AU-Blendshape for Fine-grained Stylized 3D Facial Expression Manipulation

Hao Li[1,2], Ju Dai[2,*], Feng Zhou[3], Kaida Ning[2], Lei Li[4,5], Junjun Pan[1,2,*]
[1]Beihang University, [2]Peng Cheng Laboratory, [3]North China University of Technology,
[4]University of Washington, [5]University of Copenhagen
{lih09, daij, ningkd}@pcl.ac.cn, zhoufeng@ncut.edu.cn, lilei@di.ku.dk, pan_junjun@buaa.edu.cn

Abstract

While 3D facial animation has made impressive progress, challenges still exist in realizing fine-grained stylized 3D facial expression manipulation due to the lack of appropriate datasets. In this paper, we introduce the AUBlendSet, a 3D facial dataset based on AU-Blendshape representation for fine-grained facial expression manipulation across identities. AUBlendSet is a blendshape data collection based on 32 standard facial action units (AUs) across 500 identities, along with an additional set of facial postures annotated with detailed AUs. Based on AUBlendSet, we propose AUBlendNet to learn AU-Blendshape basis vectors for different character styles. AUBlendNet predicts, in parallel, the AU-Blendshape basis vectors of the corresponding style for a given identity mesh, thereby achieving stylized 3D emotional facial manipulation. We comprehensively validate the effectiveness of AUBlendSet and AUBlendNet through tasks such as stylized facial expression manipulation, speech-driven emotional facial animation, and emotion recognition data augmentation. Through a series of qualitative and quantitative experiments, we demonstrate the potential and importance of AUBlendSet and AUBlendNet in 3D facial animation tasks. To the best of our knowledge, AUBlendSet is the first dataset, and AUBlendNet is the first network for continuous 3D facial expression manipulation for any identity through facial AUs. Our source code is available at https://github.com/wslh852/AUBlendNet.git.

1. Introduction

Along with the development of virtual reality and metaverse, 3D facial animation has become a prominent research topic [6, 9, 28]. Facial expressions, as the core expression of human emotions and intentions, play a crucial role in facial animation production. A fine-grained, controllable facial expression manipulation algorithm with customized stylization can effectively help animators accurately construct characters' emotions, thereby improving narrative effectiveness and audience resonance [32, 34].

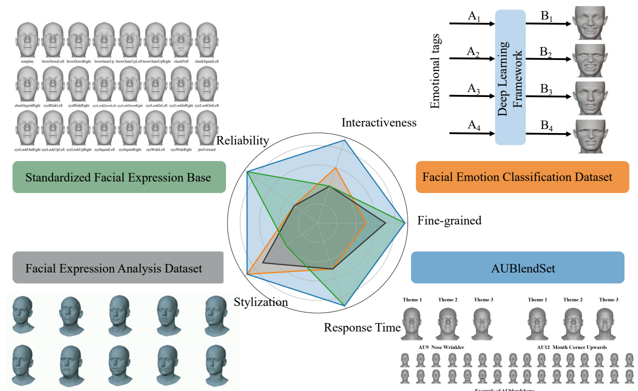

Figure 1. Comparisons of different datasets for facial expression manipulation regarding control basis reliability, stylization, interactiveness, fine-grained, and response time.

Due to diverse constraints in available facial emotion datasets, achieving fine-grained stylized facial expression manipulation confronts various challenges. As shown in Figure 1, off-the-shelf standardized facial expression control usually generates characters' expressions through predefined universal blendshapes [5, 13]. Although those methods present some flexibility, they lack precise descriptions of stylization characteristics.

Facial emotion classification datasets [37, 39] typically annotate emotions into categories that do not provide appropriate facial control bases. Besides, the lack of subtle facial action annotations during the annotation process makes it challenging to achieve fine-grained facial expression manipulation. Despite the facial expression analysis dataset using Facial Action Coding System (FACS) [8] to annotate facial details, it can not provide reliable facial manipulation like blendshape animation. Therefore, building a dataset with user-friendly interactiveness and support for stylized facial expression manipulation is essential and meaningful.

The FACS is widely adopted in behavioral science and

*Corresponding authors

psychology [8]. Based on facial muscle movements, AU-based representation offers a structured and interpretable framework capable of capturing the complexity of human expressions. with each AU representing an independent muscle. Facial expressions can be accurately described as combinations of corresponding AUs, and the FACS provides complete AU expression mapping rules. To support fine-grained and stylized 3D facial expression manipulation, we collect the facial expression control dataset, AUBlendSet, in AU-Blendshape format based on the expressive flame model. AUBlendSet contains 500 different identity themes. Each theme consists of 32 AU-Blendshape basis vectors of corresponding styles, accompanied by a set of facial poses with detailed AU annotations. Unlike existing public datasets, AUBlendSet assigns stylized AU-Blendshape control bases to each character for stylized facial expression manipulation.

