an unpaired learning framework was developed to learn the mapping between any two facial expressions in the facial blendshape space. This framework automatically transforms the source expression in an input video clip to a specified target expression. This work lacks the capability to generate personalized expressions; individual-specific expression characteristics, such as wrinkles and creases, are ignored. Also, the transitions between different expressions are not taken into consideration. Finally, this work is limited in the sense that it cannot produce highly exaggerated expressions.

Both the graphic-based methods and the genererative methods of facial expression transfer have been used to create synthetic data that are used as auxiliary data in network training, augmenting the training dataset. A synthetic data generation system with a 3D convolutional neural network (CNN) was created in Abbasnejad et al. (2017) to confidentially create faces with different levels of saturation in expression. Antoniou et al. (2017) proposed the Data Augmentation Generative Adversarial Network (DAGAN) which is based on cGAN and tested its effectiveness on vanilla classifiers and one shot learning. DAGAN is a basic framework for data augmentation based on cGAN.

Zhu et al. (2018) presented another basic framework for face data augmentation based on CycleGAN (Zhu et al. 2017). Similar to cGAN, CycleGAN is also an general-purpose solution for image-to-image translation, but it learns a dual mapping between two domains simultaneously with no need for paired training examples, because it combines a cycle consistency loss with adversarial loss. The authors used this framework to generate auxiliary data for imbalanced datasets, where the data class with fewer samples was selected as transfer target and the data class with more samples was the reference.

3 Materials and Methods

In the following, we first describe the 4DFAB database, its annotation in terms of valence-arousal and the selection of expressive categorical sequences from it. The annotated 4DFAB database has been used for constructing the 3D facial expression gallery that is the basis of our affect synthesis pipeline described in the next Section. Then we describe the methods we have used: a) for registering and correlating all components of the 3D gallery into a universal coordinate frame; b) for constructing the 3D Morphable Model used in this work.

3.1 The 4DFAB Database

The 4DFAB database (Cheng et al. 2018) is the first large scale 4D face database designed for biometric applications and facial expression analysis. It consists of 180 subjects (60 females, 120 males) aging from 5 to 75 years. 4DFAB was collected over a period of 5 years under four different sessions, with over 1,800,000 3D faces. The database was designed to capture articulated facial actions and spontaneous facial behaviors, the latter being elicited by watching emotional video clips. In each of the four sessions, different video clips were shown that stimulated different spontaneous behaviors. In this paper, we use all 1580 spontaneous expression sequences (video clips) for dimensional emotion analysis and synthesis. The frame rate of 4DFAB database is 60 FPS and the average clip length for spontaneous expression sequences is 380 frames. Consequently the 1580 expression sequences correspond to 600,000 frames, which we annotated in terms of valence and arousal (details follow in the next subsection). These sequences cover a wide range of expressions as shown in Figs. 2 and 3.

Moreover, to be able to develop the categorical emotion synthesis model, we used the 2000 expressive 3D meshes per basic expression (12,000 meshes in total) that were provided along with 4DFAB. Those 3D meshes corresponded to (annotated) apex frames of posed expression sequences in 4DFAB. Such examples are shown in Fig. 1.

3.2 4DFAB Dimensional Annotation

Targeting to develop the novel dimensional expression synthesis method, all 1580 dynamic 3D sequences (i.e., over 600,000 frames) of 4DFAB have been annotated in terms of valence and arousal emotion dimensions. In total, three experts were chosen to perform the annotation task. Each expert performed a time-continuous annotation for

Fig. 1 Examples from the 4DFAB of apex frames with posed expressions for the six basic expressions: Anger (AN), Disgust (DI), Fear (FE), Joy (J), Sadness (SA), Surprise (SU)

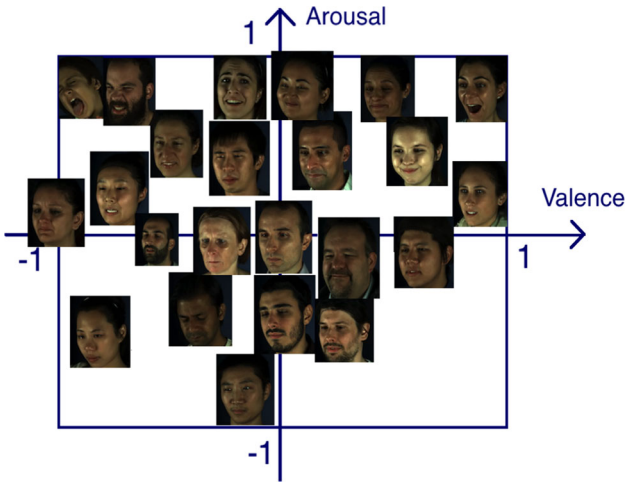


Fig. 2 The 2D valence-arousal space and some representative frames of 4DFAB

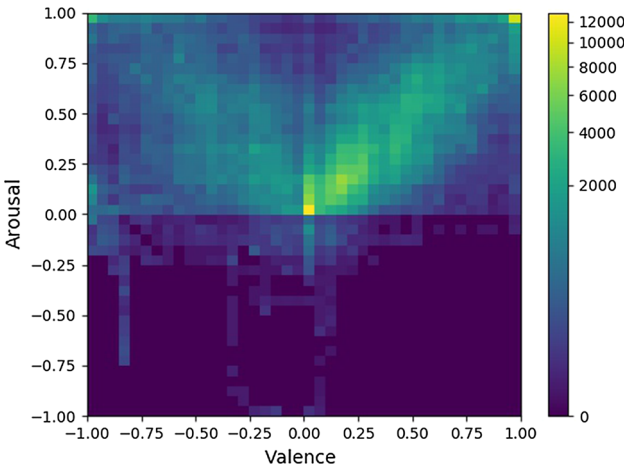


Fig. 3 The 2D histogram of annotations of 4DFAB

both affective dimensions. The application-tool described in Zafeiriou et al. (2017), was used in the annotation process.

Each expert logged into the application-annotation tool using an identifier (e.g. his/her name) and selected an appropriate joystick; then the application showed a scrolling list of all videos and the expert selected a video to annotate; then a screen appeared that showed the selected video and a slider of valence or arousal values ranging in $[-1, 1]$; the expert annotated the video by moving the joystick either up or down; finally, a file was created with the annotations. The mean inter-annotation correlation per annotator was 0.66, 0.70, 0.68 for valence and 0.59, 0.62, 0.59 for arousal. The average of those mean inter-annotation correlations was 0.68 for valence and 0.60 for arousal. Those values are high, indicating a very good agreement between annotators. As a consequence, the final label values were chosen to be the mean of those three annotations.

Examples of frames from the 4DFAB along with their annotations, are shown in Fig. 2. Figure 3 shows the 2D histogram of annotations of 4DFAB. In the rest of the paper, we refer to the 4DFAB database either as: (i) the 600,000 frames with their corresponding 3D meshes, which have been annotated with 2D valence and arousal (VA) emotion values or (ii) the 12,000 apex frames of posed expressions with their corresponding 3D meshes, which have categorical annotation.

3.3 Mesh Pre-Processing: Establishing Dense Correspondence

Each 3D mesh is first re-parameterized into a consistent form where the number of vertices, the triangulation and the anatomical meaning of each vertex are made consistent across all meshes. For example, if the vertex with index i in one mesh corresponds to the nose tip, it is required that the vertex with the same index in every mesh corresponds to the nose tip too. Meshes satisfying the above properties are said to be in dense correspondence with one another. So, correlating all these meshes with a universal coordinate frame (viz. a 3D face template) is a step to follow so as to establish dense correspondence.

In order to do so, we need to define a 2D *UV* space for each mesh, which in fact is a contiguous flattened atlas that embeds the 3D facial surface. Such a UV space is associated with its corresponding 3D surface through a bijective mapping; thus, establishing dense correspondence between two UV images implicitly establishes a 3D-to-3D correspondence for the mapped mesh. UV mapping is the 3D modelling process of projecting a 2D image to a 3D model's surface for texture mapping. The letters U and V denote the axes of the 2D

texture, since *X, Y* and *Z* are already taken to denote the axes of the 3D object in model space.

We employ an optimal cylindrical projection method (Booth and Zafeiriou 2014) to synthetically create a UV space for each mesh. A UV map (which is an image I), with each pixel encoding both spatial information (X, Y, Z) and texture information (R, G, B), is produced, on which we perform non-rigid alignment. Non-rigid alignment is performed through the UV-TPS method that utilises key landmarks fitting and Thin Plate Spline (TPS) warping (Cosker et al. 2011). Following Cheng et al. (2018), we perform several modifications to Cosker et al. (2011), to suit our data. Firstly, we build session-and-person-specific Active Appearance Models (AAMs) (Alabort-i Medina and Zafeiriou 2017) to automatically track feature points in the UV sequences. This means that 4 different AAMs are built and used separately for one subject. Main reasons behind this are:(i) textures of different sessions differ due to several facts (i.e. aging, beards, make-ups, experiment lighting condition), (ii) person-specific model is proven more accurate and robust in specific domains (Chew et al. 2012).

In total, 435 neutral meshes and 1047 expression meshes (1 neutral and 2–3 expressive meshes per person and session) in 4DFAB were selected; these contained annotations with 79 3D landmarks. They were unwrapped and rasterised to UV space, then grouped for building the corresponding AAMs. Each UV map was flipped to increase fitting robustness. Once all the UV sequences were tracked with 79 landmarks, they were then warped to the corresponding reference frame using TPS, thus achieving the 3D dense correspondence. For each subject and session, one specific reference coordinate frame from his/her neutral UV map was built. From each warped frame, we could uniformly sample the texture and 3D coordinates. Eventually, a set of non-rigidly corresponded 3D meshes under the same topology and density were obtained.

Given that meshes have been aligned to their designated reference frame, the last step was to establish dense 3D-to-3D correspondences between those reference frames and a 3D template face. This is a 3D mesh registration problem, solved by Non-rigid ICP (Amberg et al. 2007). We employed it to register the neutral reference meshes to a common template, the Large Scale Facial Model (LSFM) (Booth et al. 2018). We brought all 600,000 3D meshes into full correspondence with the mean face of LSFM. As a result, we created a new set of 600,000 3D faces that share identical mesh topology, while maintaining their original facial expressions. In the following, this set constitutes the 3D facial expression gallery which we use for facial affect synthesis.

3.4 Constructing a 3D Morphable Model

3.4.1 General Pipeline

A common 3DMM consists of three parametric models: the shape, the camera and the texture models.

To build the shape model, the training 3D meshes should be put in dense correspondence (similarly to the previous Mesh Pre-Processing subsection). Next, Generalized Procrustes Analysis is performed to remove any similarity effects, leaving only shape information. Finally, Principal Component Analysis (PCA) is applied to these meshes, which generates a 3D deformable model as a linear basis of shapes. This model allows for the generation of novel shape instances. The model can be expressed as:

$$\mathcal{S}(\mathbf{p}) = \bar{\mathbf{s}} + \mathbf{U}_s\mathbf{p} \tag{1}$$

where $\bar{\mathbf{s}} \in \mathbb{R}^{3N}$ is the mean component of 3D shape (in our case it is the mean of shape models from the LSFM model described in the next subsection) with N denoting the number of vertices in the shape model; $\mathbf{U}_s \in \mathbb{R}^{3N\times n_s}$ is the shape eigenbase (in our case it is the identity subspace of LSFM) with $n_s << 3N$ being the number of principal components (n_s is chosen to explain a percentage of the training set variance; generally, this percentage is 99.5%); and $\mathbf{p} \in \mathbb{R}^{n_s}$ is a vector of parameters which allows for the generation of novel shape instances.

The purpose of camera model is to project the object-centered Cartesian coordinates of a 3D mesh instance into 2D Cartesian coordinates in an image plane. At first, given that the camera is static, the 3D mesh is rotated and translated using a linear view transformation, which results in 3D rotation and translation components. Then, a nonlinear perspective transformation is applied. Note that quaternions (Kuipers et al 1999; Wheeler and Ikeuchi 1995) are used to parametrise the 3D rotation, which ensures computational efficiency, robustness and simpler differentiation. In this manner we construct the camera parameters (i.e., 3D translation components, quaternions and parameter of linear perspective transformation). The camera model of the 3DMM applies the above transformations on the 3D shape instances generated by the shape model. Finally, the camera model can be written as:

$$\mathcal{W}(\mathbf{p}, \mathbf{c}) = \mathcal{P}(\mathcal{S}(\mathbf{p}), \mathbf{c}), \tag{2}$$

where $\mathcal{S}(\mathbf{p})$ is a 3D face instance; $\mathbf{c} \in \mathbb{R}^{n_c}$ are the camera parameters (for rotation, translation and focal length; n_c is 7); and $\mathcal{P} : \mathbb{R}^{3N} \rightarrow \mathbb{R}^{2N}$ is the perspective camera projection.

