# Photorealistic Facial Synthesis in the Dimensional Affect Space

Dimitrios Kollias[1], Shiyang Cheng[1], Maja Pantic[1], and Stefanos Zafeiriou[1,2]

[1] Department of Computing, Imperial College London, UK
[2] Centre for Machine Vision and Signal Analysis, University of Oulu, Finland
{dimitrios.kollias15,shiyang.cheng11,s.zafeiriou}@imperial.ac.uk

**Abstract.** This paper presents a novel approach for synthesizing facial affect, which is based on our annotating 600,000 frames of the 4DFAB database in terms of valence and arousal. The input of this approach is a pair of these emotional state descriptors and a neutral 2D image of a person to whom the corresponding affect will be synthesized. Given this target pair, a set of 3D facial meshes is selected, which is used to build a blendshape model and generate the new facial affect. To synthesize the affect on the 2D neutral image, 3DMM fitting is performed and the reconstructed face is deformed to generate the target facial expressions. Last, the new face is rendered into the original image. Both qualitative and quantitative experimental studies illustrate the generation of realistic images, when the neutral image is sampled from a variety of well known databases, such as the Aff-Wild, AFEW, Multi-PIE, AFEW-VA, BU-3DFE, Bosphorus.

**Keywords:** dimensional facial affect synthesis, valence, arousal, discretization, blendshape models, 3DMM fitting, 4DFAB, Aff-Wild, AFEW, AFEW-VA, Multi-PIE, BU-3DFE, Bosphorus, deep neural networks

## 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 to various applications including face recognition [6] [28], face verification [33] [35], emotion prediction [19] [21] [22], expression database generation, augmentation and entertainment.

This paper describes a novel approach that takes an arbitrary face image with a neutral facial expression and synthesizes a new face image of the same person, but with a different expression, generated according to a dimensional emotion representation model. This problem cannot be tackled using small databases with labeled facial expressions, because it would be really difficult to disentangle facial expression and identity information through them. Our approach is based on the analysis of a large 4D facial database, the 4DFAB [8], which we appropriately annotated and used for facial expression synthesis on a given subject's face. 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 activation of the emotion) [39] [31], has been used to annotate the large amounts of facial images, since 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.

Section 2 refers to related work that has been published with reference to facial expression synthesis. Section 3 presents the proposed approach for generating facial affect. We describe the annotation and use of the 4DFAB database, and provide the pipeline of our approach in detail. In Section 4, we provide an evaluation of the Valence - Arousal discretization and modeling procedure. Then, we synthesize facial affect on a variety of neutral faces from ten different databases (annotated either using a categorical or dimensional emotion model). By using augmented data of faces from two in-the-wild databases, we train a deep neural network to predict the valence and arousal values in these databases. Experimental results show that the proposed approach manages to synthesize photorealistic facial affect, which can be used to improve the accuracy of valence and arousal prediction. Conclusions and future work are presented in Section 5.

## 2 Related Work

In the past several years, facial expression synthesis has been an active research topic. All facial expression synthesis methods that were proposed in the past two decades were roughly split into two categories. The first category is mainly using computer graphics techniques in order to directly warp input faces to target expressions [47] [42] [44] or re-use sample patches of existing images [26]. The second one synthesizes images with attributes that are predefined [10] [34] through the creation of generative models. For the first category, a lot of research efforts have been devoted to finding the correspondence between the target images and existing facial textures. Earlier approaches mostly generated new expressions by either compositing face patches from an existing expression database [26] [17], or warping face images via optical flow [42] [43] and feature correspondence [36], or creating fully textured 3D facial models [30] [3]. In particular, [44] proposed to learn the optical flow using a variational autoencoder. Although this kind of methods can usually produce realistic images with high resolution, the elaborated complex processes often result in highly expensive computations. These works have shown either how to synthesize facial expressions on virtual agents [48], or how to transfer facial expressions between different subjects, i.e., facial reenactment [37]. However, synthesizing accurately a wide variety of facial expressions on arbitrary real faces is considered an open problem and has much room for improvement.

Due to this difficulty, the second category of methods has initially focused on using deconvolutional neural networks (DeCNNs) [3] or deep belief nets (DBNs)

[3] https://zo7.github.io/blog/2016/09/25/generating-faces.html

[34], generating faces through interpolation of the facial images in their training set. This, however, makes them inherently unsuited for facial expression generation in the case of unseen subjects. With the recent development of Generative Adversarial Networks (GANs) [13], image editing has migrated from pixel-level manipulations to semantic-level ones. GANs have been successfully applied to face image editing, for modification of facial attributes [41] [12], age modeling [49] and pose adjustment [16]. These methods generally use the encoder of the GAN to find a low-dimensional representation of the face image in a latent space, manipulate the latent vector and then decode it to generate the new image.

Popular approaches shift the latent vector along a direction corresponding to semantic attributes [24] [44], or concatenate attribute labels with it [49] [41]. Adversarial discriminator networks are used, either at the encoder to regularize the latent space [25], or at the decoder to generate blur-free and realistic images [24] or at both encoder and decoder, such as the Conditional Adversarial Autoencoder. All of these approaches require large training databases so that identity information can be properly disambiguated. Otherwise, when presented with an unseen face, the network tends to generate faces which look like the closest subject in the training database. It has been proposed to handle this problem by warping images, rather than generating them from the latent vector [44]. This approach achieves a high interpolation quality, but requires that the input expression is known and fails when generating facial expressions that are far apart, e.g. angry faces from smiling ones. Moreover, it is hard to take fine-grain control of the synthesized images, e.g., widen the smile or narrow the eyes.