Based on AUBlendSet, we propose the AUBlendNet for stylized facial expression manipulation. Our key insight is to generate stylized AU-Blendshape control bases, so AUBlendNet focuses on establishing the relationship between character identity and stylized AU-Blendshape control basis. Specifically, AUBlendNet first establishes a stylized AU-Blendshape control basis codebook, which uses the template mesh of the character as control conditions and predicts feature tokens using a non-autoregressive model. These tokens are leveraged to query the codebook sequence in discrete space to reconstruct the corresponding style control basis. Thanks to a large amount of training data from different identities, AUBlendNet is capable of quickly predicting the AU control basis for the given template mesh style and achieving fine-grained facial manipulation. We comprehensively validate the abilities of AUBlendSet and AUBlendNet through diversified downstream tasks, demonstrating their potential and value in 3D facial animation tasks. The main contributions are summarized as follows:

- We collect the AUBlendSet dataset for stylized emotional facial animation. It includes stylized AU-Blendshape control bases and detailed facial AU poses to support stylized 3D facial expression manipulation.
- We propose the AUBlendNet model guided by character template meshes to concurrently predict all AU-Blendshpe control bases, which can perform fine-grained facial expression manipulation for any identity style.
- We formulate diverse testing tasks to demonstrate the importance and applicability of AUBlendSet dataset and the powerful capability of AUBlendNet model.

2. Related Work

2.1. Facial manipulation with expressions

Facial emotion manipulation has broad applications in virtual humans, movie animations, and emotion computing [1, 12]. Extensive endeavors have been devoted to this task. ARKit [19] provides detailed descriptions of faces through 51 universal predefined blendshapes. However, with a large number of control bases and a lack of clear mapping rules for expression, it is problematic to achieve efficient, user-friendly facial manipulation. The FACS mapping table defines the mapping rules between AU and facial expression, which enables FACSGEN [13] with AU as the control basis vector to modulate the corresponding expression quickly. However, due to the consistent style of the generated effects of the universal control basis, the obtained characters lack personalized characterizations. Li et al. [15] achieve personalized facial expression manipulation to some extent by assigning mesh weights to predefined controllers. However, this method is restricted by adaptability and has limited generalization ability when facing new facial topology structures. Some deep learning methods [23, 27] use category labels as generation conditions to guide characters in generating corresponding facial emotions. Although characters can generate facial expressions of corresponding styles, the finite and discrete category labels significantly diminish the diversity and freedom of editing. Therefore, considering the current state of research, investigating facial expression manipulation techniques that have good interaction and can achieve fine-grained stylization is valuable and compelling.

2.2. Facial datasets with expressions

Existing facial datasets with expressions mainly focus on expression analysis and emotion calculation. AffectNet [20] contains approximately 400,000 facial expression images by manually labeling eight expressions, as well as valence and arousal. CK+ [17] uses facial videos to record the expression changes from neutral to 7 basic expressions. This dataset comprises detailed facial keypoints and annotations of some AUs. DISFA [18] triggers spontaneous facial expression changes in participants and labels facial action units (AUs) with different intensities. FEAFA+ [11, 36] provides a high-quality image collection with detailed annotations for 24 AUs. Despite 2D datasets being widely used due to their ease of access, annotations often lean on category labels, which cannot be directly mapped to independent expression control parameters, limiting their use in fine-grained facial expression manipulation.

The BU-3DFE [37] dataset utilizes scanning devices to capture seven facial expressions of different subjects and annotate their intensity, which is widely used in facial expression modeling, emotion analysis, and 3D expression generation. However, this dataset has no precise annotation of facial details. BP4D [39] and D3DFACS [4] use FACS to describe facial detail information. However, each sample corresponds to multiple AU annotations, resulting in strong coupling relationships between different AUs. This makes it challenging to obtain stylized emotional facial manipulation. Although FaceWarehouse [3] includes blendshape control bases for each character, its base vectors are entangled and can only be edited to a limited degree. Li et al.[16] collect 4,000 high-resolution facial scans for character identity generation. However, this dataset utilizes a universal ARKit control base for emotional expression, with no style descriptions for character. To sum up, existing datasets have difficulty supporting fine-grained stylized facial expression manipulation in terms of data annotation, collection methods, and data formats.

2.3. Applications of facial expression manipulation

Facial emotion manipulation adjusts the emotional expression of digital characters through user editing. Shao et al. [25] propose a multi-class adversarial training method that achieves fine-grained expression transfer between unpaired images by decoupling AU-related and AU-unrelated features of facial images. Bodur et al. [2] leverage depth information and depth consistency loss to manipulate facial expressions and combined confidence regularization to improve generalization ability. In speech-driven facial generation tasks, a large amount of work [23, 27, 33] uses one-hot encoding to generate facial expressions. However, due to the diversity of emotional expressions of different identities, the same emotion may correspond to different facial expression details. One-hot encoding cannot capture these fine-grained differences and often tends towards averaging during the training process, resulting in a lack of naturalness and personalization in the generated expressions. Meanwhile, in emotion and AU recognition tasks, automated annotation of AU can provide larger-scale, high-quality training data, thereby improving the accuracy and generalization ability of AU recognition and emotion analysis. We demonstrate that AuBlendSet can offer a rich collection of expression data to facilitate and enhance these applications.

3. AUBlendSet

Blendshape is widely used in facial manipulation for its efficiency and controllability. To achieve fine-grained and stylized facial manipulation, we collect the AUBlendSet dataset, which encodes expressions using stylized AU-Blendshape representations. we first perform sampling in the FLAME parameter space based on key style factors (gender, age, and facial shape proportions), and systematically generate 500 characters to ensure comprehensive coverage of style variations. The dataset comprising 206 females and 294 males with age distributions ranging from 20 to 60 years. We adopt the expressive FLAME model to represent each character, where each mesh contains 5,023 3D vertices.