For the texture model, large facial "in-the-wild" databases annotated for sparse landmarks are needed. Let us assume that the meshes have corresponding camera and

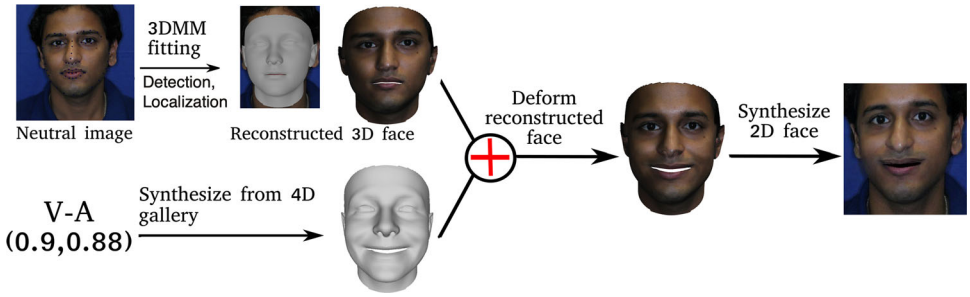


Fig. 4 The facial affect synthesis framework: the user inputs an arbitrary 2D neutral face and the affect to be synthesized (a pair of valence-arousal values in this case)

shape parameters. These images are passed through a dense feature extraction function that returns feature-based representations for each image. These are then sampled from the camera model at each vertex location so as to build a texture sample, which will be nonsensical for some regions mainly due to self occlusions present in the mesh projected in the image space. To complete the missing information of the texture samples, Robust PCA (RPCA) with missing values (Shang et al. 2014) is applied. This produces complete feature-based textures that can be processed with PCA to create the statistical model of texture, which can be written as:

$$\mathcal{T}(\boldsymbol{\lambda}) = \bar{\mathbf{t}} + \mathbf{U_t}\boldsymbol{\lambda}, \tag{3}$$

where $\bar{\mathbf{t}} \in \mathbb{R}^{3N}$ is the mean texture component (in our case it is the mean of texture model from LSFM); $\mathbf{U}_t \in \mathbb{R}^{3N \times n_t}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{n_t}$ are the texture subspace (eigenbase) and texture parameters, respectively, with $n_t << 3N$ being the number of principal components. This model can be used to generate novel 3D feature-based texture instances.

3.4.2 The Large Scale Facial Model (LSFM)

We have adopted the LSFM model constructed using the MeIn3D dataset (Booth et al. 2018). The construction pipeline of LSFM starts with a robust approach to 3D landmark localization resulting in generating 3D landmarks for the meshes. The 3D landmarks are then employed as soft constraints in Non-rigid ICP to place all meshes in correspondence with a template facial surface; the mean face of the Basel Face Model (Paysan et al. 2009) has been chosen. However, the large cohort of data could result in convergence failures. These are an unavoidable byproduct of the fact that both landmark localization and NICP are non-convex optimization problems sensitive to initialization.

A refinement post-processing step weeds out problematic subjects automatically, guaranteeing that the LSFM models are only constructed from training data for which there exist a high confidence of successful processing. Finally, the LSFM models are derived by applying PCA on the corresponding training sets, after excluding the shape vectors that have been classified as outliers. In total, 9663 subjects are used to build LSFM, which covers a wide variety of age (from 5 to over 80s), gender (48% male, 52% female), and ethnicity (82% White, 9% Asian, 5% Mixed Heritage, 3% Black and 1% other).

4 The Proposed Approach

In this section, we present the fully automatic facial affect synthesis framework. The user needs to provide a neutral image and an affect, which can be a VA pair of values, a path in the 2D VA space, or one of the basic expression categories. Our approach: (1) performs face detection and landmark localization on the input neutral image, (2) fits a 3D Morphable Model (3DMM) on the resulting image (Booth et al. 2017), (3) deforms the reconstructed face and adds the input affect, and (4) blends the new face with the given affect into the original image. Here let us note that the total time needed for the first two steps is about 400ms; this has to be performed only once, if generating multiple images from the same input image. Specific details regarding the described steps of our approach follow. This procedure is shown in Fig. 4.

4.1 Face Detection and Landmark Localization

The first step to edit an image is to locate landmark points that will be used for fitting the 3DMM. We first perform face detection with the face detection model from Zhang et al. (2016) and then utilize (Deng et al. 2018) to localize 68 2D facial landmark points which are aware of the 3D structure of the face, in the sense that points on occluded parts of the face (most commonly part of the jawline) are correctly localized.

4.2 3DMM-Fitting: Cost Function and Optimization

The goal of this step is to retrieve a reconstructed 3D face with the texture sampled from the original image. In order to do so, we first need a 3DMM; we select the LSFM.

Fitting a 3DMM on face images is an inverse graphics approach to 3D reconstruction and consists of optimizing three parametric models of the 3DMM, the *shape*, *texture* and *camera* models. The optimization aims at rendering a 2D

image which is as close as as possible to the input one. In our pipeline we follow the 3DMM fitting approach of Booth et al. (2017). As is already noted, we employ the LSFM (Booth et al. 2018) $\mathcal{S}(\mathbf{p})$ to model the identity deformation of faces. Moreover, we adopt the robust, feature-based texture model $\mathcal{T}(\boldsymbol{\lambda})$ of Booth et al. (2017), built from in-the-wild images. The employed camera model is a perspective transformation $\mathcal{W}(\mathbf{p}, \mathbf{c})$, which projects shape $\mathcal{S}(\mathbf{p})$ on the image plane.

Consequently, the objective function that we optimize can be formulated as:

$$\begin{aligned} &\text{argmin}_{\mathbf{p},\boldsymbol{\lambda},\mathbf{c}} \, \|\mathbf{F}(\mathcal{W}(\mathbf{p}, \mathbf{c})) - \mathcal{T}(\boldsymbol{\lambda})\|^2 + c_l \|\mathcal{W}_l(\mathbf{p}, \mathbf{c}) - \mathbf{s}_l\|^2 \\ &+ \; c_s \|\mathbf{p}\|^2_{\Sigma_s^{-1}} + c_t \|\boldsymbol{\lambda}\|^2_{\Sigma_t^{-1}}, \end{aligned} \tag{4}$$

where the first term denotes the pixel loss between the feature based image $\mathbf{F}$ sampled at the projected shape's locations and the model generated texture; the second term denotes a sparse landmark loss between the image 2D landmarks and the corresponding 2D projected 3D points, where the 2D shape, $\mathbf{s}_l$, is provided by Deng et al. (2018); the rest two terms are regularization terms which serve as counter overfitting mechanism, where $\mathbf{6}_s$ and $\mathbf{6}_t$ are diagonal matrices with the main diagonal being eigenvalues of the shape and texture models respectively; c_l, c_s and c_t are weights used to regularize the importance of each term during optimization and were empirically set to 10^5, 3×10^6 and 1, respectively, following Booth et al. (2017). Note also, that the 2D landmarks term is useful as it drives the optimization to converge faster. Problem of Eq. 4 is solved by the Project-Out variation of Gauss-Newton optimization as formulated in Booth et al. (2017).

From the optimized models, the optimal shape instance constitutes the neutral 3D representation of the input face. Moreover, by utilizing the optimal shape and camera models, we are able to sample the input image at the projected locations of the recovered mesh and extract a UV texture, that we later use for rendering.

4.3 Deforming Face and Adding Affect

Given an affect and an arbitrary 2D image $\mathbf{I}$, we first fit the LSFM to this image using the aforementioned 3DMM fitting method. After that, we can retrieve a reconstructed 3D face $\mathbf{s}_{orig}$ with the texture sampled from the original image (texture sampling is simply extracting image pixel value for each projected 3D vertex in image plane). Let us assume that we have created an affect synthesis model $\mathbf{M}_{Aff}$ that takes the affect as input and generates a new expressive face (denoted as $\mathbf{s}_{gen}$), i.e., $\mathbf{s} = \mathbf{M}_{Aff}(affect)$ (specific details regarding the generation of the expressive face, can be found in Sect. 4.5). Next, we calculate the facial deformation $\Delta\mathbf{s}$ by subtracting the synthesized face $\mathbf{s}_{gen}$ from the LSFM template $\bar{\mathbf{s}}$, i.e., $\Delta\mathbf{s} = \mathbf{s}_{gen} - \bar{\mathbf{s}}$, and impose this deformation on the reconstructed mesh, i.e., $\mathbf{s}_{new} = \mathbf{s}_{orig} + \Delta\mathbf{s}$. Therefore, we obtain a 3D face (dubbed $\mathbf{s}_{new}$) with facial affect.

4.4 Synthesizing 2D Face

The final step in our pipeline is to render the new 3D face $\mathbf{s}_{new}$ back to the original 2D image. To do that we employ the mesh that we have deformed according to the given affect, the extracted UV texture and the optimal camera transformation of the 3DMM. For rendering, we pass the three model instances to a renderer and we use as background the background of the input image. Lastly, the rendered image is fused with the original image via poisson blending (Pérez et al. 2003) to smooth the boundary between foreground face and image background so as to produce a natural and realistic result. In our experiments, we used both a CPU-based renderer (Alabort-i-Medina et al. 2014) and a GPU-based renderer (Genova et al. 2018). The GPU-based renderer greatly decreases the rendering time, as it needs 20 ms to render a single image, while the CPU-based renderer needs 400 ms.

4.5 Synthesizing Expressive Faces with Given Affect

4.5.1 VA and Basic Expression Cases: Building Blendshape Models and Computing Mean Faces

Let us first describe the VA case. We have 600,000 3D meshes (established into dense correspondence) and their VA annotations. We want to appropriately discretize the VA space into classes, so that each class contains a sufficient number of data. This is due to the fact that if classes contain only few examples, it is more likely to include identity information. However, the synthesized facial affect should only describe the expression associated with the VA pair of values, rather than information for the person's identity, gender, or age. We have chosen to perform agglomerative clustering (Maimon and Rokach 2005) on the VA values, using the euclidean distance as metric and the ward as linkage criterion (keeping the correspondence between VA values and 3D meshes). In this manner, we created 550 clusters, i.e., classes. Then we built blendshape models and computed the mean face per class. Figure 5 illustrates the mean faces of various classes. It should be mentioned that the majority of classes correspond to the first two quadrants of the VA space, namely the regions of positive valence (as can be seen in the 2D histogram of Fig. 3).

As far as the basic expression case is concerned, based on the derived 12,000 3D meshes, 2000 for each of the six basic expressions, we built six blendshape models and six corresponding mean faces.

Fig. 5 Some mean faces of the 550 classes in the VA space

4.5.2 User Selection: VA/Basic Expr and Static/Temporal Synthesis

The user first chooses the type of affect that our approach will generate. The affect could be either a point, or a path in the VA space, or one the six basic expression categories. If the user chooses the latter, then we retrieve the mean face of this category and add it on the 3D face reconstructed from the user's input neutral image. In this case, the only difference in Fig. 4 would be for the user to input a basic expression, the happy one, instead of a VA pair of values. If the user chooses the former, then (s)he needs to additionally clarify if our approach should generate an image ('static synthesis') or a sequence of images ('temporal synthesis') with this affect.

Static synthesis If the user selects 'static synthesis', then the user should input a specific VA pair of values. Then, we retrieve the mean face of the class to which this VA value belongs. We use this mean face as the affect to be added on the 3D face reconstructed from the provided neutral image. Figure 4 shows the proposed approach for this specific case. Figure 6 illustrates the procedure described in Sect. 4.5.1 given that the 550 VA classes are already created.

Temporal synthesis If the user selects 'temporal synthesis', then, (s)he should provide a path in the VA space (for instance by drawing) that the synthesized sequence should follow. Then, we retrieve the mean faces of the classes to which the VA values of the path belong. We use each of these mean faces as the affect to be added on the 3D faces reconstructed from the provided neutral image. As a consequence, an expressive sequence is generated that shows the evolution of affect on the VA path specified by the user.

Here let us mention that the fact that the 4DFAB used in our approach is a temporal database, ensures that successive video frames' annotations are adjacent in the VA space, since they generally show the same or slightly different states of affect. Thus, the 3D meshes of successive video frames will lie in the same and in adjacent classes in the 2-D VA space. Thus mean faces from adjacent classes can be used to show temporal evolution of affect as was above described.

4.5.3 Expression Blendshape Models

Expression blendshape models provide an effective way to parameterize facial behaviors. The localized blendshape model (Neumann et al. 2013) has been used to describe the selected VA samples. To build this model, we first bring all meshes into full correspondence following the dense registration approach described in Sect. 3.3. As a result, we have a set of training meshes with the same number of vertices and identical topology. Note that we have also selected one neutral mesh for each subject, which should have full correspondence with the rest data. Next, we subtract each 3D mesh from the respective neutral mesh, and create a set of m difference vectors $\mathbf{d}_i \in \mathbb{R}^{3N}$. We then stack them into a matrix $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_m] \in \mathbb{R}^{3N \times m}$, where N is number of vertices in the mesh. Finally, a variant of sparse Principal Component Analysis (PCA) is applied to the data matrix $\mathbf{D}$, so as to identify sparse deformation components $\mathbf{C} \in \mathbb{R}^{h \times 1}$:

$$\arg\min \|\mathbf{D} - \mathbf{B}\mathbf{C}\|_F^2 + \Omega(\mathbf{C}) \quad \text{s.t. } \mathcal{V}(\mathbf{B}), \tag{5}$$

where the constraint $\mathcal{V}$ can be either $\max(|\mathbf{B}_k|) = 1, \ \forall k$ or $\max(\mathbf{B}_k) = 1, \ \mathbf{B} \geq 1, \ \forall k$, with $\mathbf{B}_k \in \mathbb{R}^{3N \times 1}$ denoting the kth components of sparse weight matrix $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_h]$. Selection of these two constraints depends on the actual usage; the major difference is that the latter one allows for negative weights and therefore enables deformation towards both directions, which is useful for describing shapes like muscle bulges. In this paper, we have selected the latter constraint, as we wish to enable bidirectional muscle movement and synthesise a rich variety of expressions. The regularization of sparse components $\mathbf{C}$ was performed with $\ell 1/\ell 2$ norm (Wright et al. 2009; Bach et al. 2012). To permit more local deformations, additional regularization parameters were added into $\Omega(\mathbf{C})$. To compute optimal $\mathbf{C}$ and $\mathbf{B}$, an iterative alternating optimization was employed (please refer to Neumann et al. 2013 for more details).