The proposed approach has quite a few novelties. First of all, it is the first time, to the best of our knowledge, that the dimensional model of affect is taken into account when synthesizing images. All other models are producing synthesized images according to the seven basic, or a few more, expressions. Our approach, as verified in the experimental section of this paper, produces a large number of different expressions given a valence and arousal pair of values in the continuous 2D domain. Also, 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. What is more, until now, there has not been any attempt to use the blendshape models like we propose for the synthesis of the data. Finally, the proposed approach works well, when presented with a neutral image either from a controlled or from an in-the-wild database and with different head poses of the person appearing in that image.

## 3 The Proposed Approach

### 3.1 The 4DFAB database

The 4DFAB database [8] is the first large scale 4D face database designed for biometrics applications and facial expression analysis. It consists of 180 subjects (60 females, 120 males) aging from 5 to 75. 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, where spontaneous expressions are elicited by emotional video clips watching. In this paper, we use all the 1,580 spontaneous expression sequences for our emotion analysis and synthesis; these sequences cover a wide range of expressions as defined in [11].

To be able to develop the novel expression synthesis method, we annotate these dynamic 3D sequences (over 600,000 frames), in terms of valence and arousal emotion dimensions, using the tool described in [46]. Valence and arousal values range in [-1,1]. Examples are shown in Fig. 1. In the rest of the paper, when we refer to the 4DFAB database, we mean the 600,000 frames which are annotated with categorical expressions, as well as 2D valence and arousal (V-A) emotion values.

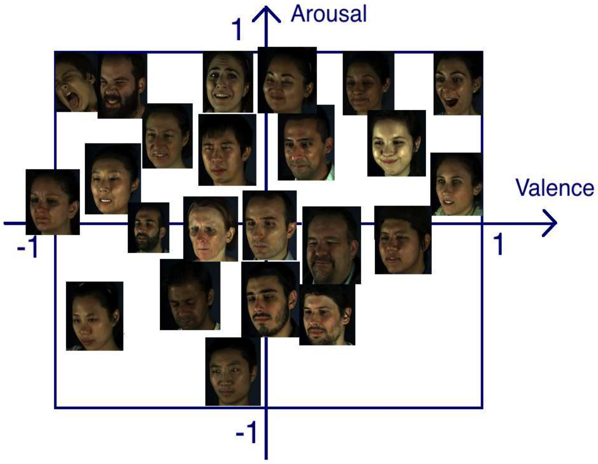
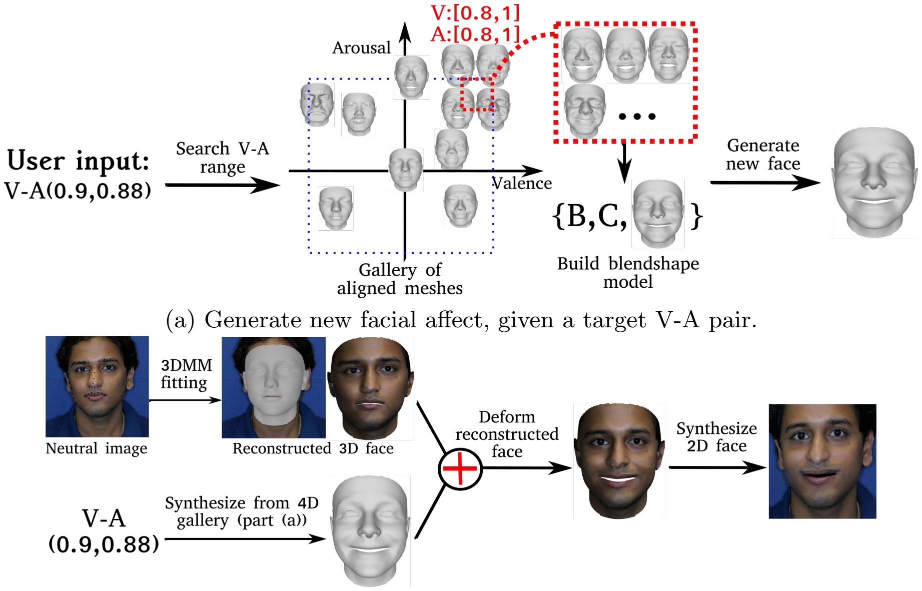


**Fig. 1.** The 2D Valence-Arousal Space and some representatives frames of 4DFAB

As each 3D face in 4DFAB differs in the number, as well as topology of vertex, we need to first correlate all these meshes to an universal coordinate frame - namely a 3D face template. This step is usually called establishing dense correspondence. We follow the same UV-based registration approach in [8] to bring all the 600,000 meshes into full correspondence with the mean face of LSFM [5]. As a result, we create a new set of 600,000 3D faces that share identical mesh topology, while maintaining their original facial expressions; we will use them as our 3D facial expression gallery for the facial affect synthesis.