The facial AU describes facial muscle movement and provides a reasonable interactive control basis for facial expression manipulation. Users can quickly manipulate and

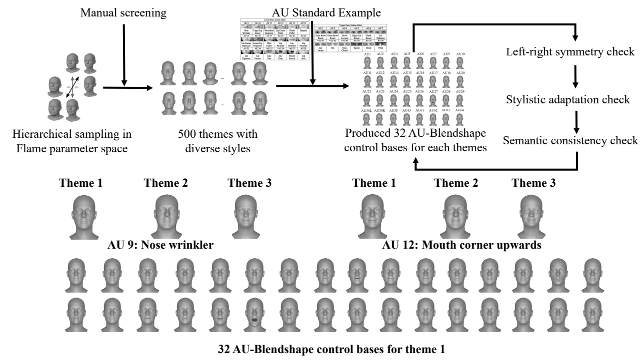

Figure 2. AUBlendSet production flow and examples. The top figure illustrates the production flow of the Blendshape. The bottom figure visualizes AU 9 (nose wrinkle) and AU 12 (mouth corner upwards) across three different themes, alongside the full set of 32 AU-Blendshape for Theme 1.

generate corresponding target expressions using the expression mapping rules defined in the FACS coding table [8]. However, due to the high similarity of some eye AU movements, they only involve varying degrees of movement. To achieve maximum fine-grained expression manipulation while ensuring good interaction, we simplify the description of eye AU to reduce redundancy while removing the AU descriptions of tongue and head posture. Thus, AUBlendSet ultimately retains 32 AU control bases. *For specific AUs and FACS table, refer to the supplementary material (SM).*

Table 1. Statistics of key indicators in the dataset.

Name	Identifiers	Blendshape	AU Annotation	Annotation Type	Editing Suitability
FACSD3D [4]	10	✗	✓	Discrete	low
CK+ [17]	97	✗	✓	Discrete	low
BP4D [38]	23	✗	✓	Discrete	low
FaceWarehouse [3]	150	✓	✗	Continuous	low
AUBlendSet	500	✓	✓	Continuous	high

For each of the 500 characters in AUBlendSet, As shown in Figure 2, following standard animation production procedures, professional artists manually create blendshape control bases for each character, guided by AU exemplars defined in Facial Action Coding System (FACS). During this process, we ensure symmetry, stylistic adaptation, and semantic consistency with AU definitions, with all results manually reviewed for clarity and structural correctness. This ensures that each theme contains a template mesh and 32 AU-Blendshape control bases that match the character's style. These control bases can accurately capture the unique facial dynamics of characters, making expression generation more natural and personalized. In addition, the dataset provides detailed annotated AU examples for each character theme, which can be used for expression generation and model performance evaluation, We present several key statistics of our dataset and related datasets in Table 1. supporting a wide range of experimental and analytical needs.

4. AUBlendNet

To achieve fine-grained, stylized, user-friendly interactive manipulation, the goal of AUBlendNet is to generate AU-Blendshape control bases that match the style of the character template, as shown in Figure 3. AUBlendNet mainly consists of the AUCodebook and StyleBlendNet. The AU-Codebook constructs an AU-style library in the hidden space using learned motion prior information. The non-autoregressive StyleBlendNet maps the input facial template to the corresponding encoding vector in the Codebook in parallel, thereby predicting the AU-Blendshape control basis that matches the target character style.

4.1. AUCodebook

Inspired by CodeTalker[35], our AUCodebook is a VAE-based Transformer architecture. It is optimized through a combined perceptual loss and a quantified training loss during training [29]. Precisely, given AU-Blendshape control bases $\mathbf{B} \in \mathbb{R}^{N\times V}$ for the identity in mesh space, the encoder $\mathcal{E}$ transforms AU-Blendshap base vectors $\mathbf{B}$ into latent representations $\mathbf{Z} = \mathcal{E}(\mathbf{B})$ and the decoder $\mathcal{D}$ reconstructs the AU-blendshape base vector from the latent representation $\hat{\mathbf{B}} = \mathcal{D}(\mathbf{Z}) = \mathcal{D}(\mathcal{E}(\mathbf{B}))$, $\mathbf{Z} \in \mathbb{R}^{N\times D}$. N is the number of AUs involved in the control, V is the 5023 facial vertices, and D is the latent space dimension.

4.2. StyleBlendNet

StyleBlendNet first expands the temporal dimension to fill in the missing time for input character templates. Subsequently, the extended sequence is projected onto the latent space, which can be described as:

$$\hat{\mathbf{Z}} = \mathrm{MLP_f}(\mathrm{TCN}(\mathbf{M})), \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{1\times W}$ is the mesh template for the input character, $\hat{\mathbf{Z}} \in \mathbb{R}^{N\times D}$ is the input token. $\mathrm{MLP_f}$ denotes the feature projection Module.