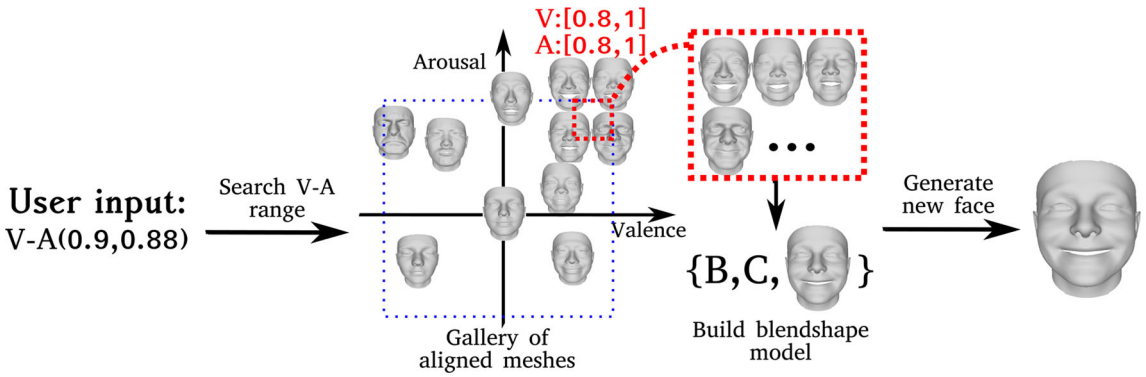


Fig. 6 Generation of new facial affect from the 4D face gallery; the user provides a target VA pair

5 Databases

To evaluate our facial affect synthesis method in different scenarios (e.g. controlled laboratory environment, uncontrolled in-the-wild setting), we utilized neutral facial images from as many as 15 databases (both small and large in terms of size). Table 1 briefly presents the Multi-PIE (Gross et al. 2010), Aff-Wild (Kollias et al. 2019; Zafeiriou et al. 2017), AFEW 5.0 (Dhall et al. 2017), AFEW-VA (Kossaifi et al. 2017), BU-3DFE (Yin et al. 2006), RECOLA (Ringeval et al. 2013), AffectNet (Mollahosseini et al. 2017), RAF-DB (Li et al. 2017), KF-ITW (Booth et al. 2017), Face place, FEI (Thomaz and Giraldi 2010), 2D Face Sets and Bosphorus (Savran et al. 2008) databases that we used in our experimental study. Let us note that for AffectNet no test set is released and thus we use the released validation set to test on and randomly divide the training set into a training and a validation subset (with a 85/15 split).

Table 1 presents these databases by showing:(i) the model of affect they use, their condition, their type (static images or audiovisual image sequences), the total number of frames and (male/female) subjects that they contain and the range of ages of the subjects, and (ii) the total number of images that we synthesized using our approach (both in the valence-arousal and the six basic expressions cases).

6 Experimental Study

This section describes the experiments performed so as to evaluate the proposed approach. At first, we provide a qualitative evaluation of our approach by showing many synthesized images or image sequences from all fifteen databases described in the previous Section; as well as by comparing images generated by state-of-the-art GANs (StarGAN, GANimation) and our approach. Next, a quantitative evaluation is performed by using the synthesized images as additional data to train Deep Neural Networks (DNNs); it is shown that the trained DNNs outperform current state-of-the-art networks and GAN-based methods on each database. Finally an ablation study is performed in which:(i) the synthesized data are considered and used as a training (test) dataset, while the original data are respectively used as test (training) dataset, (ii) the effect of the amount of synthesized data on network performance is studied, (iii) an analysis is performed based on subjects' age.

6.1 Qualitative Evaluation of Achieved Facial Affect Synthesis

We used all databases mentioned in Sect. 5 to supply the proposed approach with 'input' neutral faces. We then synthesized the emotional state corresponding to specific affects (both in VA case and in the six basic expressions one) for these images. At first we show many generated images (static synthesis) according to different VA values, then we illustrate examples of generated image sequences (temporal synthesis) and next we present some synthesized (static) images according to the six basic expressions. Finally, we visually compare images generated by our approach with synthesized images by StarGAN and GANimation.

6.1.1 Results on Static and Temporal Affect Synthesis

Figure 7 shows representative results of facial affect synthesis, when user inputs a VA pair and selects to generate a static image. These results are organized in three age groups:

Table 1 Databases used in our approach, along with their properties and the number of synthesized images in the valence-arousal case and the six basic expressions one; 'static' means images, 'A/V' means audiovisual sequences, i.e., videos

Databases (DBs)	DB type	Model of affect	Condition	DB size	# of subjects	Age range	Total # of synthesized images	
							VA	Basic Expr
MULTI-PIE (Gross et al. 2010)	Static	Neutral, Surprise, Disgust, Smile + Squint, Scream	Controlled	755,370	337 Male: 235 Female: 102	–	52,254	5520
Kinect fusion ITW (Booth et al. 2017)	Static	Neutral, Happiness , Surprise	In-the-wild	3264	17	–	116,235	12,236
FEI (Thomaz and Giraldi 2010)	Static	Neutral, Smile	Controlled	2800	200 Male: 100 Female: 100	19–40	11,400	1200
Face place[a]	Static	6 Basic Expr, Neutral, Confusion	Controlled	6574	235 Male: 143 Female: 92	–	59,736	6288
AFEW 5.0 (Dhall et al. 2017)	A/V	6 Basic Expr, Neutral	In-the-wild	41,406	> 330	1–77	705,649	56,514
RECOLA (Ringeval et al. 2013)	A/V	VA	Controlled	345,000	46 Male: 19 Female: 27	–	46,455	4890
BU-3DFE (Yin et al. 2006)	Static	6 Basic Expr, Neutral	Controlled	2500	100 Male: 56 Female: 44	18–70	5700	600
Bosphorus (Savran et al. 2008)	Static	6 Basic Expr	Controlled	4666	105 Male: 60 Female: 45	25–35	17,018	1792
AffectNet (Mollahosseini et al. 2017)	Static	VA + 6 Basic Expr, Neutral + Contempt	In-the-wild	450,000 manually annotated	–	0 to >50	2,476,235	176,425
Aff-wild (Kollias et al. 2019; Zafeiriou et al. 2017)	A/V	VA	In-the-wild	1,224,094	200 Male: 130 Female: 70	–	60,135	6330
AFEW-VA (Kossaifi et al. 2017)	A/V	VA	In-the-wild	30,050	<600	–	108,864	11,460
RAF-DB (Li et al. 2017)	Static	6 Basic, Neutral + 11 Compound Expr	In-the-wild	15,339 + 3954	–	0–70	121,866	12,828
2D face sets[b]: Pain	Static	6 Basic, Neutral + 10 Pain Expr	Controlled	599	23 Male: 13 Female: 10	–	2736	288
2D face sets: Iranian	Static	Neutral, Smile	Controlled	369	34 Male: 0 Female: 34	–	2679	282
2D face sets: Nottingham scans	Static	Neutral	Controlled	100	100 Male: 50 Female: 50	–	5700	600

[a] Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/
[b] http://pics.stir.ac.uk

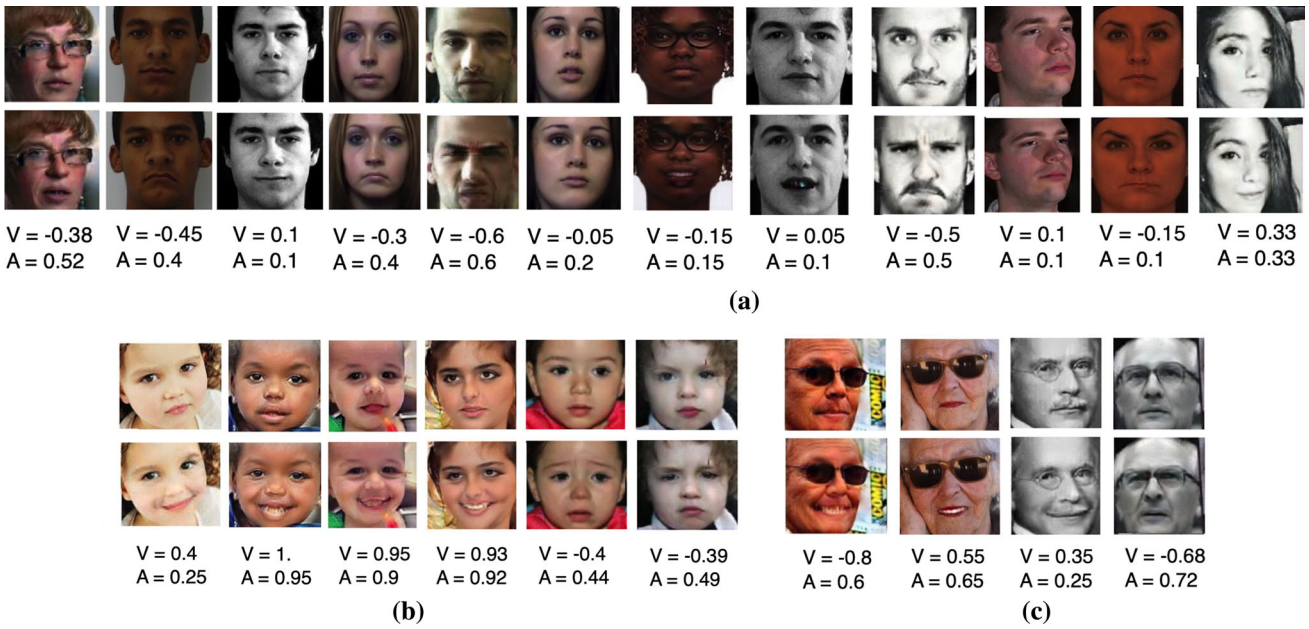


Fig. 7 VA Case of static (facial) synthesis across all databases; first rows show the neutral, second ones show the corresponding synthesized images and third rows show the corresponding VA values. Images of: **b** kids, **c** elderly people and **a** in-between ages, are shown

Fig. 7b kids, Fig. 7c elderly people and Fig. 7a in-between ages. In each part, the first row illustrates neutral images sampled from each of the aforementioned databases, the second one shows the respective synthesized images and the third shows the respective VA values that were synthesized. Moreover, Fig. 8 shows neutral images on the left hand side (first column) and synthesized images, with various valence and arousal values, on the right hand side (following columns). It can be observed that the synthesized images are identity preserving, realistic and vivid. Figure 9 refers to the basic expression case; it shows neutral images on the left hand side of (a) and (b) and synthesized images with basic expressions on the right hand side. Figure 10 illustrates the VA case for temporal synthesis, as was described in Sect. 4.5.2. Neutral images are shown on the left hand side, while synthesized face sequences with time-varying levels of affect are shown on the right hand side.

All these figures show that the proposed framework works well, when using images from either in-the-wild, or controlled databases. This indicates that we can effectively synthesize facial affect irregardless of image conditions (e.g., occlusions, illumination and head poses).

6.1.2 Comparison with GANs

In order to characterize the value that the proposed approach imparts, we provide qualitative comparisons with two state-of-the-art GANs, namely StarGAN (Choi et al. 2018) and GANimation. Like CycleGAN (referenced in Sect. 2), StarGAN performs image-to-image translation, but adopts a unified approach such that a single generator is trained to map an input image to one of multiple target domains, selected by the user. By sharing the generator weights among different domains, a dramatic reduction of the number of parameters is achieved. GANimation was described in Sect. 2.

At first, it should be mentioned that, the original StarGAN synthesized images according to the basic expressions (apart from facial attributes) and the GANimation synthesized images according to AUs. However, in psychology, there does not exist any mapping between AUs–VA and no consistent mapping (across studies) between AUs-expressions, or VA-expressions. In order to achieve a fair comparison of our method with these networks, we applied them—for the first time—to the VA space; we trained them with the same 600,000 frames of 4DFAB that we used in our approach. In both networks, pre-processing was conducted, which included face detection and alignment. For a fair comparison, in all presented results (both qualitative and quantitative), the GANs were provided with the same neutral images and the same VA values.

Figure 11 presents a visual comparison between images generated by our approach, StarGAN and GANimation. It shows the neutral images, the synthesized VA values and the resulting images. It is evident that our approach synthesizes samples that:(i) look much more natural and realistic, (ii) maintain the degree of sharpness of the original neutral image, and (iii) combine visual accuracy with spatial resolution.

Some further deductions can be made from Fig. 11. StarGAN does not perform well when tested on different in-

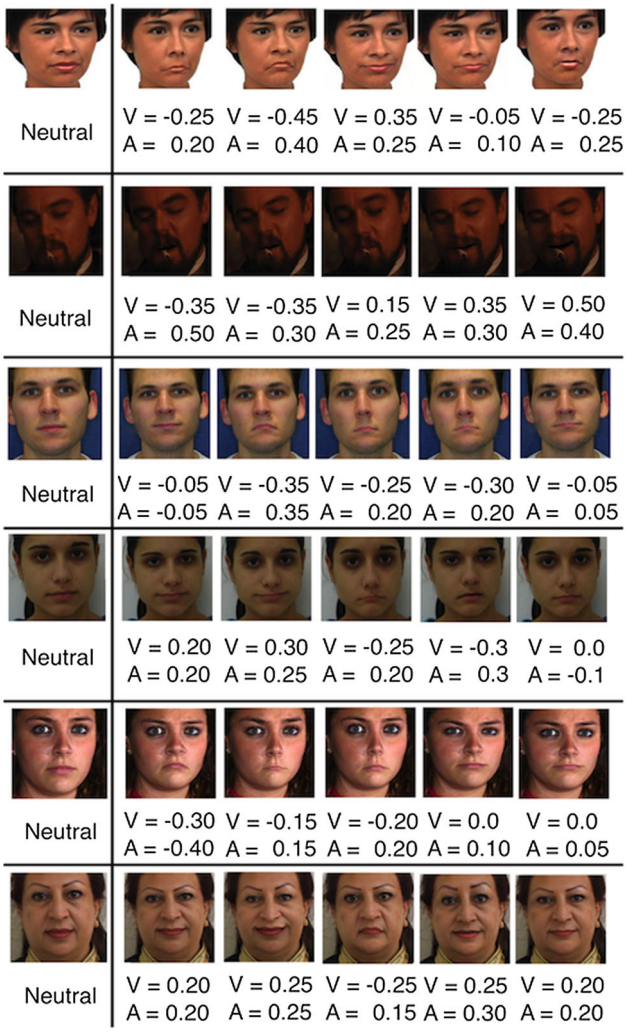


Fig. 8 VA case of facial synthesis: on the left hand side are the neutral 2D images and on the right the synthesized images with different levels of affect

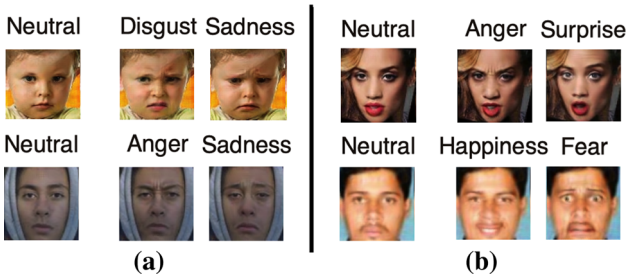


Fig. 9 Basic Expression Case of facial synthesis: on the left hand side of **a** and **b** are the neutral 2D images and on the right the synthesized images with some basic expressions

the-wild and controlled databases that include variations in illumination conditions and head poses. StarGAN is unable to reflect detailed illumination; unnatural lighting changes were observed on the results. These can be explained because in the original StartGAN paper (Choi et al. 2018), its capability to generate affect has not been tested on in-the-wild facial analysis (we refer only to the case of emotion recognition). In general, StarGAN yields more realistic results when it is trained simultaneously with multiple datasets annotated for different tasks.