### 3.2 The methodology pipeline

The main novelty and contribution of this paper comes from the development of a fully automatic facial affect synthesis framework (depicted in Fig. 2). In the first part (Fig. 2(a)), assuming that the user inputs a target V-A pair, we aim at generating semantically correct 3D facial affect from our 4D gallery. There are two key stages in this pipeline. The first includes the data selection from the 4D face gallery and the utilization of these data. To this end, we discretize the 2D Valence-Arousal (V-A) Space into 100 classes (see Fig. 3 for visualization). Each class contains aligned meshes that are associated with the corresponding V-A pairs; all these V-A pairs lie within the area of this class. Therefore, when a user provides us with a V-A pair, we find its class and retrieve the data belonging to this class. We then build a blendshape model using these data and compute the mean face. Eventually, using this blendshape model, we can generate an unseen 3D face with affect. The details of this part are described in Fig. 2(a) and Section 3.3.



(a) Generate new facial affect, given a target V-A pair.

(b) synthesize facial affect on a 2D neutral face.

**Fig. 2.** Two main parts in our facial affect synthesis framework: (a) generating new facial affect from our 4D face gallery, given a target V-A value pair provided by the user; (b) synthesizing the facial affect (from part (a)) on an arbitrary 2D neutral face.

Fig. 2(b) describes the procedure of synthesizing a new facial affect to an arbitrary 2D face. As described previously, given a target V-A pair, we create an unseen expressive face without any identity, gender and age information. In this part, we want to transfer the affect of this expressive face to the face of another person, after which, we render a 2D expressive face without loss of identity. Three processing steps are needed to achieve this goal. The first is to perform 3DMM fitting [4] to estimate the 3D shape of target face. The second step is to transfer the facial affect from synthetic 3D face to the reconstructed 3D face. Finally, we rasterize the new 3D face with affect to the original image frame. We will describe this procedure in details in Section 3.4.

### 3.3 Generation of new 3D facial affect from 4DFAB

**Discretizing the 2D Valence-Arousal Space** At first, we discretize the 2D Valence-Arousal Space into 100 classes, with each one covering a square of size 0.2 × 0.2 and including a sufficient number of data. Although the number of classes can be increased to further categorize the facial affect, it might not provide a better result. This is because, if each class contained few examples, it would be more likely that the identity information is incorporated. However, our synthetic facial affects should only describe the expression associated with the designated V-A value pair, rather than any of the identity, gender and age information. Fig. 3 shows on the right side the histogram of annotations (of 4DFAB database) of the discretized Valence-Arousal Space and on the left side the corresponding mean blendshapes of various classes of this Space. Expression blendshape models provide an effective way to parameterize facial behaviors and are frequently used in many computer vision applications. We choose to build the localized blendshape model [27] to describe our selection of V-A examples. For each 3D mesh, we subtracted it from the neutral mesh of the corresponding sequence and created a set of $m$ difference vectors $\mathbf{d}_i \in \mathbb{R}^{3n}$ which were then stacked into a matrix $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_m] \in \mathbb{R}^{3n \times m}$, where $n$ is number of vertices in our mesh. Afterwards, a variant of sparse Principal Component Analysis (PCA) was applied to our data matrix $\mathbf{D}$ to identify sparse deformation components $\mathbf{C} \in \mathbb{R}^{h \times 1}$:

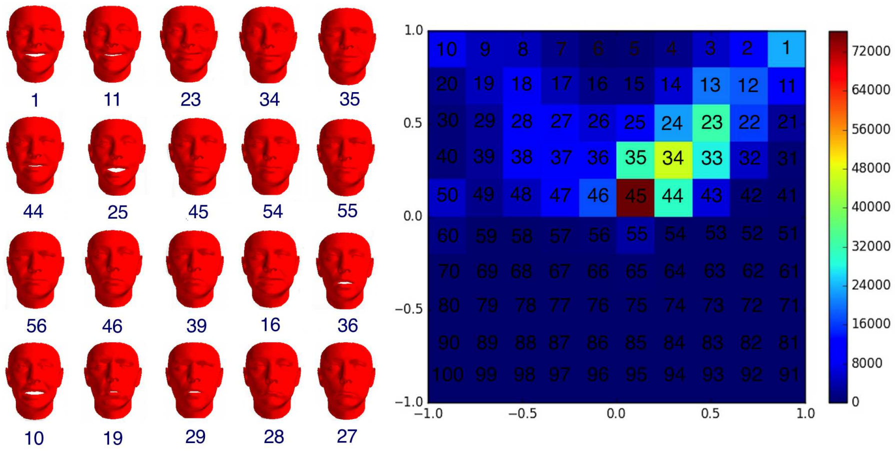


**Fig. 3.** The mean shapes of our blendshape models and their corresponding areas in the 2D Valence-Arousal Space, which is shown as a 2D histogram of annotations of the 4DFAB database.

$$\arg\min \|\mathbf{D} - \mathbf{BC}\|_F^2 + \Omega(\mathbf{C}) \quad \text{s.t. } \mathcal{V}(\mathbf{B}), \tag{1}$$

here, the constraint $\mathcal{V}$ can be either $\max(|\mathbf{B}_k|) = 1, \forall k$ or $\max(\mathbf{B}_k) = 1, \mathbf{B} \geq 1, \forall k$, where $\mathbf{B}_k \in \mathbb{R}^{3n \times 1}$ denotes the $k^{th}$ components of sparse weight matrix

$\mathbf{B} = [\mathbf{B}_1, \cdots, \mathbf{B}_h]$. The selection of these two constraints depends on our actual usage; the major difference is that the latter one allows negative weights and therefore enables deformation towards both directions, which is useful for describing shapes like muscle bulges. The regularization of sparse components $\mathbf{C}$ is performed with $\ell 1/\ell 2$ norm $[40, 1]$. To permit more local deformations from the model, additional regularization parameters were added into $\Omega(\mathbf{C})$. To solve for the optimal $\mathbf{C}$ and $\mathbf{B}$, an iterative alternating optimization is employed, please refer to [27] for more details.