The style embedding module obtains the corresponding template style encoding, and we leverage the AdaLN-Zero block to inject the style encoding as a generation condition into the prediction process. The process is formulated as:

$$\boldsymbol{\gamma}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_2, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_2 = \mathrm{MLP_c}(\mathrm{MLP_s}(\mathbf{M})), \tag{2}$$

where $\mathrm{MLP_s}$ denote style embeding, and $\mathrm{MLP_c}$ is the AdaLN-Zeros block. The control parameters $\boldsymbol{\gamma_1}$, $\boldsymbol{\beta_1}$, $\boldsymbol{\alpha_1}$, $\boldsymbol{\gamma_2}$, $\boldsymbol{\beta_2}$, $\boldsymbol{\alpha_2}$ are then integrated into a prediction network with a transformer backbone using AdaLN-Zero [22]:

$$\hat{\mathbf{Z}}_{t,k'} = \hat{\mathbf{Z}}_{t,k-1} + \boldsymbol{\alpha}_1 \mathrm{MHSA}(\mathrm{LN}(\hat{\mathbf{Z}}_{\mathrm{t,k-1}})\boldsymbol{\gamma}_1 + \boldsymbol{\beta}_1), \tag{3}$$

$$\hat{\mathbf{Z}}_{t,k} = \hat{\mathbf{Z}}_{t,k'} + \boldsymbol{\alpha}_2 \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{Z}}_{\mathrm{t,k'}})\boldsymbol{\gamma}_2 + \boldsymbol{\beta}_2), \tag{4}$$

where MHSA denotes multi-head self-attention, LN refers to layer normalization [30]. $\hat{\mathbf{Z}}_{t,k-1}$ and $\hat{\mathbf{Z}}_{t,k}$ are the outputs of $(k-1)$-th and k-th layer of prediction network. $\hat{\mathbf{Z}}_{t,k'}$ is the intermediate variable in the k-th layer of the network.

4.3. Loss functions

We first train the AUCodebook. Similar to CodeTalker [35], we adopt a facial mesh loss and two intermediate latent-level losses to optimize our AUCodeBook:

$$\mathcal{L}_{auc} = \|\mathbf{B}-\hat{\mathbf{B}}\|_1 + \|\mathrm{SG}(\hat{\mathbf{Z}})-\mathbf{Z}_q\|_2^2 + \beta\|\hat{\mathbf{Z}}-\mathrm{SG}(\mathbf{Z}_\mathrm{q})\|_2^2, \tag{5}$$

where the first term is a facial mesh reconstruction loss, $\hat{\mathbf{Z}}$ is the AU-Blendshape embedded features. $\mathbf{Z}_q$ is the quantized AU-Blendshape embedded features, SG stands for a stop-gradient operation, and β refers to the weighting factor controlling the update rate.

Secondly, we start to optimize the StyleBlendNet. We leverage a facial mesh loss and an intermediate latent-level loss to train StyleBlendNet:

$$\mathcal{L}_{syn} = \|\mathbf{B}-\hat{\mathbf{B}}\|_2^2 + \|\hat{\mathbf{Z}} - \mathrm{SG}(\mathbf{Z}_\mathrm{q})\|_2^2. \tag{6}$$

5. Experiments

We explore the capabilities and potential of AUBlendSet and AUBlendNet through diverse tasks, including stylized facial expression manipulation, speech-driven emotional facial animation, and AUBlendNet for AU detection.

5.1. Implementation details

Our framework is built on the PyTorch platform and trained on RTX 6000 Ada. First, we optimize the AUCodebook, setting its layer numbers to 8 and latent spatial dimensions to 1024. The hyperparameter β equals 0.1. Adam optimizer is used to train AUCodebook with the learning rate 1×10^{-4}, and a batch size of 1 for 200 epochs. Subsequently, we optimize StyleBlendNet by setting its hidden space dimension to 1024. The head number and layer number are both set to 8. We train StyleBlendNet with a batch size of 1 for 400 epochs and a learning rate set of 1×10^{-5}.

5.2. Stylized facial expression manipulation

Although the universal facial blendshape control base can be applied to most tasks, in more realistic character expression manipulation processes, it often fails to reflect the unique style of character emotions. The key to stylized facial expression manipulation lies in constructing control bases that can adapt to each character's personalized characteristics and emotional expressions, thereby achieving natural and accurate emotional style shaping for any character.

Experiments on different datasets for expression manipulation. To validate the superiority of AUBlendSet with AU-Blendshape representation for stylized facial expression control, we implement AUBledNet on different datasets to validate the generation effects. We select two representative datasets with AU annotation with our dataset for comparison. The D3DFACS [4] is constructed based on the FLAME model. It consists of 110 single-AU and 352 multi-AU combination data for 10 characters. Each piece

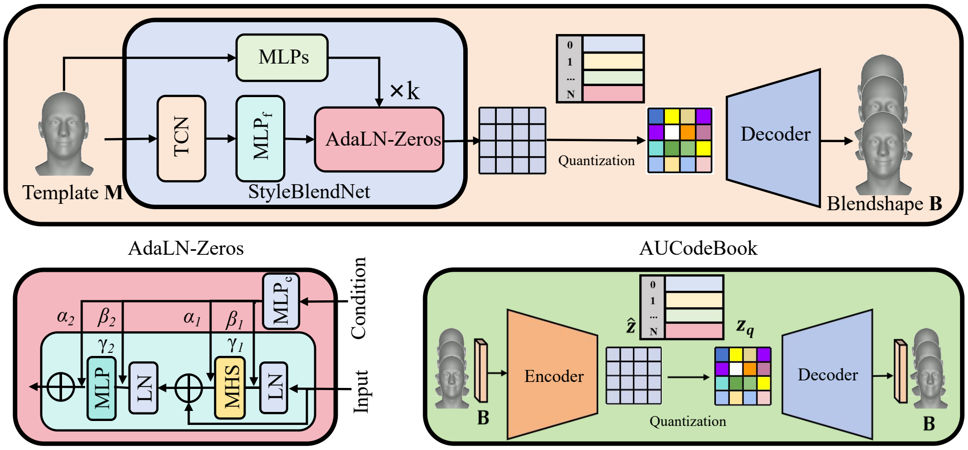

Figure 3. The pipeline of AUBlendNet, consisting of the AUCodebook and StyleBlendNet. AUCodebook is first pre-trained to construct an AU-style library in the hidden space. Given a character Template $\mathbf{M}$, StyleBlendNet first predicts stylized AU-Blendshape control bases code and then generates corresponding control bases $\mathbf{B}$ through the prior decoder in the AUCodeBook.