Additionally, in Choi et al. (2018), when referring to emotion recognition, StarGAN was trained and evaluated on Radboud Faces Database (RaFD) (Langner et al. 2010) which:(i) is very small in terms of size (around 4800 images) and (ii) is a lab-controlled and posed expression database. Last but not least, StarGAN has been tested to change only a particular aspect of a face among a discrete number of attributes/emotions defined by the annotation granularity of the dataset. As can be seen in Fig. 11, StarGAN cannot accurately provide realistic results when tested in the much broader and more difficult task of valence and arousal generation (and estimation).

As far as GANimation is concerned, its results are also worse than the results of our approach. In most cases, it shows artifacts and in some cases certain levels of blurriness. When compared to StarGAN, GANimation seems more robust to changing backgrounds and lighting conditions; this is due to the attention and color masks that it contains. Nevertheless, in general, errors in the attention mechanism occur when the input contains extreme expressions. The attention mechanism does not seem to sufficiently weight the color transformation, causing transparencies. It is interesting to note that on the Leonardo DiCaprio image, the synthesized image by GANimation shows open eyes, whereas on the neutral image (and the one synthesized by our approach) eyes are closed; this illustrates errors of the mask. For example, in Fig. 11, images produced by GANimation in columns 1, 3, 4, 5, 6, 9 show the discussed problems.

6.2 Quantitative Evaluation of the Facial Affect Synthesis Through Data Augmentation

It is generally accepted that using more training data—of good quality—leads to better results in supervised training. Data augmentation increases the effective size of the training dataset. In this section we present a data augmentation strategy which uses the synthesized data produced by our approach, as additional data to train DNNs, for both valence-arousal prediction, as well as classification into the basic expression categories. In particular, we describe experiments performed on eight databases, presenting the adopted evaluation criteria, the networks we used and the obtained results. We also report the performances of the networks trained—in a data augmentation manner—with synthesized images from StarGAN and GANimation. It is shown that the DNNs trained with the proposed data augmentation methodology outperform both the state-of-the-art techniques and the

Fig. 10 VA case of temporal (facial) synthesis: on the left hand side are the neutral 2D images and on the right the synthesized image sequences

DNNs trained with StarGAN and GANimation, in all experiments, validating the effectiveness of the proposed facial synthesis approach. Let us first explain some notations. In the followings, by reporting 'network_name trained using StarGAN', 'network_name trained using GANimation' and 'network_name trained using the proposed approach', we refer to networks trained with the specific database's training set augmented with data synthesized by StarGAN, GANimation and the proposed approach, respectively.

6.2.1 Leveraging Synthesized Data for Training Deep Neural Networks: Valence-Arousal Case

In this set of experiments we consider four facial affect databases annotated in terms of valence and arousal, the Aff-Wild, RECOLA, AffectNet and AFEW-VA data-bases. At first, we selected neutral frames from these databases, i.e., frames with zero valence and arousal values (human inspection was also conducted to make sure that they represented neutral faces). For every frame, we synthesized facial affect according to the methodology described in Sect. 4. We start by first describing the evaluation criteria used in our experiments.

6.2.2 The Adopted Evaluation Criteria

The main evaluation criterion that we use is the Concordance Correlation Coefficient (CCC) (Lawrence, and Lin 1989), which has been widely used in related Challenges (e.g., Valstar et al. 2016); we also report the Mean Squared Error (MSE), since this has been also frequently used in related research.

CCC evaluates the agreement between two time series by scaling their correlation coefficient with their mean square difference. CCC takes values in the range $[-1, 1]$, where $+1$ indicates perfect concordance and -1 denotes perfect discordance. Therefore high values are desired. CCC is defined as follows:

Fig. 11 Generated results by our approach, StarGAN and GANimation

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}, \tag{6}$$

where s_x and s_y are the variances of the ground truth and predicted values respectively, $\bar{x}$ and $\bar{y}$ are the corresponding mean values and s_{xy} is the respective covariance value.

The Mean Squared Error (MSE) provides a simple comparative metric, with a small value being desirable. MSE is defined as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2, \tag{7}$$

where x and y are the ground truth and predicted values respectively and N is the total number of samples.

In some cases we also report the Pearson-CC (P-CC) and the Sign Agreement Metric (SAGR), since they have been reported by respective state-of-the-art methods.

The P-CC takes values in the range $[-1, 1]$ and high values are desired. It is defined as follows:

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}, \tag{8}$$

where s_x and s_y are the variances of the ground truth and predicted values respectively and s_{xy} is the respective covariance value.

The SAGR takes values in the range [0,1], with high values being desirable. It is defined as follows:

$$SAGR = \frac{1}{N}\sum_{n=1}^{N}\delta(sign(x_i), sign(y_i)), \tag{9}$$

Table 2 Aff-Wild: CCC and MSE evaluation of valence and arousal predictions provided by the VGG-FACE-GRU trained using our approach versus state-of-the-art networks and methods

Networks	CCC		MSE	
	Valence	Arousal	Valence	Arousal
FATAUVA-Net (Chang et al. 2017)	0.396	0.282	0.123	0.095
VGG-FACE-GRU trained using StarGAN	0.556	0.424	0.085	0.060
VGG-FACE-GRU trained using GANimation	0.576	0.433	0.077	0.057
AffWildNet (Kollias et al. 2017, 2019)	0.570	0.430	0.080	0.060
VGG-FACE-GRU trained using the proposed approach	**0.595**	**0.445**	**0.074**	**0.051**

Valence and arousal values are in $[-1, 1]$
Bold values correspond to the results of the best methods

where N is the total number of samples, x and y are the ground truth and predicted values respectively, δ is the Kronecker delta function and $\delta(sign(x), sign(y))$ is defined as:

$$\delta(sign(x), sign(y)) = \begin{cases} 1, & \text{x} \geqslant 0 \text{ and y} \geqslant 0 \\ 1, & \text{x} \leqslant 0 \text{ and y} \leqslant 0 \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

6.2.3 Experiments on Dimensional Affect

Aff-Wild We synthesized 60,135 images from the Aff-Wild database and added those images to the training set of the first Affect-in-the-wild Challenge. The employed network architecture was the AffWildNet (VGG-FACE-GRU) described in Kollias et al. (2017, 2019).

Table 2 shows a comparison of the performance of: the VGG-FACE-GRU trained using: (i) our approach, (ii) StarGAN, and (iii) GANimation; the best performing network, AffWildNet, reported in Kollias et al. (2017, 2019); the winner of the Aff-Wild Challenge (Chang et al. 2017) (FATAUVA-Net).

From Table 2, it can be verified that the network trained on the augmented dataset, with synthesized by our approach images, outperformed all other networks. It should be noted that the number of synthesized images (around $60K$) was small compared to the size of Aff-Wild's training set (around $1M$), the latter being already sufficient for training the best performing DNN; consequently, the improvement was not large, about 2%. An interesting observation is that the network trained using StarGAN displayed worse performance than AffWildNet. This means that the 68 landmark points that were passed as additional input to the AffWildNet helped the network in reaching a better performance than just adding a small amount (compared to the training set size) of auxiliary synthesized data. The MSE error improvement on Valence and Arousal estimation provided by the augmented training versus the AffWildNet one, over the different areas of the VA space, is shown through the 2D histograms presented in Fig. 12. It can be seen that the improvement on MSE was better in areas in which a larger number of new samples was generated, i.e., in the positive valence regions.

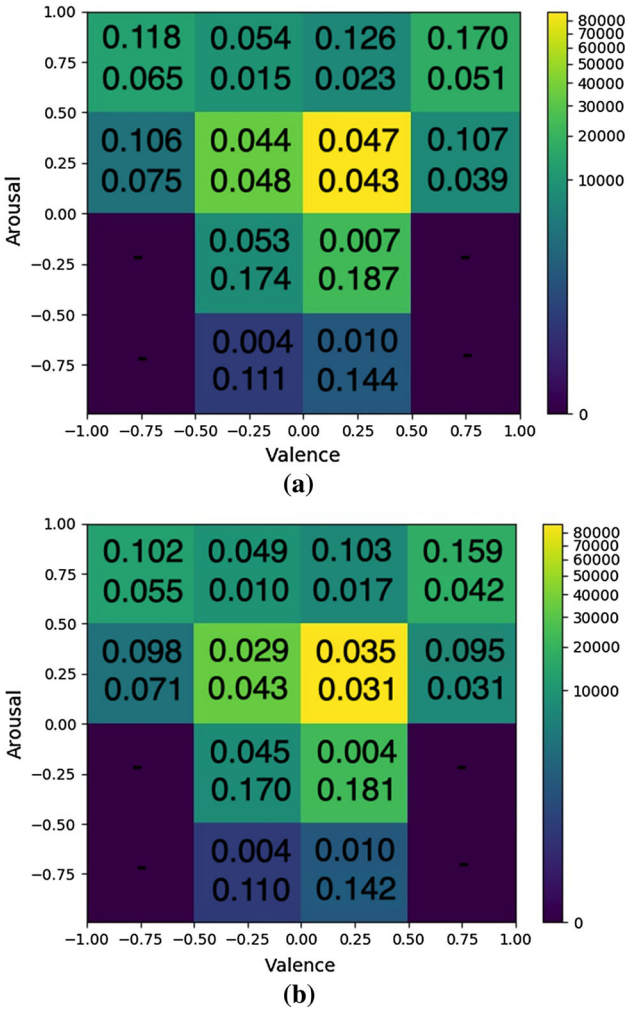


Fig. 12 The 2D histogram of valence and arousal Aff-Wild's test set annotations, along with the MSE per grid area, in the case of **a** AffWildNet and **b** VGG-FACE-GRU trained using the proposed approach

RECOLA We generated 46,455 images from RECOLA; this number corresponds to around 40% of its training data set size. The employed network architecture was the ResNet-GRU described in Kollias et al. (2019).

Table 3 RECOLA: CCC evaluation of valence and arousal predictions provided by the ResNet-GRU trained using the proposed approach versus other state-of-the-art networks and methods

Networks	CCC	
	Valence	Arousal
ResNet-GRU (Kollias et al. 2019)	0.462	0.209
ResNet-GRU trained using StarGAN	0.503	0.245
ResNet-GRU trained using GANimation	0.486	0.222
Fine-tuned AffWildNet (Kollias et al. 2019)	0.526	0.273
ResNet-GRU trained using the proposed approach	**0.554**	**0.312**

Bold values correspond to the results of the best methods

Table 3 shows a comparison of the performance of: the ResNet-GRU network trained using:(i) our approach, (ii) StarGAN, and (iii) GANimation; the AffWildNet fine-tuned on the RECOLA, as reported in Kollias et al. (2019); a ResNet-GRU directly trained on RECOLA, as reported in Kollias et al. (2019).

From Table 3, it can be verified that the network trained using the proposed approach outperformed all other networks. The above gains in performance can be justified by the fact that the number of synthesized images (around 46,500) was significant compared to the size of RECOLA's training set (around 120,000) and that the original training set size was not very sufficient to train the DNNs. It is worth mentioning that the GAN based methods have not managed to provide a sufficiently enriched dataset so that a similar boost in the achieved performances could be obtained. The MSE error improvement on Valence and Arousal estimation provided by the augmented training versus the original one (which was 0.045–0.100 versus. 0.055–0.160), over the different areas of the VA space, is shown through the 2D histograms presented in Fig. 13. Big reduction of MSE value was achieved in all covered VA areas.

AffectNet The AffectNet database contains around 450,000 manually annotated images and around 550,000 automatically annotated images for valence-arousal. We only used the manually annotated images so as to be consistent with the state-of-the-art networks that were also trained using this set. Additionally, the manually annotated set ensures that the images used by our approach to synthesize new, are indeed neutral. We created 2,476,235 synthesized images from the AffectNet database, a number that is more than 5 times bigger than the training data size. The employed network architecture was VGG-FACE. For comparison purposes, we trained the network using the original training data set (let us call this network 'the VGG-FACE baseline').

Table 4 shows a comparison of the performance of: the VGG-FACE baseline; the VGG-FACE trained using:(i) our approach, (ii) StarGAN, and (iii) GANimation; AlexNet, which is the baseline network of the AffectNet database (Mollahosseini et al. 2017).

From Table 4, it can be verified that the network trained by the proposed methodology outperformed all other networks.