### 3.4 Facial affect synthesis for arbitrary 2D image

Given a facial expression synthesis based on the valence-arousal value pair, we aim at modifying the face in an arbitrary 2D image and generating a new facial image with affect. This procedure consists of three steps: (1) fit a 3D morphable model on the image; (2) generate facial affect on the reconstructed 3D face; (3) blend the new face into the original image. Specifically, we started by performing a 3DMM fitting [4] on a 2D facial image, and retrieved a reconstructed 3D face with the texture sampled from the original image. Next, we calculated the facial deformation by subtracting the synthetic face with the LSFM template, and imposed this deformation on the reconstructed mesh. This far, we have generated a new 3D face with certain affect; the last step would be rendering it back to the original 2D image, where a Poisson image blending [29] is employed to produce a natural and realistic result.

## 4 Experimental Study

### 4.1 Discovering shared information between 3D data and Valence-Arousal

In the first experiment, we wanted to prove the validity of our Valence-Arousal modeling and synthesis approach. This could be verified by showing that there is shared information between our 3D data and Valence-Arousal through a correlation analysis.

Due to the high volume and dimensionality of our 3D data, it is intractable to directly perform typical correlation analysis. Hence, we first built a powerful expression blendshape model using the apex frames of posed expression sequences from the 4DFAB; in total, 12,000 expressive 3D meshes are selected, with 2,000 for each of the six basic expressions. Then, we projected our 3D data to its subspace and retrieved the sparse representations for future analysis. We experimented with different number of components (i.e. 84, 150, 200, 300, 500) of the blendshape model to select the best configuration.

Next, we split the data into 2 sets: the training and the test set, containing 480,000 and 120,000 frames respectively, in a subject independent manner, meaning that one person could only appear in the training or test set, but not on both of them. As we have found a compact representation of our data, Canonical

Correlation Analysis (CCA) [15] can be performed on the training set and their corresponding valence and arousal values. CCA is a shared-space component analysis method, which recovers the loadings to project two data matrices on a subspace where the linear correlation is maximized. This can be interpreted as discovering the shared information conveyed by all the data (or views).

After CCA, we reduced the dimensions of our data to 2. Then, on the training set, we performed Support Vector Regression (SVR) [2] with Radial Basis Function (RBF) kernel to map those 2 dimensions to the valence and arousal values. In order to examine whether our 3D data highly correlate to the Valence-Arousal labels, we predicted the V-A values of the test data using the aforementioned models (CCA and SVR), and compare the predictions with our annotated V-A labels. This comparison was performed with respect to two criteria: Concordance Correlation Coefficient and the usual Mean Squared Error. The Concordance Correlation Coefficient (CCC) can be defined as follows:

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

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

Table 1 shows those two criteria for the test set when we keep different numbers of principal components for our expression blendshape model. We can observe that with 200 components, highest correlation between the data and V-A labels was achieved, as well as lowest prediction error. By selecting this value, we ensured that the proposed synthesis approach is valid.

**Table 1.** CCC and MSE evaluation of valence & arousal predictions on the test set when we keep different number of principal components in PCA

| No. of principal components to keep | CCC | | MSE | |
|---|---|---|---|---|
| | Valence | Arousal | Valence | Arousal |
| 84 | 0.63 | 0.68 | 0.107 | 0.046 |
| 150 | 0.65 | 0.68 | 0.099 | 0.041 |
| **200** | **0.66** | **0.69** | **0.097** | **0.040** |
| 300 | 0.35 | 0.30 | 0.127 | 0.058 |
| 500 | 0.31 | 0.22 | 0.129 | 0.061 |

### 4.2 Databases used for affect synthesis evaluation

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 10 databases.

**1) Multi-PIE** [14]: It contains 755,370 images (3072x2048) of 337 people. Pose, illumination, and expression are the key factors of the database. 15 view points, 19 illuminations and 7 expressions are recorded in a controlled environment.

**2) Face place:** This database [4] contains photographs of many different individuals in various types of disguises, such that, for each individual, there are multiple photographs in which hairstyle and/or eyeglasses have been changed/added. It consists of 1,284 images of Asian, 937 images of African-American, 3,362 images of Caucasian, 494 images of Hispanic and 497 images of multiracial people. All images show posed expression.

**3) 2D Face Sets:** We used 3 subsets from the 2D Face Sets database[5].

Iranian women: It consists of 369 color images (1200x900) of 34 women. People display mostly smile and neutral expression in each of five poses.

Nottingham scans: It has 100 monochrome images (50 men, 50 women) in neutral and frontal pose. The image resolution varies from 358x463 to 468x536.

Pain expressions: It consists of 599 color images (720x576) of 13 women and 10 men. They usually display two of the six basic emotions (anger, disgust, fear, sad, happy, surprise) plus pain 10 expressions. Profile neutral and 45 degrees images are available.

**4) FEI:** The FEI database [38] is a Brazilian face database that contains a set of face images taken between June 2005 and March 2006. 200 individuals were recorded, and each one has 14 images, resulting in 2,800 images of size 640x480. All images were color and taken against a white background in an upright frontal position with profile rotation of up to 180°. The subjects are mostly students and staff at FEI, between 19 and 40 years old with distinct appearance, hairstyle and adorns. The number of male and female subjects are both 100.