of data gradually changes from a neutral posture to a fully activated state to accurately capture the dynamic changes in facial expressions. The CK+ [17] contains 97 different characters, each with at least 5 different expressions and accompanied by AU annotations. CK+ has 55 single-AU and 435 multi-AU combination annotations, providing rich annotation data for facial expression analysis. We leverage EMOCA [7] to transform 2D images into 3D Flame model for subsequent model training. Since D3DFACS [4] and CK+ contain AU labels, we inject AU labels as a generation condition into the prediction process for fair comparisons. We divide each dataset into training, validation, and testing sets with an 8:1:1 ratio and train them using AUBlendNet.

Table 2. Comparison results of AUBlendNet trained on D3DFACS, CK+, and AUBlendSet on the AUBlendSet test set.

Method	$\text{MSE}_S\downarrow$ ($\times10^{-9}$)	$\text{MES}_M\downarrow$ ($\times10^{-9}$)	Inference $\downarrow$ (s)
D3DFACS[4]	226.41	68.21	2.7
CK+[17]	8490.27	2028.14	2.7
AUBlendSet	**4.55**	**7.62**	**0.3**

Table 2 records the results of AUBlendNet optimized by three datasets on the AUBlendSet test set. MSE_S denotes the generation error of a single-AU control, and MSE_M is the generation error of multi-AU controls. From Table 2, we observe that although CK+ has more identities, D3DFACS contains more samples for each theme, resulting in higher control accuracy for D3DFACS than CK+. Due to the direct AU annotations in CK+ and D3DFACS, such enumeration labels make it difficult to avoid errors when facing multi-AU-driven control processes. Furthermore, we can find that our AUBlendSet dataset is much more effective in generating results for single-AU control and multi-AU collaborative control. The superior performance can be attributed to AU-Blendshape representation, which simplifies the process of generating controllable facial postures. In the face of complex multiple AUs control, we only need to blend the offsets generated by each AU-Blendshape. By comparing the CK and D3DFACS datasets, which rely on direct AU annotations, we observe that when encountering AU control signals not seen in the training set, the optimized AUBlendNet tends to lead to biased results, favoring those generations within the training set. This phenomenon can be observed intuitively in Figure 4. The facial meshes generated from our AUBlendSet are much more consistent with the given control signals, whereas the CK+ and D3DFACS exhibit significant errors and unwanted motions. This continuous control enables fine-grained expression adjustments and smooth transitions, overcoming the limitations of discrete or static datasets and supporting more natural and flexible

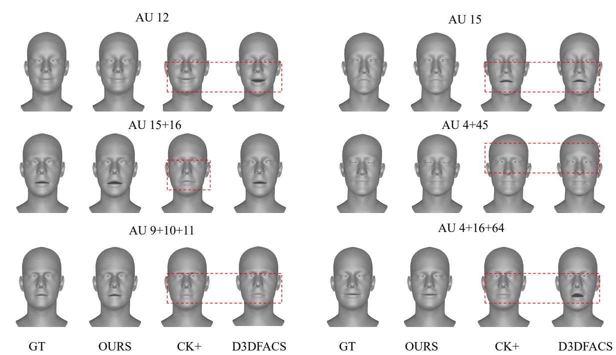

Figure 4. Visualization Comparisons of AU-guided emotional face generation results for AUBlendNet trained on different datasets.

facial animation. Those quantitative and qualitative comparison results demonstrate the superiority and generalization of our AUBlendSet dataset.

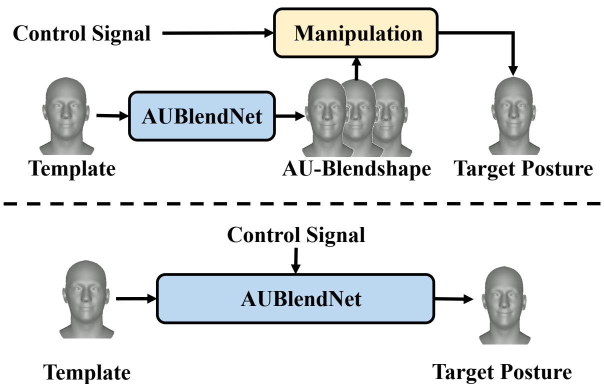

Figure 5. Workflow comparisons of AUBlendNet for AU-guided facial expression control. The up first predicts AU-Blendshape basis, then blends them with control signal to obtain target posture. The bottom directly predicts target posture given control signal.