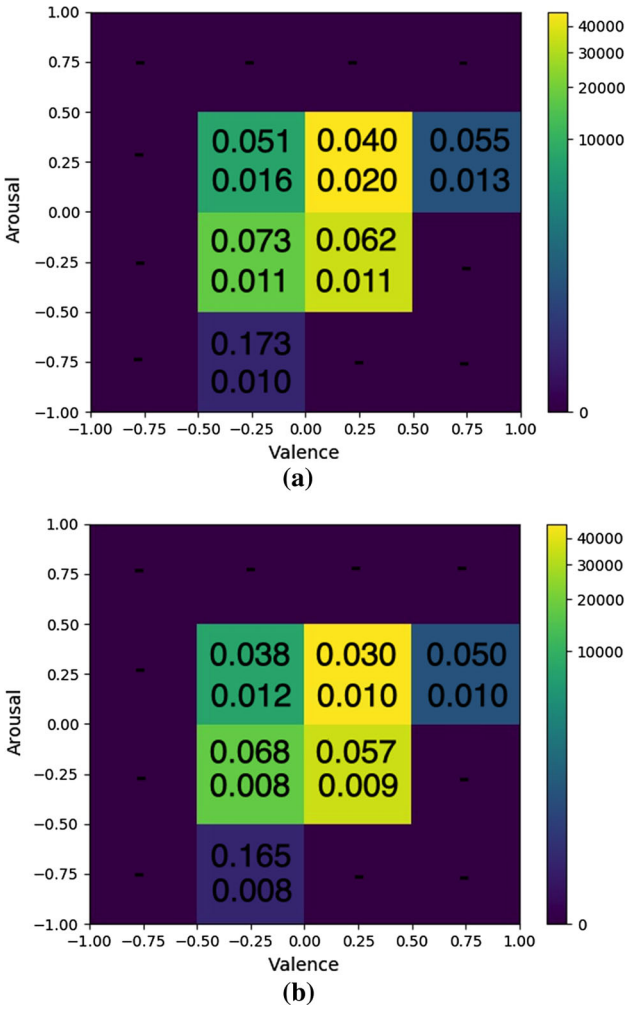


Fig. 13 The 2D histogram of valence and arousal RECOLA's test set annotations, along with the MSE per grid area, in the case of **a** ResNet-GRU and **b** ResNet-GRU trained using the proposed approach

This boost in performance has been large, in all evaluation criteria, compared to the VGG-FACE baseline network, with spread of this improvement over the VA space shown in Fig. 14. The explanation arises from the large number of synthesized images that helped the network train and generalize better, since in the training set there existed a lot of ranges that were poorly represented. This is shown in

Table 4 AffectNet: CCC, P-CC, SAGR and MSE evaluation of valence and arousal predictions provided by the VGG-FACE trained using the proposed approach versus state-of-the-art networks and methods

Networks	CCC		P-CC		SAGR		MSE	
	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
AlexNet (Mollahosseini et al. 2017)	0.60	0.34	**0.66**	0.54	0.74	0.65	**0.14**	0.17
The VGG-FACE baseline	0.50	0.37	0.54	0.48	0.65	0.60	0.19	0.18
VGG-FACE trained using StarGAN	0.55	0.42	0.58	0.49	0.74	0.73	0.17	0.16
VGG-FACE trained using GANimation	0.56	0.45	0.59	0.51	0.74	0.74	0.15	0.16
VGG-FACE trained using the proposed approach	**0.62**	**0.54**	**0.66**	**0.55**	**0.78**	**0.75**	**0.14**	**0.15**

Valence and arousal values are in $[-1, 1]$
Bold values correspond to the results of the best methods

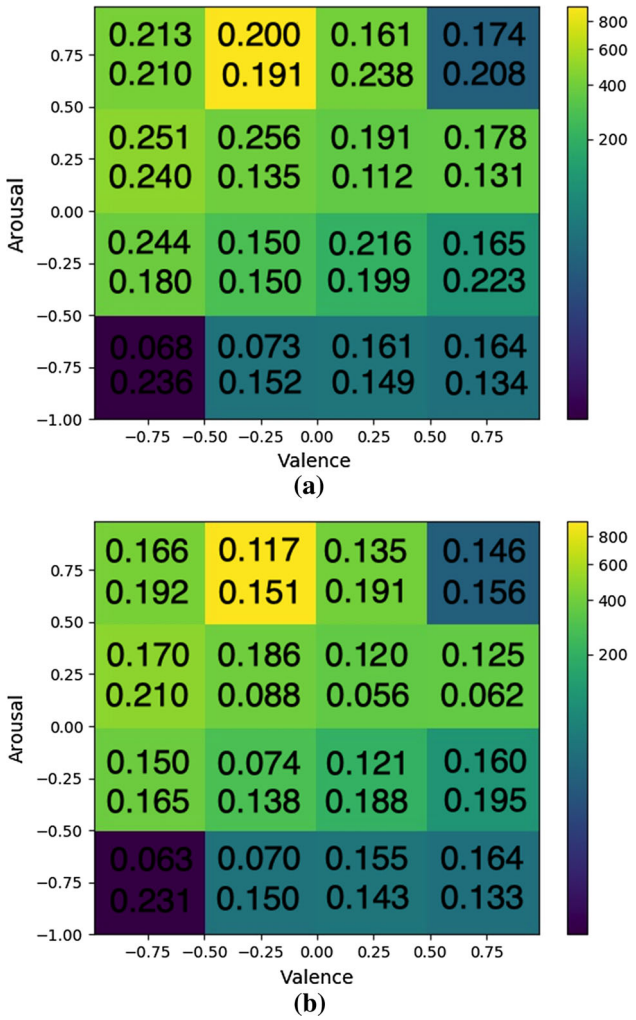


Fig. 14 The 2D histogram of valence and arousal AffectNet's test set annotations, along with the MSE per grid area, in the case of **a** VGG-FACE baseline, **b** VGG-FACE trained using the proposed approach

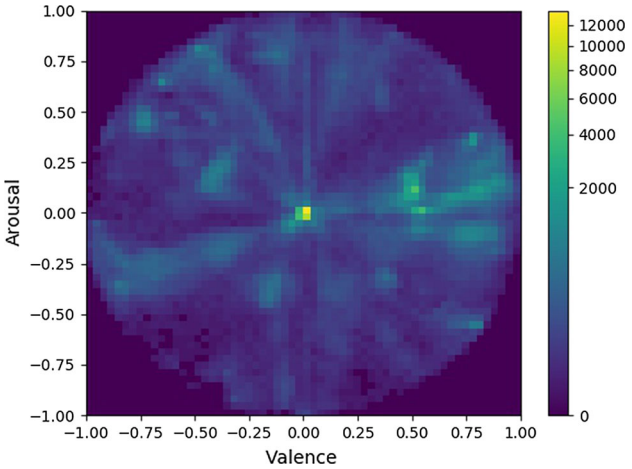


Fig. 15 The 2D histogram of valence and arousal AffectNet's annotations for the manually annotated training set

the histogram of the—manually annotated—training set, for valence and arousal, in Fig. 15. Our network also outperformed the AffectNet's database baseline. For the arousal estimation, the performance gain was remarkable, mainly in CCC and SAGR evaluation criteria, whereas for the valence estimation the performance gain was also significant.

AFEW-VA We synthesized 108,864 images from the AFEW-VA database, a number that is more than 3.5 times bigger than its original size. For training, we used the VGG-FACE-GRU architecture described in Kollias et al. (2019). Similarly to Kossaifi et al. (2017), we used a 5-fold person-independent cross-validation strategy and at each fold we augmented the training set with the synthesized images of people appearing only in that set (preserving the person independence).

Table 5 shows a comparison of the performance of: the VGG-FACE-GRU network trained using: (i) our approach, (ii) StarGAN, and (iii) GANimation; the best performing network as reported in Kossaifi et al. (2017).

From Table 5, it can be verified that the network trained using the proposed approach outperformed all other networks. Great boost in performance was achieved. The general gain in performance can be justified by the fact that the number of synthesized images (around 109,000) is much greater than the number of images in the dataset (around 30,000), with the latter being rather small for effectively training the

Table 5 AFEW-VA: P-CC and MSE evaluation of valence and arousal predictions provided by the VGG-FACE trained using the proposed approach versus state-of-the-art network and methods

Networks	Pearson CC		MSE	
	Valence	Arousal	Valence	Arousal
Best of Kossaifi et al. (2017)	0.407	0.450	0.484	0.247
VGG-FACE trained using StarGAN	0.512	0.489	0.262	0.097
VGG-FACE trained using GANimation	0.491	0.453	0.308	0.151
VGG-FACE-GRU trained using the proposed approach	**0.562**	**0.614**	**0.226**	**0.075**

Valence and arousal values are in $[-1, 1]$
Bold values correspond to the results of the best methods

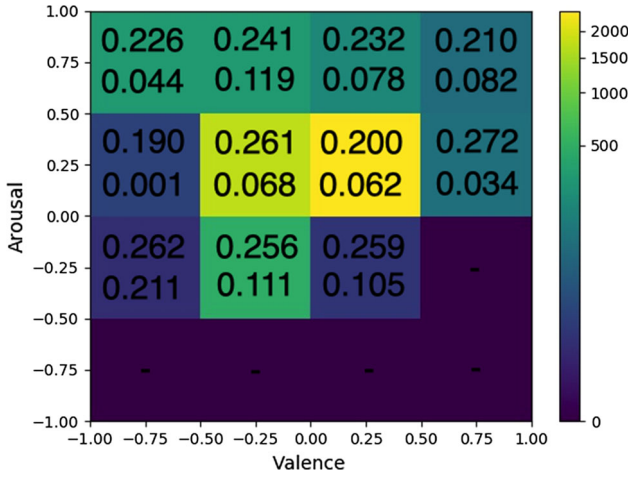


Fig. 16 The 2D histogram of valence and arousal AFEW-VA's test set annotations, along with the MSE per grid area, in the case of the VGG-FACE trained using the proposed approach

DNNs. The 2D histogram in Fig. 16 shows the achieved MSE when using the proposed approach over the different areas of the VA space.

6.2.4 Leveraging Synthesized Data for Training Deep Neural Networks: Basic Expressions Case

In the following experiments we used the synthesized faces to train DNNs, for classification into the six basic expressions, over four facial affect databases, RAF-DB, AffectNet, AFEW and BU-3DFE. Our first step has been to select neutral frames from these four databases. Then, for each frame, we synthesized facial affect according to the methodology described in Sect. 4. We start by first describing the evaluation criteria used in our experiments.

6.2.5 The Adopted Evaluation Criteria

One evaluation criterion used in the experiments is total accuracy, defined as the total number of correct predictions divided by the total number of samples. Another criterion is the F_1 score, which is a weighted average of the recall (= the ability of the classifier to find all the positive samples) and precision (= the ability of the classifier not to label as positive a sample that is negative). The F_1 score reaches its best value at 1 and its worst score at 0. In our multi-class problem, F_1 score is the unweighted mean of the F_1 scores of the expression classes. F_1 score of each class is defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{11}$$

Another criterion that is used is the average of the diagonal values of the confusion matrix for the seven basic expressions.

One, or more of the above criteria are used in our experiments, so as to illustrate the comparison with other state-of-the-art methods.

6.2.6 Experiments on Categorical Affect

RAF-DB In this database we only considered the six basic expression categories, since our approach synthesizes images based on these categories; we ignored compound expressions that were included in the original dataset. We created 12,828 synthesized images, which are slightly more than the training images (12,271). We employed the VGG-FACE network. For comparison purposes, we trained the network using the original training dataset (let us call this network 'the VGG-FACE baseline').

For further comparison purposes, we used the networks defined in Li et al. (2017):(i) mSVM-VGG-FACE: first the VGG-FACE was trained on the RAF-DB database and then features from the penultimate fully connected layer were extracted and fed into a Support Vector Machine (SVM) that performed the classification, (ii) LDA-VGG-FACE: same as before: LDA was applied on the features which were extracted from the penultimate fully connected layer and performed the final classification and (iii) mSVM-DLP-CNN: the designed Deep Locality Preserving CNN network (we refer the interested reader for more details to Li et al. (2017)) was first trained on the RAF-DB database and then a SVM performed the classification using the features extracted from the penultimate fully connected layer of this architecture.