**5) Aff-Wild:** Aff-Wild [20] [46] consists of 298 Youtube videos, with $1,200,000$ frames in total. The length of each video varies from 10 seconds to 15 minutes. These videos contain spontaneous facial behaviors elicited by a variety of stimuli in arbitrary recording conditions. There are 200 subjects (130 males and 70 females) from different ethnicities. Aff-Wild serves as the benchmark of the first Affect-in-the-wild Challenge[6] [18]. For each video, there are 8 annotators to annotate the valence and arousal, in the range of $[-1, +1]$.

**6) AFEW 5.0:** This database is a dynamic facial expressions corpus (used in EmotiW Challenge 2017 [9]) consisting of 1,809 nearly real world scenes from movies and reality TV shows. There are over 330 subjects aging from 1 to 77. The database is split into three sets: training (773 videos), validation (383 videos) and test set (653 videos). It is a challenging database because both training and validation sets are mainly from the movies, while 114 out of 653 test videos are from TV. Annotations of neutral and 6 basic expressions are provided.

[4] 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/

[5] http://pics.stir.ac.uk

[6] https://ibug.doc.ic.ac.uk/resources/first-affect-wild-challenge

**7) AFEW-VA:** Recently, a part of the AFEW database has been annotated in terms of Valence and Arousal, thus creating the AFEW-VA [23] database. It includes 600 video clips selected from films with real-world conditions, i.e., occlusions, illumination and body movements. The length of each video ranges from around 10 frames to over 120 frames. This database consists of per-frame annotations of V-A. In total, more than 30,000 frames were annotated for affect prediction of V-A, using discrete values in the range of $[-10, +10]$.

**8) BU-3DFE:** BU-3DFE database [45] is the first 3D facial expression database, which includes 2,500 expressive meshes from 100 subjects (56 females, 44 males) with age from 18 to 70. The subjects are from various ethnic/racial ancestries. They recorded 6 articulated expressions (happiness, disgust, fear, angry, surprise and sadness) with 4 intensities; also, there is a neutral 3D scan per subject.

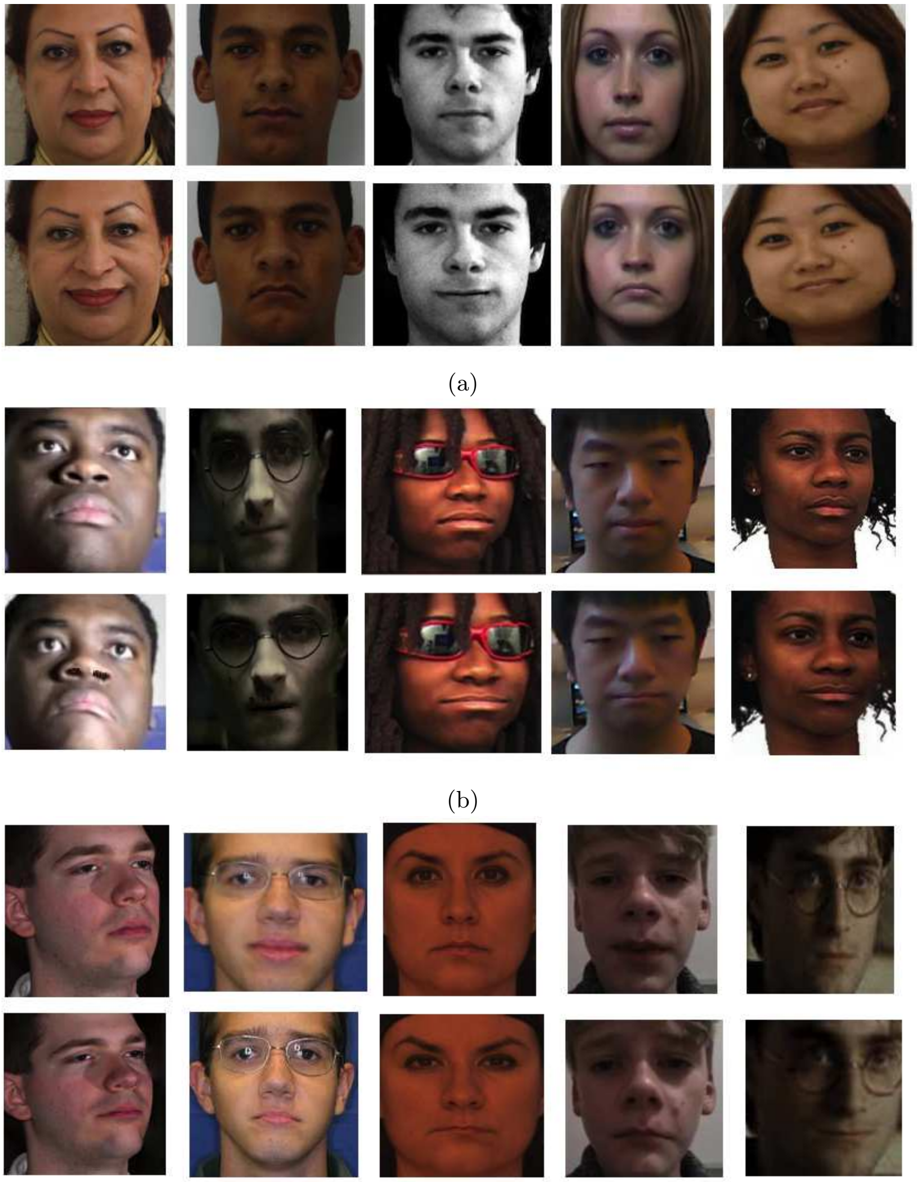
**9) Kinect Fusion ITW:** The KF-ITW database [4] is the first Kinect 3D database captured under relatively unconstrained conditions. This database consists of 17 different subjects performing some expressions (neutral, happy, surprise) under various illumination conditions.