Computation efficiency analysis. Inference generation speed is a crucial criterion for facial expression manipulation. We compare the workflows based on AU-Blendshape control basis and AU category labels in Figure 5. The response times of facial postures for inferring 7 different emotions are reported in the last column of Table 2. For the CK+[17] and D3DFACS[4] datasets labeled with AU categories, given changed AU control signals, model requires individual predictions each time, resulting in cumbersome and inefficient inference processes. In contrast, AUBlendNet trained on AUBlendSet with AU-Blendshape representations only requires one-step prediction to generate the corresponding style of AU-BlendShape base vectors and then conducts expression manipulation through blendshape different AU control signals. Specifically, on the RTX 6000 Ada, AUBlendNet trained on AUBlendSet takes 0.3 seconds to infer the character's stylized AU-BlendShape, while manipulation through AU-Blendshape basis takes only 0.002 seconds. However, the AUBlendNet trained with category labels on CK+[17] and D3DFACS[4] takes 2.7 seconds to infer 7 emotional postures. This indicates that AUBlendNet significantly improves the real-time efficiency of facial manipulation tasks by generating control bases and completing manipulation through blending AU-Blendshape bases.

Comparisons with other models. We validate the superiority of AUBlendNet for stylized facial expression control. As far as we know, there are no available methods for AU-guided 3D facial expression manipulation in deep learning. Thus, we implement the current state-of-the-art conditional generation models based on our collected AUBlendSet. The contrast models consist of a Transformer [30], a VQ-VAE [35], and a Diffusion model [28]. All compared methods establish the relationship between identity mesh and the AU-Blendshape basis vectors conditioned on character styles. We also utilize MSE_S and MSE_M as evaluation metrics to quantitatively analyze the generation quality of single-AU and multi-AU control signals. Table 3 illustrates the comparison results. From Table 3, we observe that the proposed AUBlendNet significantly outperforms the compared methods in both single-AU and multi-AU facial expression manipulation.

Table 3. Quantitative comparison of different models for AU-guided 3D facial expression manipulation on AUBlendSet.

Method	$\mathrm{MSE}_S \downarrow$ ($\times 10^{-9}$)	$\mathrm{MES}_M \downarrow$ ($\times 10^{-9}$)
Diffusion[28]	5.66	8.71
VQ-VAE[35]	8.48	12.79
Transformer[31]	86.12	96.14
AUBlendNet	**4.55**	**7.62**

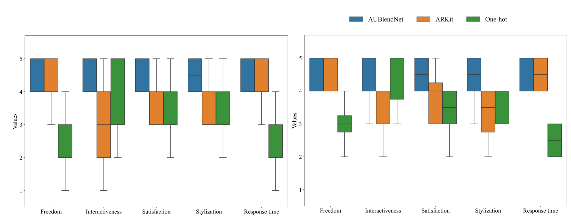

Figure 6. User study evaluation for facial expression manipulation. Right is 8 animation experts, Left is 32 general participants.

User study evaluation. Controllability and generative validity are crucial for evaluating model performance. To compare different methods, we recruited 32 general participants and 8 animation experts to evaluate the emotional face generation results of the ARKit-based method [19], the one-hot-based method [27], and our method. Participants were asked to edit facial expressions (e.g., happiness, anger) using different methods and provide qualitative ratings. We use a 5-point scale to assess different methods regarding manipulation freedom, interactiveness, generation satisfaction, style consistency, and response time (*Refer to SM*). We report the statistical results in Figure 6.

The experimental results indicate that the recruited participants and animation experts prefer our AUBlendNet, which performs well for all the criteria. Users can quickly manipulate corresponding emotions through the FACS coding table, and the offline generation of stylized control bases significantly enhances the user's interactive experience. Thanks to AUBlendNet's direct prediction of AU-BlendShape that matches the character's style, our method is comparable to the ARKit-based BlendShape method in terms of real-time and interactivity. At the same time, the