Table 6 shows a comparison of the performance of the above described networks. From Table 6, it can be veri-

Table 6 RAF-DB: the diagonal values of the confusion matrix for the seven basic expressions and their average, using the VGG-FACE trained using the proposed approach, as well as using other state-of-the-art networks

Networks	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral	Average
LDA-VGG-FACE (Li et al. 2017)	0.661	0.250	0.378	0.731	0.515	0.535	0.472	0.506
mSVM-VGG-FACE (Li et al. 2017)	0.685	0.275	0.351	0.853	0.649	0.663	0.599	0.582
The VGG-FACE baseline	0.691	0.287	0.363	0.853	0.661	0.666	0.600	0.589
mSVM-DLP-CNN (Li et al. 2017)	0.716	0.522	**0.622**	**0.928**	0.801	0.812	0.803	0.742
VGG-FACE trained using the proposed approach	**0.784**	**0.644**	**0.622**	0.911	**0.812**	**0.845**	**0.806**	**0.775**

Bold values correspond to the results of the best methods

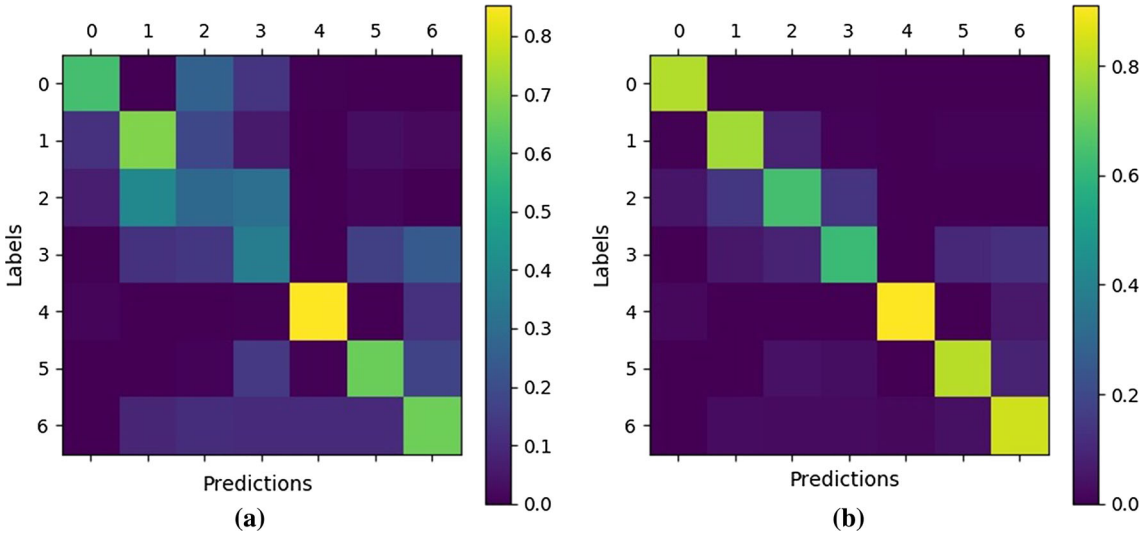


Fig. 17 The confusion matrix of **a** VGG-FACE baseline and **b** VGG-FACE trained using the proposed approach for the RAF-DB database; 0: Neutral, 1: Anger, 2: Disgust, 3: Fear, 4: Joy, 5: Sadness, 6: Surprise

fied that the network trained using the proposed approach outperformed all state-of-the-art nets. When compared to the mSVM-VGG-FACE and LDA-VGG-FACE networks, the boost in performance has been significant. This can be explained by the fact that the disgust and fear classes, originally, did not contain a lot of training images, but after adding the synthesized data, they did. This resulted in obtaining a better performance in the other classes, as well. Interestingly, there was also a considerable performance gain in the neutral class, that did not contain any synthesized images. This can be explained by considering the fact that the network trained with the augmented data could distinguish better the classes, since it had more samples in the two above described categories. Figure 17 illustrates the whole confusion matrix of the VGG-FACE baseline and the VGG-FACE trained using the proposed approach, giving a better insight on the improved performance and verifying the above explanations.

AffectNet We synthesized 176,425 images from the AffectNet database, a number that is almost 40% of its size. It should be mentioned that the AffectNet database contained the six basic expressions and another one, contempt. Our approach synthesized images only for the basic expressions, so for the contempt class we only kept the original training data. The network architecture that we employed here was VGG-FACE. For comparison purproses, we trained a VGG-FACE network using the training set of the AffectNet database (let us call this network 'the VGG-FACE baseline').

Table 7 shows a comparison of the performance of:(i) the VGG-FACE baseline, (ii) the VGG-FACE network trained using the proposed approach and (iii) AlexNet, the baseline network of the AffectNet database (Mollahosseini et al. 2017).

From Table 7, it can be verified that the network trained using the proposed approach outperformed all the other networks. In more detail, when compared to the VGG-FACE baseline network, the boost in performance was significant, as also shown in Fig. 18 in terms of the confusion matrices obtained by the two networks. This can be explained by the big size of the added synthesized images. When compared to the AffectNet's baseline, a slightly improved performance was also obtained; this could be higher, if we had synthesized images for the contempt category as well.

AFEW We synthesized 56,514 images from the AFEW database; this number was almost 1.4 times bigger than its

Table 7 AffectNet: total accuracy and F_1 score of the VGG-FACE trained using the proposed approach versus state-of-the-art networks

Networks	Total accuracy	F_1 score
AlexNet (Mollahosseini et al. 2017)	0.58	0.58
The VGG-FACE baseline	0.52	0.51
VGG-FACE trained using the proposed approach	**0.60**	**0.59**

Bold values correspond to the results of the best methods

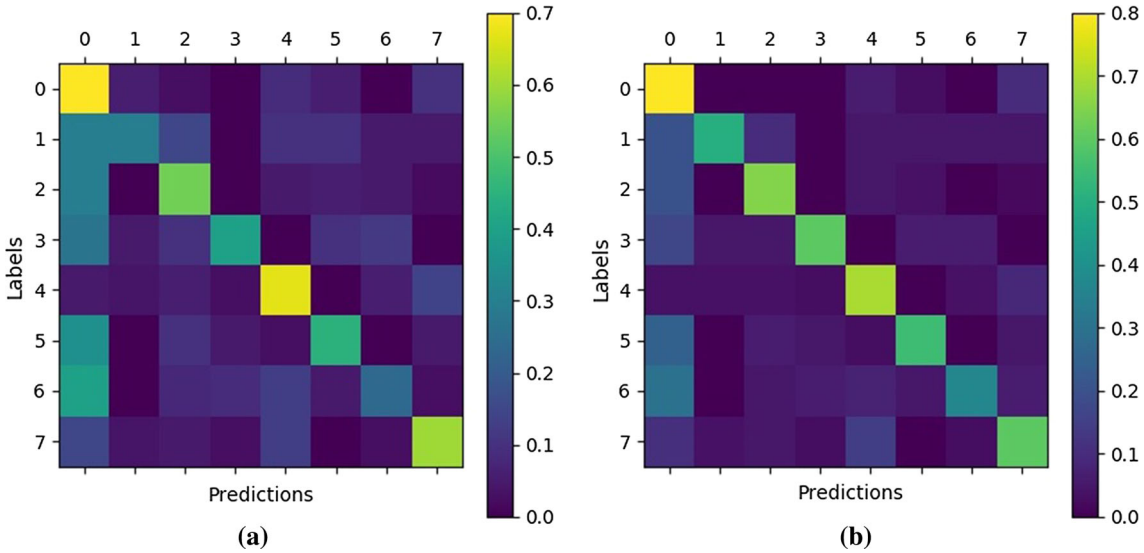


Fig. 18 The confusion matrix of **a** VGG-FACE baseline and **b** VGG-FACE trained using the proposed approach for the AffectNet database; 0: Neutral, 1: Anger, 2: Disgust, 3: Fear, 4: Joy, 5: Sadness, 6: Surprise, 7: Contempt

training set size (41,406). The employed network architecture was VGG-FACE. For comparison purposes, we first trained a baseline network on AFEW's training set, which we call the VGG-FACE baseline. For further comparisons, we used the following networks developed by the three winning methods of the EmotiW 2017 Grand Challenge:(i) VGG-FACE-FER: the VGG-FACE was first fine-tuned on the FER2013 database (Goodfellow et al 2013) and then trained on the AFEW as described in Knyazev et al. (2017), (ii) VGG-FACE-external: the VGG-FACE was trained on the union of the AFEW database and some external data as described in Vielzeuf et al. (2017) and (iii) VGG-FACE-LSTM-external-augmentation: the VGG-FACE-LSTM was trained on the union of the AFEW database and some external data; then data augmentation was performed, as described in Vielzeuf et al. (2017).

Table 8 shows a comparison of the performance of the above described networks. From Table 8, one can see that the VGG-FACE trained using the proposed approach performed much better than the same network trained on, either only the AFEW database, or the union of the AFEW database with some external data whose size in terms of videos was the same as that of AFEW. The boost in performance can be explained taking into account the fact that the fear, disgust and surprise classes contained few data in AFEW and that our approach augmented the data size of those classes; in total the large number of synthesized images assisted to improve the performance of the network. This is evident when comparing the confusion matrix of the VGG-FACE baseline to the one of VGG-FACE trained using the proposed approach, as can be seen in Fig. 19. The diagonal of the two confusion matrices indicates that there is an increase in the performance in almost all basic categories.

Additionally, performance of our network is slightly better than the performance of the same VGG-FACE network first fine-tuned on the FER2013 database and then trained on the AFEW. FER2013 is a database of around 35,000 still images and different identities, annotated with the six basic expressions. In this case, the network that was first fine-tuned on the FER2013 database has seen more faces, since the tasks were similar. However, still our network provided a slightly better performance. On the other hand, our network had a slightly worse performance than a VGG-FACE-LSTM network that was trained with the same external data mentioned before and was also trained with data augmentation. Here, it was the LSTM network, which due to the time recurrent nature could better exploit the fact that AFEW consists of video sequences.

BU-3DFE We synthesized 600 images from the BU-3DFE database. This number was almost one fourth of its size

Table 8 AFEW: total accuracy of the VGG-FACE trained using the proposed approach versus state-of-the-art networks

Networks	Total accuracy
The VGG-FACE baseline	0.379
VGG-FACE-external (Vielzeuf et al. 2017)	0.414
VGG-FACE-FER (Knyazev et al. 2017)	0.483
VGG-FACE-LSTM-external-augmentation (Vielzeuf et al. 2017)	**0.486**
VGG-FACE trained using the proposed approach	0.484

Bold values correspond to the results of the best methods

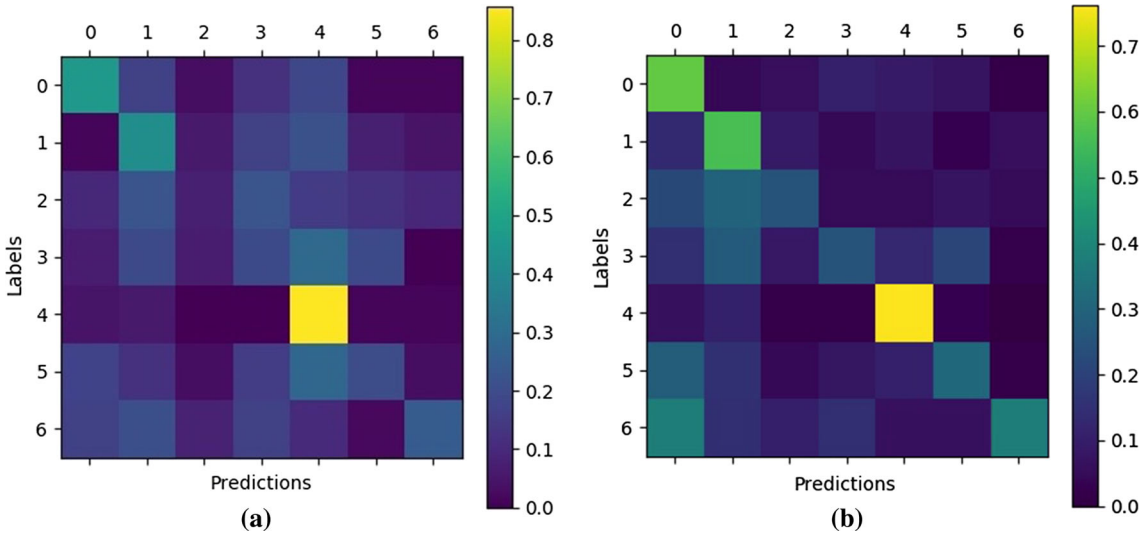


Fig. 19 The confusion matrix of **a** VGG-FACE baseline and **b** VGG-FACE trained using the proposed approach for the AFEW database; 0: Neutral, 1: Anger, 2: Disgust, 3: Fear, 4: Joy, 5: Sadness, 6: Surprise

(2500). BU-3DFE is a small database and is not really suited for training DNNs. The network architecture that we employed here was VGG-FACE, with a modification in the number of hidden units in the two first fully connected layers. Since we did not have a lot of data for training the network, we(i) used 256 and 128 units in the two fully connected layers and (ii) kept the convolutional weights fixed, training only the fully connected ones. For training the network on this database, we used a 10-fold person-independent cross-validation strategy; in each fold, we augmented the training set with the synthesized images of people appearing only in that set (preserving person independence). The reported total accuracy of the model has been the average of the total accuracies over the 10-folds.

At first, we trained the above described VGG-FACE network (let us call this network 'the VGG-FACE baseline'). Next, we trained the above described VGG-FACE network, but also applied on-the-fly data augmentation techniques, such as: small rotations, left and right flipping, first resize and then random crop to original dimensions, random brightness and saturation (let us call this network 'VGG-FACE-augmentation'). Finally, we trained the above described VGG-FACE network using the proposed approach.

Table 9 shows a comparison of the performance of those networks. From Table 9, it can be verified that the network trained using the proposed approach greatly outperformed the networks trained without it. This indicates that the proposed approach for synthesizing images can be used for data augmentation in cases of small amount of DNN training data, being able to significantly improve the obtained performances.

6.3 Quantitative Evaluation of the Facial Affect Synthesis Used in Testing or Training Tasks

Results in the previous section show that the data generated using our approach provide improvements in network performance in both valence-arousal and basic expressions settings, when used for data augmentation. In the following, we perform further analysis (two different settings) to assess the quality of our generated data, compared to the data synthesized by StarGAN and GANimation, focusing only on the synthesized data.

In the first setting, the synthesized data are evaluated as a test set, for each database, against models trained on real data/images.

Table 9 BU-3DFE: total accuracy of the VGG-FACE trained using the proposed approach versus the VGG-FACE baseline and the VGG-FACE trained with on-the-fly data augmentation

Networks	Total accuracy
The VGG-FACE baseline	0.528
VGG-FACE-augmentation	0.588
VGG-FACE trained using the proposed approach	**0.768**

Bold values correspond to the results of the best methods

The AffWildNet that has been trained solely on Aff-Wild's training set, the ResNet-GRU trained on the RECOLA's training set and the VGG-FACE baseline trained on AffectNet's training set (all described in Sect. 6.2.3), have been used as emotion regressors and are being evaluated on each of the three afore-mentioned synthesized datasets. From Table 10 it is evident that the networks trained on the afore mentioned databases displayed a much better performance (in all databases) when tested on the synthesized data from the proposed approach in comparison to the synthesized data from StarGAN and GANimation.

We further conducted a second setting, using the synthesized data to train respective DNN models. These models are then evaluated on the real test set of Aff-Wild, RECOLA and AffectNet. Table 11 shows the results of this setting. The performance in terms of both CCC and MSE is much higher in all databases when the networks are trained with the data synthesized by the proposed approach. This difference in the compared performances, along with the former results, reflect the direct value of our generated data in enhancing regression performance.