**10) Bosphorus:** The Bosphorus database [32] consists of 105 subjects in various poses, expressions and occlusion conditions. 18 men had beard/moustache and 15 others had short facial hair. There are 60 men and 45 women, they are mostly between 25 and 35. Majority of them are Caucasian. 27 professional actors/actresses are incorporated in the database. The number of total face scans is 4,652, each scan has been manually labeled with 24 facial landmarks.

### 4.3 Qualitative evaluation of the facial affect synthesis

We used all the above-mentioned databases to supply the proposed approach with 'input' neutral faces. We synthesized the emotional state of specific V-A value pairs for these images. One important task during this facial affect synthesis procedure is to preserve identity, age and gender of the original face. Instead of finding the closest matching sample (or K-nearest samples) for the given V-A pair, we categorized our 3D data based on the 2D Valence-Arousal Space (as shown in Fig. 3) and employed the mean expression of the area that contains the target V-A pair.

Fig. 4 is split into three parts. In each part, the top row illustrates some neutral images sampled from each of the aforementioned databases and the bottom one shows the respective synthesized images. Fig. 5 shows the neutral images on the left side, and the synthesized images of different valence and arousal values on the right. It could be observed that our synthetic images are identity preserving, realistic and vivid. We showed that the proposed framework works well for images from both in-the-wild and controlled databases. This suggests that we could effectively synthesize facial affect regardless of different image conditions (e.g., occlusions, illumination and head poses).



(a)

(b)

(c)

**Fig. 4. (a)-(c).** Synthesis of facial affect across all databases: on top rows are the neutral and on the bottom are the corresponding synthesized images.

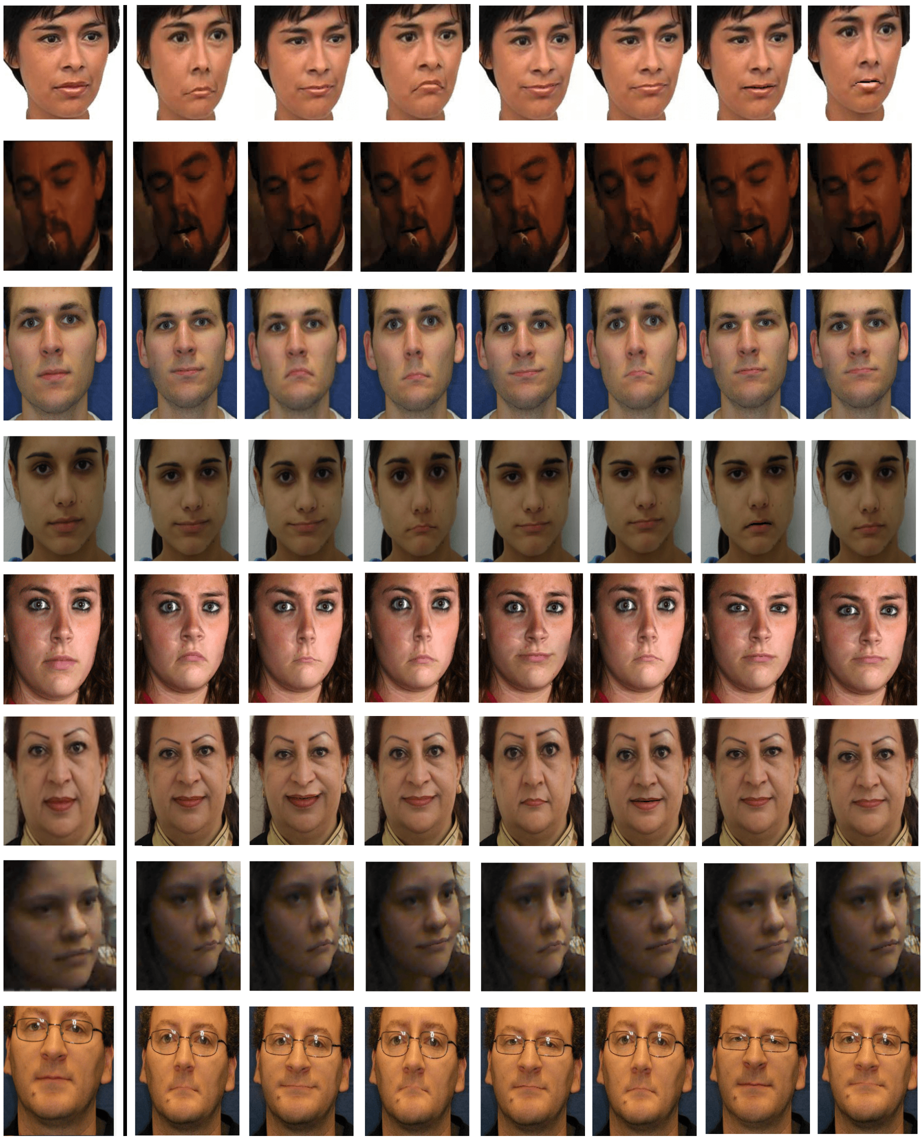


**Fig. 5.** Synthesis of facial affect: on the left side are the neutral 2D images and on the right the synthesized images with different levels of affect