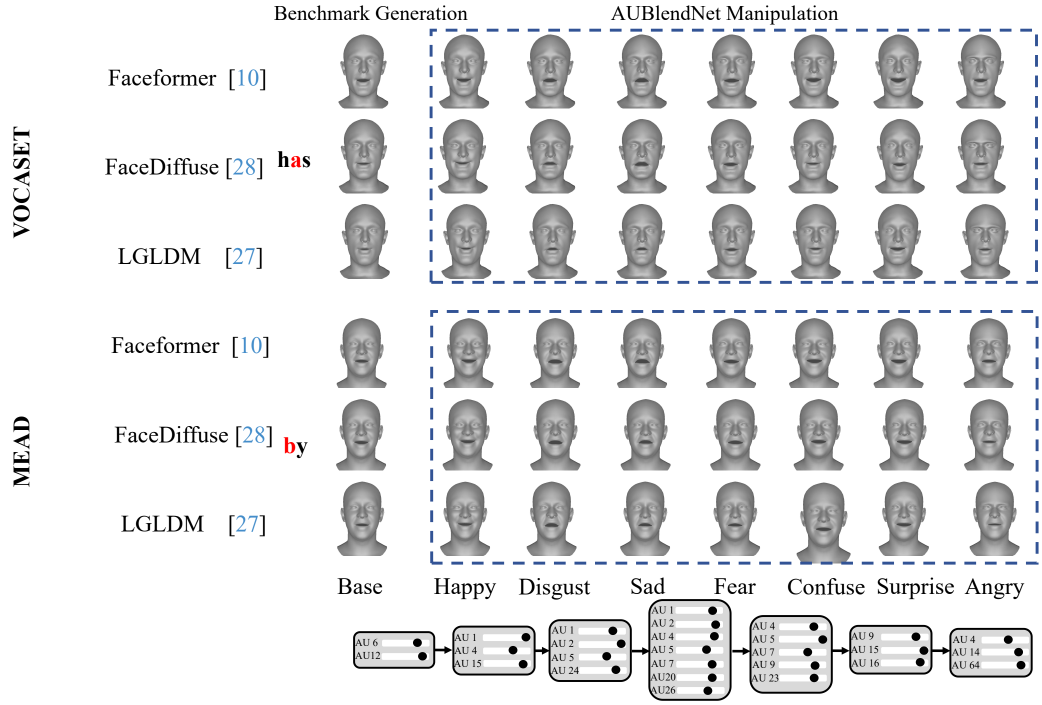

Figure 7. Visualization results of different speech-driven methods combined with AUBlendNet. The red font represents the speech information corresponding to the current frame. The lower section displays fine-grained AU-driven facial expression manipulation parameters.

AU-BlendShape control bases generated by AUBlendNet not only accurately match the corresponding AU but also conform to the style characteristics of the corresponding character, ensuring the naturalness and consistency of the generated results.

5.3. Speech-driven emotional facial animation

In existing emotional speech-driven facial animation methods, expressions are usually injected as category labels into the generation process, often leading to almost consistent expression results among all characters. However, characters often exhibit diversity and personalization in their emotional expressions. A single expression category label makes it difficult to generate unique emotional responses for different characters. To address this issue, we demonstrate an intriguing speech-driven emotional facial motion generation method based on AUBlendSet and AUBlendNet.

Through observation, we find that the leading facial changes are concentrated on the lips, and the expression of character emotions mainly focuses on the upper face and corners of the mouth in a speech-driven process. This enables the AUBlendNet predicted AU-Blendshape control basis to be easily integrated into common speech-driven facial frameworks for facial expression manipulation through blendshape control basis. This process can be described as:

$$\mathbf{M} = \mathbf{M_0} + \mathbf{Offset_{spe}} + \mathbf{Offset_{Exp}}, \tag{7}$$

where $\mathbf{M}$ is the final facial posture, $\mathbf{M_0}$ is the facial template posture, $\mathbf{Offset_{spe}}$ is the lip deviation driven by speech, and $\mathbf{Offset_{Exp}}$ is the facial expression deviations based on AU-Blendshape representation.

Visualization analysis. We select three representative speech-driven models, Faceformer [10], FaceDiffuser [28], and LGLDM [27] as benchmarks and compare the emotional generation results with and without our AUBlendNet. As shown in Figure 7, we demonstrate the neutral expression (Base) generation results without AUBlendNet. Meanwhile, we illustrate the fine-grained continuous manipulation effects of the AU-Blendshape control bases predicted by AUBlendNet on seven emotions. The stylized AU-Blendshape control base supports the adjustment of the corresponding AU and its activation intensity, making facial expression control more refined. For VOCASET (without emotion labels), AUBlendNet achieves continuous manipulation of emotions based on the FACS coding table, generating diverse facial expressions. For MEAD (with emotion labels), such as happy expressions, compared to the generated results of using One-hot labels, the programmatic control base provided by AUBlendNet can not only finely adjust the current emotion but also flexibly achieve natural transitions between different emotions.

Quantitative comparison. The stylized AU-Blendshape control bases provided by AUBlendNet can achieve di-

Table 4. Comparisons of different speech-driven models with and without combined with AUBlendNet on MEAD and VOCASET. “N/A” denotes Not Applicable, since VOCASET has no emotions.

Methods	LVE ↓ ($\times10^{-4}$)	FDD ↓ ($\times10^{-6}$)	V-LVE↓ ($\times10^{-5}$)	Diversity ↑ ($\times10^{-6}$)
MEAD				
Faceformer[10]	**2.83**	1.57	2.64	2.16
Faceformer+AUBlendNet	3.04	**1.55**	2.64	**2.74**
FaceDiffuse[28]	**1.47**	1.84	3.11	2.19
FaceDiffuse+AUBlendNet	1.76	**1.81**	3.11	**2.67**
LGLDM[27]	**2.01**	**5.79**	2.15	2.42
LGLDM+AUBlendNet	2.35	5.96	2.15	**2.91**
VOCASET				
Faceformer[10]	**4.11**	**4.67**	2.91	N/A
Faceformer+AUBlendNet	4.45	4.96	2.91	**2.17**
FaceDiffuse[28]	**3.79**	4.26	4.08	N/A
FaceDiffuse+AUBlendNet	4.04	**4.22**	4.08	**2.06**
LGLDM[27]	**3.79**	**3.92**	3.12	N/A
LGLDM+AUBlendNet	3.93	3.95	3.12	**2.31**

verse facial expression manipulation. Following [35], we adopt the commonly used evaluation metrics LVE (Lip Vertex Error), V-LVE (Velocity of Lip Vertex Error), and FDD (Upper-face Dynamics Deviation) to quantitatively evaluate the generated results. Meanwhile, we use the indicator proposed by Ren et al.[24] to evaluate the diversity of facial expression generation results.