6.4 Effect of Synthesized Data Granularity on Performance Improvement

In this subsection we performed experiments using a subset of our synthesized data for augmenting the data-bases. Our aim is to see if all synthesized data are needed for augmenting network training and more generally to see how the improvement in classification and regression scale with the granularity of synthesized data. In more detail, for each database used in our experiments, we used a subset of N synthesised data from this database to augment its training set. Table 12 shows the databases and its corresponding N values.

Figure 20 shows the improvement in network performance when training using additionally auxiliary data; the improvement shown per database is the difference in the performances when training networks with only the database's training set and when training them with the union of the training set and auxiliary data. Figure 20 illustrates for each database the difference in network performance, when N synthesized data generated by our approach (N defined in Table 12) are used as auxiliary data.

The performance measure for Aff-Wild, RECOLA, AffectNet and AFEW-VA is the average of valence CCC and arousal CCC. The performance measure for the rest databases depends on the database. More details follow.

Dimensional affect generation

For the Aff-Wild database, we use the VGG-FACE-GRU network. When augmenting the dataset with $30K$ or less synthesized images, no performance improvement is seen, whereas when augmenting it with more than $30K$, the performance is increasing, following the increase in the granularity of synthesized data. Adding synthesized data to the training set seems to be beneficial for improving the performance and thus the improvement would be much greater if we added more than $60K$ (if we had more neutral expressions), although probably at a given point, a plateau would be reached (considering the large training set that consists of around $1M$ images).

For the RECOLA database, we use the ResNet-GRU network. When augmenting the dataset with up to $30K$ synthesized images, there exists small performance improvement, whereas when augmenting it with more than $30K$, the performance is continuously increasing following the increase in the granularity of synthesized data; this increase is large. This is expected, since $120K$ frames are not sufficient for training a network for regression and additionally, $170K$ frames are not either.

For the AffectNet database, we use the VGG-FACE network. After adding $10K$ synthesized images, the performance starts to increase. This increase continues to happen as more data are added until the training set has been augmented with $1.5M$ data. If more data are added, the performance does not change, implying that a plateau has been reached. The final performance improvement is large.

For the AFEW-VA database, we use the VGG-FACE-GRU network. The improvement is systematically very significant. When adding more than $30K$ data, the increase in performance is more rapid. The performance is expected to continue increasing while more data are added, as both the initial training set of around $23K$ frames and the augmented set of around $135K$ frames are not large enough to train a DNN for regression.

Categorical affect generation

For the RAF-DB database, we use the VGG-FACE network and the performance is measured in terms of the mean

Table 10 CCC and MSE evaluation of valence and arousal predictions provided by the: (i) AffWildNet (trained on Aff-Wild), (ii) ResNet-GRU (trained on RECOLA) and (iii) the VGG-FACE baseline (trained on AffectNet); these networks are tested on the synthesized images by StarGAN, GANimation and our approach

Databases	Methods	Evaluation metrics	Networks		
			AffWildNet (Kollias et al. 2019)	ResNet-GRU (Kollias et al. 2019)	the VGG-FACE baseline
Aff-Wild	StarGAN	CCC	0.33–0.26	–	–
		MSE	0.21–0.19		
	GANimation	CCC	0.35–0.28	–	–
		MSE	0.19–0.16		
	Ours	CCC	**0.43–0.33**	–	–
		MSE	**0.15–0.13**		
RECOLA	StarGAN	CCC	–	0.29–0.23	–
	GANimation	CCC	–	0.28–0.22	–
	Ours	CCC	–	**0.34–0.33**	–
AffectNet	StarGAN	CCC	–	–	0.23–0.23
		MSE			0.34–0.37
	GANimation	CCC	–	–	0.26–0.21
		MSE			0.31–0.38
	Ours	CCC	–	–	**0.39–0.31**
		MSE			**0.27–0.28**

Each score is shown in the format: valence value-arousal value
Bold values correspond to the results of the best methods

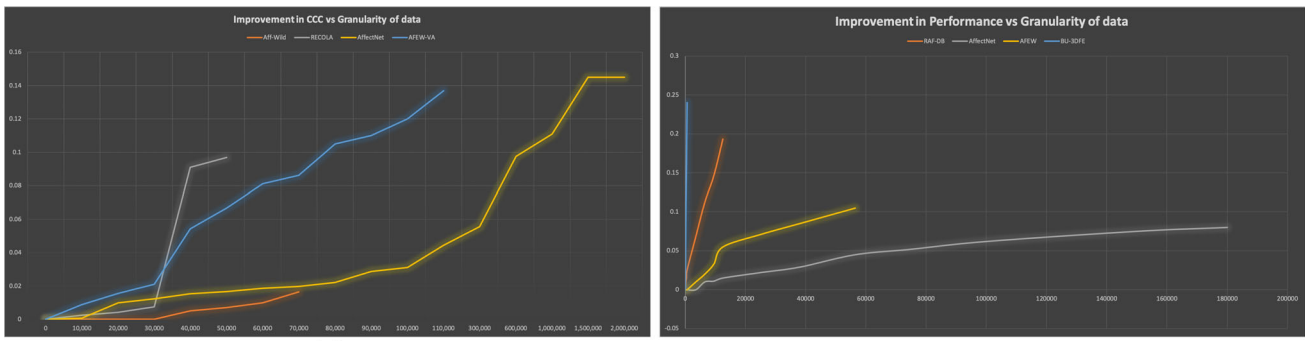
Table 11 CCC and MSE evaluation of valence and arousal predictions provided by the: (i) AffWildNet, (ii) ResNet-GRU and (iii) the VGG-FACE baseline; these networks are trained on the synthesized images by StarGAN, GANimation and our approach; these networks are evaluated on the Aff-Wild, RECOLA and AffectNet test sets

Databases	Methods	Evaluation metrics	Networks		
			AffWildNet	ResNet-GRU	VGG-FACE baseline
Aff-Wild	StarGAN	CCC	0.16–0.13	–	–
		MSE	0.18–0.17		
	GANimation	CCC	0.17–0.14	–	–
		MSE	0.17–0.15		
	Ours	CCC	**0.21–0.20**	–	–
		MSE	**0.15–0.12**		
RECOLA	StarGAN	CCC	–	0.19–0.10	–
	GANimation	CCC	–	0.17–0.10	–
	Ours	CCC	–	**0.23–0.14**	–
AffectNet	StarGAN	CCC	–	–	0.37–0.29
		MSE			0.23–0.21
	GANimation	CCC	–	–	0.40–0.31
		MSE			0.20–0.19
	Ours	CCC	–	–	**0.45–0.35**
		MSE			**0.18–0.17**

Each score is shown in the format: valence value-arousal value
Bold values correspond to the results of the best methods

Table 12 Databases used in our approach and the different values of N for each one; N denotes a subset of the synthesized data (per database) by the proposed approach

Databases	N synthesized data
Aff-Wild	$N \in \{10K, 20K, 30K, 40K, 50K, 60K\}$
RECOLA	$N \in \{10K, 20K, 30K, 40K, 50K\}$
AffectNet (VA)	$N \in \{10K, 20K, 30K, 40K, 50K, 60K, 70K, 80K, 90K, 100K, 110K, 300K, 600K, 1M, 1.5M, 2M, 2.5M\}$
AFEW-VA	$N \in \{10K, 20K, 30K, 40K, 50K, 60K, 70K, 80K, 90K, 100K, 110K\}$
RAF-DB	$N \in \{200, 400, 600, 3.5K, 6.5K, 9.5K, 12.5K\}$
AffectNet (Expressions)	$N \in \{6.5K, 12.5K, 25K, 38K, 56.5K, 75K, 100K, 150K, 180K\}$
AFEW	$N \in \{3.5K, 6.5K, 12.5K, 25K, 38K, 56.5K\}$
BU-3DFE	$N \in \{200, 400, 600\}$



(a) **(b)**

Fig. 20 Improvement in network performance versus amount of synthesized data; criteria: **a** mean/average CCC of VA in Aff-Wild, RECOLA, AffectNet, AFEW-VA and **b** mean diagonal value of the confusion matrix for RAF-DB, F1 score for AffectNet, Total Accuracy for AFEW and BU-3DFE

Table 13 Age analysis in terms of CCC and MSE for the dimensionally annotated databases

Databases	Ages	# Test samples	# Synthesized samples	Network-augmented		Network	
				CCC	MSE	CCC	MSE
Aff Wild	20–29	29,013	5301	**0.61–0.38**	**0.101–0.063**	0.59–0.37	0.102–0.066
	30–39	99,962	23,427	**0.66–0.47**	**0.077–0.054**	0.61–0.44	0.088–0.066
	40–49	44,727	21,831	**0.50–0.48**	**0.048–0.033**	0.46–0.44	0.054–0.044
	50–59	41,748	9120	**0.58–0.40**	**0.074–0.054**	0.57–0.38	0.075–0.057
	Total	215,450	59,679	**0.60–0.45**	**0.074–0.051**	0.57–0.43	0.080–0.060
RECOLA	30–39	90,000	11,001	**0.61–0.38**	–	0.60–0.34	–
	40–49	15,000	16,188	**0.43–0.24**	–	0.36–0.19	–
	50–59	7500	11,742	**0.49–0.20**	–	0.44–0.10	–
	Total	112,500	38,931	**0.55–0.31**	–	0.53–0.27	–
AffectNet	0–19	172	118,902	**0.67–0.55**	**0.105–0.156**	0.61–0.41	0.127–0.181
	20–29	1179	714,232	**0.60–0.53**	**0.128–0.159**	0.51–0.36	0.170–0.193
	30–39	1218	814,588	**0.64–0.54**	**0.139–0.145**	0.50–0.39	0.193–0.169
	40–49	762	452,504	**0.64–0.61**	**0.149–0.134**	0.49–0.44	0.202–0.166
	50–59	569	229,938	**0.58–0.53**	**0.161–0.149**	0.47–0.34	0.216–0.181
	60–89	600	146,091	**0.62–0.44**	**0.145–0.167**	0.51–0.29	0.200–0.195
	Total	4500	2476,235	**0.62–0.54**	**0.141–0.150**	0.50–0.37	0.190–0.180
AFEW-VA	20–29	766	17,466	0.46–0.60	0.192–0.084	–	–
	30–39	1990	36,388	0.51–0.62	0.254–0.080	–	–
	40–49	1558	34,906	0.59–0.47	0.211–0.076	–	–
	50–59	946	15,102	0.74–0.85	0.215–0.045	–	–
	60–79	396	4102	0.63–0.45	0.236–0.100	–	–
	Total	5646	108,864	0.57–0.59	0.226–0.075	–	–

Bold values correspond to the results of the best methods

diagonal value of the confusion matrix. The increase in performance is almost linear as more data are used. The final performance gain is great. RAF-DB is a very small database (of size about $12K$ images) and therefore if we had more data to add, the performance would further improve.

In the AffectNet database, we use the VGG-FACE network and performance is measured in terms of the F1 score. Increasing the amount of added data provides a respective increase in the performance. After adding $60K$ images the performance is increasing at a lower rate. It should be mentioned that the results include erroneous classification of the contempt class. If we synthesized samples of the contempt class as well, the network would provide a higher performance; but this is beyond the scope of the current paper.

In the AFEW database, we use the VGG-FACE network; the performance measure is total accuracy. The performance is increasing with the addition of more data. The performance increase is significant. The AFEW database is a small database (of size about $40K$ images) and therefore adding data is expected to increment the performance.

In the BU-3DFE database, we use the VGG-FACE network; the performance measure is total accuracy. There is a huge and rapid increase in network performance with the addition of data. This is explained by the very small size of BU-3DFE (around $2K$) which makes it impossible to train a neural network on it.

General deductions that can be made from Fig. 20:

- the smaller the size of the database, the bigger and faster the increase in performance would be, when augmenting it with synthesized data from our approach
- the improvement in performance is small if we augment the training set with few data in proportion to its size
- in dimensionally annotated databases, a plateau is reached and no further improvement is seen when a lot of data (about $\geq 1.5M$ in our case) are added
- the performance due to data augmentation does not increase commensurately; in the AffectNet database (mainly in the valence-arousal case) the gain yielded by data augmentation saturates as N increases
- generally, the performance increase is larger in categorically annotated databases in comparison to dimensionally annotated ones. This is an interesting result, since it indicates that synthesizing more data is needed in the latter case, to make the data distribution more dense.

Table 14 Age analysis for the categorically annotated databases; criterion for RAF-DB & AffectNet is F1 score, for AFEW & BU-3DFE is total accuracy; AFEW test samples refer to: number of videos (frames)

Databases	Ages	# Test samples	# Synthesized samples	VGG-FACE-augmented Performance metric	VGG-FACE Performance metric
RAF-DB	10–19	168	210	**0.631**	0.446
	20–29	911	2250	**0.813**	0.556
	30–39	998	4320	**0.739**	0.498
	40–49	516	3606	**0.744**	0.511
	50–59	258	1776	**0.709**	0.440
	60–69	149	552	**0.657**	0.550
	70–79	68	128	**0.904**	0.635
	Total	3068	12,828	**0.738**	0.505
AffectNet	0–19	152	12,516	**0.593**	0.453
	20–29	882	45,182	**0.584**	0.477
	30–39	962	55,513	**0.593**	0.518
	40–49	594	27,632	**0.586**	0.532
	50–59	431	20,204	**0.648**	0.606
	60–69	289	11,178	**0.564**	0.498
	70–79	161	3582	**0.466**	0.398
	80–89	29	618	**0.448**	0.410
	Total	3500	176,425	**0.590**	0.510
AFEW	20–29	29 (1536)	6474	**0.379**	0.241
	30–39	156 (8568)	22,518	**0.455**	0.333
	40–49	132 (7803)	17,934	**0.553**	0.439
	50–59	57 (3202)	7482	**0.474**	0.456
	60–79	16 (764)	2106	**0.438**	0.313
	Total	390 (21,873)	56,514	**0.484**	0.379
BU–3DFE	20–29	115	192	**0.800**	0.600
	30–39	100	240	**0.820**	0.570
	40–49	100	120	**0.800**	0.550
	50–59	100	30	**0.790**	0.490
	60–70	85	18	**0.600**	0.400
	Total	500	600	**0.768**	0.528

Bold values correspond to the results of the best methods

6.5 Effect of Subjects' Age in Classification and Regression Results

It is interesting to quantitatively assess the effect of age on the performance of the proposed approach. However, not all databases contain age information about their subjects. To achieve this, we trained an age estimator on them. In more detail, we trained a Wide Residual Network (WideResNet) (Zagoruyko and Komodakis 2016) on the union of IMDB (Rothe et al. 2015) and Adience datasets (Eidinger et al. 2014) (so that the training dataset contained an adequate number of images of people under the age of 25) and tested it on WIKI (Rothe et al. 2015). Then we applied this estimator on the test sets of the examined databases.