### 4.4 Quantitative evaluation of the facial affect synthesis

**Leveraging synthetic data for training Deep Neural Networks** We used the synthetic faces to train deep neural networks for valence and arousal prediction on two facial affect databases annotated in terms of valence and arousal, the Aff-Wild and AFEW-VA. Our first step is to select neutral frames from these two databases. Specifically, we selected frames with zero valence and arousal (human inspection was also conducted to make sure they are neutral faces), then, for each frame, we synthesized facial affect using the mean blendshape (as shown in Fig. 3) and assigned the median valence and arousal value of that class.

**Experiments and data augmentation on the AFEW-VA** Following our approach, we created 108,864 synthetic images from the AFEW-VA database, a number that is 3.5 times bigger than its original size. For training, we used the CNN-RNN (VGG-Face-GRU) architecture described in [18]. Similarly to [23], 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 2 shows a comparison of the performance of our network with the best results reported in [23]. Those results are in terms of the Pearson Correlation Coefficient criterion (Pearson CC), defined as follows:

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3}$$

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.

**Table 2.** Pearson Correlation Coefficient evaluation of valence & arousal predictions provided by the best architecture in [23] vs the network trained on the augmented dataset created by our approach. Note that valence and arousal values are in $[-10, 10]$.

| Group | Pearson CC | | MSE | |
|---|---|---|---|---|
| | Valence | Arousal | Valence | Arousal |
| best of [23] | 0.407 | 0.45 | 6.96 | 4.97 |
| Our network (trained on the augmented dataset) | **0.542** | **0.589** | 4.75 | 2.74 |

**Experiments and data augmentation on the Aff-Wild** : Following our approach, we created 60,135 synthetic images from the Aff-Wild database. We added those images to the training set of the first Affect-in-the-wild Challenge. It should be noticed that these images were synthesized from neutral faces found only in the training set of the challenge. The network we employed here was the the same CNN-RNN (VGG-Face-GRU) architecture described in [18]. Table 3 shows a comparison of the performance of our network trained with the augmented data with the best results reported in [18] and the results of the winner of the Aff-Wild Challenge [7] (Method FATAUVA-Net).

**Table 3.** Concordance Correlation Coefficient evaluation of valence & arousal predictions provided by the CNN-RNN trained on the Aff-Wild dataset augmented with images synthesized by our approach vs methods [7] & [18]. Note that valence and arousal values are in $[-1, 1]$.

| | CCC | | MSE | |
|---|---|---|---|---|
| | Valence | Arousal | Valence | Arousal |
| FATAUVA-Net [7] | 0.396 | 0.282 | 0.123 | 0.095 |
| [18] | 0.570 | 0.430 | 0.080 | 0.060 |
| Our network trained on the augmented dataset | **0.591** | **0.442** | 0.074 | 0.051 |

From both tables, it can be verified that the network trained on the augmented, with synthetic images, dataset, outperformed the networks trained without them. This implies that, by augmenting the original training set, our methodology improved the network performance. It should be noted that the boost in performance is greater when the number of augmented images is much greater than the number of images in the dataset (which is the case of AFEW-VA that contains 30,000 frames, while the augmented set included 109,000 more frames).

## 5 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, and a large scale 4D face database, the 4DFAB. An efficient method has been developed for matching different blendshape models on large amounts of images extracted from the database and using these to render the appropriate facial affect on a selected face. A variety of faces and facial expressions has been examined in the experimental study, from ten databases showing expressions according to dimensional, but also categorical emotion models. The proposed approach has been successfully applied to faces from all databases, being able to render photorealistic facial expressions on them.

In our future work we will extend this approach to synthesize, not only dimensional affect in faces, but also Facial Action Units. 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.

**Acknowledgements.** The authors would also like to thank Evangelos Ververas for assisting with the affect synthesis. The work of Dimitris Kollias was funded by a Teaching Fellowship of Imperial College London. The work of S. Cheng is funded by the EPSRC project EP/J017787/1 (4D-FAB) and EP/N007743/1 (FACER2VM).

## References

1. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning **4**(1), 1–106 (Jan 2012). https://doi.org/10.1561/2200000015, http://dx.doi.org/10.1561/2200000015
2. Basak, D., Pal, S., Patranabis, D.C.: Support vector regression. Neural Information Processing-Letters and Reviews **11**(10), 203–224 (2007)
3. Blanz, V., Basso, C., Poggio, T., Vetter, T.: Reanimating faces in images and video. In: Computer graphics forum. vol. 22, pp. 641–650. Wiley Online Library (2003)
4. Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S.: 3d face morphable models "in-the-wild". In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017), https://arxiv.org/abs/1701.05360
5. Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S.: Large scale 3d morphable models. International Journal of Computer Vision **126**(2-4), 233–254 (2018)
6. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. pp. 67–74. IEEE (2018)
7. Chang, W.Y., Hsu, S.H., Chien, J.H.: 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 (2017)
8. Cheng, S., Kotsia, I., Pantic, M., Zafeiriou, S.: 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City, Utah, US (June 2018)
9. Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T.: 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 (2017)
10. Ding, H., Sricharan, K., Chellappa, R.: Exprgan: Facial expression editing with controllable expression intensity. arXiv preprint arXiv:1709.03842 (2017)
11. Du, S., Tao, Y., Martinez, A.: Compound facial expressions of emotion. Proceedings of the National Academy of Sciences of the United States of America **111**(15), 1454–1462 (2014). https://doi.org/10.1073/pnas.1322355111
12. Ghodrati, A., Jia, X., Pedersoli, M., Tuytelaars, T.: Towards automatic image editing: Learning to see another you. arXiv preprint arXiv:1511.08446 (2015)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
14. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28**(5), 807–813 (2010)
15. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis; an overview with application to learning methods. Technical report, Royal Holloway, University of London (May 2003), http://eprints.soton.ac.uk/259225/
16. Huang, R., Zhang, S., Li, T., He, R., et al.: Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. arXiv preprint arXiv:1704.04086 (2017)
17. Jonze, S., Cusack, J., Diaz, C., Keener, C., Kaufman, C.: Being John Malkovich. Universal Studios (1999)

18. Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G., Zafeiriou, S.: Recognition of affect in the wild using deep neural networks. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. pp. 1972–1979. IEEE (2017)
19. Kollias, D., Tagaris, A., Stafylopatis, A.: On line emotion detection using retrainable deep neural networks. In: Computational Intelligence (SSCI), 2016 IEEE Symposium Series on. pp. 1–8. IEEE (2016)
20. Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond (2018)
21. Kollias, D., Yu, M., Tagaris, A., Leontidis, G., Stafylopatis, A., Kollias, S.: Adaptation and contextualization of deep neural network models. In: Computational Intelligence (SSCI), 2017 IEEE Symposium Series on. pp. 1–8. IEEE (2017)
22. Kollias, D., Zafeiriou, S.: Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm (2018)
23. Kossaifi, J., Tzimiropoulos, G., Todorovic, S., Pantic, M.: Afew-va database for valence and arousal estimation in-the-wild. Image and Vision Computing (2017)
24. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300 (2015)
25. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015)
26. Mohammed, U., Prince, S.J., Kautz, J.: Visio-lization: generating novel facial images. ACM Transactions on Graphics (TOG) **28**(3), 57 (2009)
27. Neumann, T., Varanasi, K., Wenger, S., Wacker, M., Magnor, M., Theobalt, C.: Sparse localized deformation components. ACM Transactions on Graphics (TOG) **32**(6), 179 (2013)
28. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. vol. 1, p. 6 (2015)
29. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers. pp. 313–318. SIGGRAPH '03, ACM, New York, NY, USA (2003). https://doi.org/10.1145/1201775.882269, http://doi.acm.org/10.1145/1201775.882269
30. Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.H.: Synthesizing realistic facial expressions from photographs. In: ACM SIGGRAPH 2006 Courses. p. 19. ACM (2006)
31. Russell, J.A.: Evidence of convergent validity on the dimensions of affect. Journal of personality and social psychology **36**(10), 1152 (1978)
32. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: European Workshop on Biometrics and Identity Management. pp. 47–56. Springer (2008)
33. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems. pp. 1988–1996 (2014)
34. Susskind, J.M., Hinton, G.E., Movellan, J.R., Anderson, A.K.: Generating facial expressions with deep belief nets. In: Affective Computing. InTech (2008)
35. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: 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 (2014)
36. Theobald, B.J., Matthews, I., Mangini, M., Spies, J.R., Brick, T.R., Cohn, J.F., Boker, S.M.: Mapping and manipulating facial expression. Language and speech **52**(2-3), 369–386 (2009)

37. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. **34**(6), 183–1 (2015)
38. Thomaz, C.E., Giraldi, G.A.: A new ranking method for principal components analysis and its application to face image analysis. Image and Vision Computing **28**(6), 902–913 (2010)
39. Whissell, C.M.: The dictionary of affect in language. In: The measurement of emotions, pp. 113–131. Elsevier (1989)
40. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing **57**(7), 2479–2493 (July 2009). https://doi.org/10.1109/TSP.2009.2016892
41. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: European Conference on Computer Vision. pp. 776–791. Springer (2016)
42. Yang, F., Bourdev, L., Shechtman, E., Wang, J., Metaxas, D.: Facial expression editing in video using a temporally-smooth factorization. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 861–868. IEEE (2012)
43. Yang, F., Wang, J., Shechtman, E., Bourdev, L., Metaxas, D.: Expression flow for 3d-aware face component transfer. ACM Transactions on Graphics (TOG) **30**(4), 60 (2011)
44. Yeh, R., Liu, Z., Goldman, D.B., Agarwala, A.: Semantic facial expression editing using autoencoded flow. arXiv preprint arXiv:1611.09961 (2016)
45. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on. pp. 211–216. IEEE (2006)
46. Zafeiriou, S., Kollias, D., Nicolaou, M., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-wild: Valence and arousal 'in-the-wild' challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (2017)
47. Zhang, Q., Liu, Z., Quo, G., Terzopoulos, D., Shum, H.Y.: Geometry-driven photorealistic facial expression synthesis. IEEE Transactions on Visualization and Computer Graphics **12**(1), 48–60 (2006)
48. Zhang, S., Wu, Z., Meng, H.M., Cai, L.: Facial expression synthesis based on emotion dimensions for affective talking avatar. In: Modeling machine emotions for realizing intelligence, pp. 109–132. Springer (2010)
49. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2 (2017)