Table 13 shows, for each dimensionally annotated database (Aff-Wild, RECOLA, AffectNet and AFEW-VA), the estimated age groups (we split the age values into appropriate groups so that each group contained a significant amount of samples), the number of test samples that are within the age groups, the number of synthesized by our approach samples for each age group, different evaluation metrics (CCC and MSE) for each age group in two cases: when a network trained only with the training set of each database was used (denoted as 'Network' in Table 13) and when the same network was trained with the training set augmented with our approach's synthesized data (denoted as 'Network-Augmented' in Table 13). For Aff-Wild and AFEW-VA, the VGG-FACE-GRU network was used, for RECOLA the ResNet-GRU and for AffectNet the VGG-FACE.

Table 14 is similar to Table 13 with the difference being that it refers to categorically annotated databases (RAF-DB, AffectNet, AFEW and BU-3DFE). In this case, the evaluation metrics are the F1 score for RAF-DB and AffectNet, and the total accuracy for AFEW and BU-3DFE. The 'VGG-FACE-Augmented' refers to the case in which the VGG-FACE network is trained on the union of training set of each database and data synthesized by our approach.

By observing the two Tables 13 and 14, it is seen that augmenting the training dataset with the images generated by our approach is beneficial in all age groups, both for regression and classification. It would be interesting to focus on specific groups, such as very young (< 20 years old) in RAF-DB and AffectNet, each containing more than 150 subjects, or elderly (e.g., 70–79 years old) in AffectNet, also containing more than 150 subjects. In the former case, the F1 value improved from about 0.45 to 0.6; the F1 values over all categories improved from about 0.51 to 0.66. Although the F1 values in the very young category were lower than the mean F1 values over all ages, the improvement in both cases was similar. A similar observation can be made in the latter case, of elderly persons, with the F1 value in the category being improved from about 0.4 to 0.47. Although these values were lower than the total F1 values over all ages, which were 0.51 and 0.59 respectively, the improvement in these cases was similar as well. This verifies the above-mentioned observation that the proposed approach for data augmentation can be also beneficial in cases where the number of available samples is rather small.

7 Conclusions and Future Work

A novel approach to generate facial affect in faces has been presented in this paper. It leverages a dimensional emotion model in terms of valence and arousal or the six basic expressions, and a large scale 4D face database, the 4DFAB. We performed dimensional annotation of the 4DFAB and used the facial images with their respective annotations to generate mean faces on a discretized 2-D affect space.

A methodology has been proposed using these mean faces to synthesize faces with affect, both categorical or dimensional, static or dynamic. Using a given neutral image and the desired affect, which can be a Valence Arousal pair of values, a path in the 2D VA space, or one of the basic expression categories, the proposed approach performs face detection and landmark localization on the input neutral image, fits a 3D Morphable Model on the resulting image, deforms the reconstructed face, adds the input affect and blends the new face with the given affect into the original image.

An extensive experimental study has been conducted, providing both qualitative and quantitative evaluations of the proposed approach. The qualitative results show the achieved higher quality of the synthesized data compared to GAN-generated facial affect. The quantitative results are based on using the synthesized facial images for data augmentation and training of Deep Neural Networks over eight databases, annotated with either dimensional or categorical affect labels. It has been shown that, over all databases, the achieved performance is much higher than(i) the performance of the respective state-of-the-art methods, (ii) the performance of the same DNNs with data augmentation provided by the StarGAN and GANimation networks.

In our future work we will extend this approach to synthesize, not only dimensional, but also Facial Action Units in faces. In this way a Global Local synthesis of facial affect will be possible, through a unified modeling of global dimensional emotion and local action unit based facial expression synthesis. Another future direction will be to generate faces of different genders and human races.

Acknowledgements The works of Dimitrios Kollias, as well as Evangelos Ververas were funded by Teaching Fellowships of Imperial College London. We also thank the NVIDIA Corporation for donating a Titan X GPU. Additionally, we would like to thank the reviewers for their valuable comments that helped us to improve this paper.

References

Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., & Lucey, S. (2017). Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 1609–1618).

Alabort-i-Medina, J., Antonakos, E., Booth, J., Snape, P., & Zafeiriou, S. (2014). Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the ACM international conference on multimedia, MM'14* (pp. 679–682). New York, NY, USA: ACM. https://doi.org/10.1145/2647868.2654890.http://doi.acm.org/10.1145/2647868.2654890.

Alabort-i Medina, J., & Zafeiriou, S. (2017). A unified framework for compositional fitting of active appearance models. *International Journal of Computer Vision*, *121*(1), 26–64.

Amberg, B., Romdhani, S., & Vetter, T.: Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE (2007).

Antoniou, A., Storkey, A., & Edwards, H. (2017). Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340.

Averbuch-Elor, H., Cohen-Or, D., Kopf, J., & Cohen, M. F. (2017). Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, *36*(6), 196.

Bach, F., Jenatton, R., Mairal, J., & Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, *4*(1), 1–106. https://doi.org/10.1561/2200000015.

Blanz, V., Basso, C., Poggio, T., & Vetter, T. (2003). Reanimating faces in images and video. In *Computer graphics forum* (Vol. 22, pp. 641–650). Wiley Online Library.

Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., & Zafeiriou, S. (2017). 3d face morphable models "in-the-wild". In IEEE *Conference on computer vision and pattern recognition (CVPR)*. https://arxiv.org/abs/1701.05360.

Booth, J., Roussos, A., Ponniah, A., Dunaway, D., & Zafeiriou, S. (2018). Large scale 3d morphable models. *International Journal of Computer Vision*, *126*(2–4), 233–254.

Booth, J., & Zafeiriou, S. (2014). Optimal uv spaces for facial morphable model construction. In *2014 IEEE international conference on image processing (ICIP)* (pp. 4672–4676). IEEE.

Cao, C., Hou, Q., & Zhou, K. (2014). Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, *33*(4), 43.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 67–74). IEEE.

Chang, W. Y., Hsu, S. H., & Chien, J. H. (2017). Fatauva-net : An integrated deep learning framework for facial attribute recognition, action unit (au) detection, and valence-arousal estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshop*.

Cheng, S., Kotsia, I., Pantic, M., Zafeiriou, S., & (2018). 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *IEEE conference on computer vision and pattern recognition (CVPR 2018)*. Utah, US: Salt Lake City.

Chew, S. W., Lucey, P., Lucey, S., Saragih, J., Cohn, J. F., Matthews, I., et al. (2012). In the pursuit of effective affective computing: The relationship between features and registration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *42*(4), 1006–1016.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8789–8797).

Cosker, D., Krumhuber, E., & Hilton, A. (2011). A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *2011 international conference on computer vision* (pp. 2296–2303). IEEE.

Deng, J., Zhou, Y., Cheng, S., & Zaferiou, S.: Cascade multi-view hourglass model for robust 3d face alignment, pp. 399–403 (2018). https://doi.org/10.1109/FG.2018.00064.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., & Gedeon, T. (2017). From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction* (pp. 524–528). ACM.

Ding, H., Sricharan, K., & Chellappa, R. (2018). Exprgan: Facial expression editing with controllable expression intensity. In *Thirty-second AAAI conference on artificial intelligence*.

Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, *9*(12), 2170–2179.

Fried, O., Shechtman, E., Goldman, D. B., & Finkelstein, A. (2016). Perspective-aware manipulation of portrait photos. *ACM Transactions on Graphics (TOG)*, *35*(4), 128.

Garrido, P., Valgaerts, L., Rehmsen, O., Thormahlen, T., Perez, P., & Theobalt, C. (2014). Automatic face reenactment. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4217–4224).

Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., & Freeman, W. T. (2018). Unsupervised training for 3d morphable model regression. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117–124). Springer.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image and Vision Computing*, *28*(5), 807–813.

Knyazev, B., Shvetsov, R., Efremova, N., & Kuharenko, A. (2017). Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. arXiv preprint arXiv:1711.04598.

Kollias, D., Nicolaou, M. A., Kotsia, I., Zhao, G., & Zafeiriou, S. (2017). Recognition of affect in the wild using deep neural networks. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1972–1979). IEEE.

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., et al. (2019). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, *127*(6–7), 907–929.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, *65*, 23–36.

Kuipers, J. B., et al. (1999). *Quaternions and rotation sequences* (Vol. 66). Princeton: Princeton University Press.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and Emotion*, *24*(8), 1377–1388.

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*(1), 255–268.

Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2584–2593). IEEE.

Liu, X., Mao, T., Xia, S., Yu, Y., & Wang, Z. (2008). Facial animation by optimized blendshapes from motion capture data. *Computer Animation and Virtual Worlds*, *19*(3–4), 235–245.

Ma, L., & Deng, Z. (2019). Real-time facial expression transformation for monocular rgb video. In *Computer Graphics Forum* (Vol. 38, pp. 470–481). Wiley Online Library.

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook*. Springer.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Mohammed, U., Prince, S. J., & Kautz, J. (2009). Visio-lization: Generating novel facial images. *ACM Transactions on Graphics (TOG)*, *28*(3), 57.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. arXiv preprint arXiv:1708.03985.

Neumann, T., Varanasi, K., Wenger, S., Wacker, M., Magnor, M., & Theobalt, C. (2013). Sparse localized deformation components. *ACM Transactions on Graphics (TOG)*, *32*(6), 179.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *BMVC* (Vol. 1, p. 6).

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance* (pp. 296–301). IEEE.

Pérez, P., Gangnet, M., & Blake, A. (2003). Poisson image editing. In *ACM SIGGRAPH 2003 Papers, SIGGRAPH'03* (pp. 313–318). ACM, New York, NY, USA. https://doi.org/10.1145/1201775.882269.http://doi.acm.org/10.1145/1201775.882269.

Pham, H. X., Wang, Y., & Pavlovic, V. (2018). Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. arXiv preprint arXiv:1803.07716.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 818–833).

Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., & Wang, H. (2018). Geometry-contrastive gan for facial expression transfer. arXiv preprint arXiv:1802.01822.

Reed, S., Sohn, K., Zhang, Y., & Lee, H. (2014). Learning to disentangle factors of variation with manifold interaction. In *International conference on machine learning* (pp. 1431–1439).

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (pp. 1–8). IEEE.

Rothe, R., Timofte, R., & Van Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 10–15).

Russell, J. A. (1978). Evidence of convergent validity on the dimensions of affect. *Journal of Personality and Social Psychology*, *36*(10), 1152.

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., et al. (2008). Bosphorus database for 3d face analysis. In *European workshop on biometrics and identity management* (pp. 47–56). Springer.

Shang, F., Liu, Y., Cheng, J., & Cheng, H. (2014). Robust principal component analysis with missing data. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management, CIKM'14* (pp. 1149–1158). New York, NY, USA: ACM. https://doi.org/10.1145/2661829.2662083. http://doi.acm.org/10.1145/2661829.2662083.

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems* (pp. 3483–3491).

Song, L., Lu, Z., He, R., Sun, Z., & Tan, T. (2018). Geometry guided adversarial facial expression synthesis. In *2018 ACM multimedia conference on multimedia conference* (pp. 627–635). ACM.

Sun, Y., Chen, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification–verification. In *Advances in neural information processing systems* (pp. 1988–1996).

Susskind, J. M., Hinton, G. E., Movellan, J. R., & Anderson, A. K. (2008). Generating facial expressions with deep belief nets. In *Affective computing*. IntechOpen.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701–1708).

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2387–2395).

Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., & Nießner, M. (2018). Headon: real-time reenactment of human portrait videos. *ACM Transactions on Graphics (TOG)*, *37*(4), 164.

Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, *28*(6), 902–913.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., et al. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 3–10). ACM.

Vielzeuf, V., Pateux, S., & Jurie, F. (2017). Temporal multimodal fusion for video emotion classification in the wild. arXiv preprint arXiv:1709.07200.

Wheeler, M. D., & Ikeuchi, K. (1995). *Iterative estimation of rotation and translation using the quaternion*. Department of Computer Science, Carnegie-Mellon University.

Whissell, C. M. (1989). The dictionary of affect in language. In *The measurement of emotions* (pp. 113–131). Elsevier.

Wright, S. J., Nowak, R. D., & Figueiredo, M. A. T. (2009). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, *57*(7), 2479–2493. https://doi.org/10.1109/TSP.2009.2016892.

Wu, W., Zhang, Y., Li, C., Qian, C., & Change Loy, C. (2018). Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 603–619).

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition, 2006. FGR 2006* (pp. 211–216). IEEE.

Zafeiriou, S., Kollias, D., Nicolaou, M. A., Papaioannou, A., Zhao, G., & Kotsia, I. (2017). Aff-wild: Valence and arousal 'in-the-wild'challenge. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1980–1987). IEEE.

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. arXiv preprint arXiv:1605.07146.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342.

Zhou, Y., & Shi, B. E. (2017). Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In *2017 seventh international conference on affective computing and intelligent interaction (ACII)* (pp. 370–376). IEEE.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).

Zhu, X., Liu, Y., Li, J., Wan, T., & Qin, Z. (2018). Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 349–360). Springer.