We evaluate the speech-driven facial emotional generation results based on 7 different emotions provided by FACS on VOCASET and MEAD. As shown in Table 4. The results indicate that due to the involvement of mouth corner movements in some AUs, the LVE performance decreases slightly during facial emotion manipulation, but this does not affect the offset of lip movements velocity. In the V-LVE and FDD indicators, it can be seen that using AU-Blendshape to manipulate facial emotions can ensure lip synchronization consistency and preserve the character's speaking style. However, the additional facial emotion manipulation significantly enhances the diversity and emotional expressiveness of facial expressions, resulting in more prosperous and natural emotional responses, further enhancing the realism and interactivity of facial animations.

5.4. AUBlendNet for AU detection

In facial AU detection, manually annotating activated facial AUs is time-consuming and costly, restricting dataset expansion and diversity. Capturing high-quality facial geometric information requires professional equipment, while fine-grained AU annotation demands considerable human effort. To address the issue, we leverage AUBlendNet as a data augmentation tool. AUBlendNet is capable of effectively improving AU detection performance while enhancing the model's generalization ability and robustness.

We demonstrate the effectiveness of AUBlendNet as a data augmentation tool for AU recognition. We compare the experimental results of HMP-PS [26], LP-Net [21], and SRERL [14] on D3DFACS and DISFA [18] datasets. For the DISFA dataset, we convert it to the Flame model using EMOCA [7] and train it according to the experimental setup in the paper. The accuracy of AU recognition is evaluated by the average F1 score of 8 AUs (AU1, AU2, AU4, AU6, AU9, AU12, AU25, AU26). Results are shown in Table 5. By generating stylized AU-Blendshape, we construct more diverse AU combinations based on the original data, thereby enhancing the generalization ability of AU recognition algorithms to different character styles and expression changes. The experimental results show that using AUBlendNet-based augmented data to train the AU recognition model improves recognition precision on different datasets, especially in the case of a small amount of annotated data. The augmented data effectively alleviates the problem of high AU annotation cost, further verifying the practicality of AUBlendNet in 3D facial animation tasks.

Table 5. Comparisons of different AU detection methods with and without AUBlendNet-based data augmentation.

Methods	D3DFACS F1 score ↑	DISFA F1 score ↑
HMP-PS[26]	55.04	61.96
HMP-PS + AUBlendNet	**57.96**	**64.57**
LP-Net[21]	52.76	56.61
LP-Net + AUBlendNet	**54.96**	**59.35**
SRERL[14]	52.48	55.75
SRERL + AUBlendNet	**55.96**	**56.32**

6. Conclusion

In this paper, we contribute a facial action unit blendshape-guided dataset, AUBlendSet, for fine-grained stylized facial expression manipulation. AUBlendSet comprises 500 identity themes, each containing 32 AU-Blendshape basis vectors and corresponding facial posture sets with detailed AU annotations. Meanwhile, we propose AUBlendNet for generating stylized AU-Blendshape control bases. Given an identity mesh, AUBlendNet predicts the corresponding style of AU-Blendshape basis vectors. We achieve fine-grained stylized facial expression manipulation by linearly blending AU-Blendshape bases. We thoroughly evaluate the capabilities of AUBlendSet and AUBlendNet through diverse tasks, including stylized facial expression manipulation, speech-driven emotional facial animation, and leveraging AUBlendNet for AU detection. Extensive experiments demonstrate the potential and importance of AUBlendSet and AUBlendNet in 3D facial animation tasks.

Acknowledgement

This research is supported in part by National Key R&D Program of China (No. 2022ZD0115902), the Major Key Project of PengCheng Laboratory (PCL2023A10-2), Beijing Natural Science Foundation(4232023), China Ministry of Education Funds for Humanity, Social Sciences (24YJCZH458) and the Pioneer Centre for AI, DNRF grant number P1.

References

[1] Bita Azari and Angelica Lim. Emostyle: One-shot facial expression editing using continuous emotion parameters. In *WACV*, pages 6373–6382, 2024. 2

[2] Rumeysa Bodur, Binod Bhattarai, and Tae-Kyun Kim. 3d dense geometry-guided facial expression synthesis by adversarial learning. In *WACV*, pages 2391–2400, 2021. 3

[3] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 20(3):413–425, 2014. 3

[4] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A FACS valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *ICCV*, pages 2296–2303, 2011. 2, 3, 4, 5, 6

[5] Vittorio Cuculo and Alessandro D'Amelio. Openfacs: An open source facs-based 3d face animation system. In *ICIG*, pages 232–242, 2019. 1

[6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, pages 10101–10111, 2019. 1

[7] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: emotion driven monocular face capture and animation. In *CVPR*, pages 20279–20290, 2022. 5, 8

[8] Paul Ekman and Wallace V Friesen. Facial action coding system. *EPNB*, 1978. 1, 2, 3

[9] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, pages 18749–18758, 2022. 1

[10] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, pages 18770–18780, 2022. 7, 8

[11] Wei Gan, Jian Xue, Ke Lu, Yanfu Yan, Pengcheng Gao, and Jiayi Lyu. Feafa+: an extended well-annotated dataset for facial expression analysis and 3d facial animation. In *ICDI*, pages 307–316, 2022. 2

[12] Qi Guo and Xiaodong Gu. A styleclip-based facial emotion manipulation method for discrepant emotion transitions. In *ACCV*, pages 391–408, 2022. 2

[13] Eva Krumhuber, Lucas Tamarit, Klaus R. Scherer, and Etienne B. Roesch. Facsgen 2.0: facial expression animation based on FACS. In *FAA*, page 7, 2010. 1, 2

[14] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI*, pages 8594–8601, 2019. 8

[15] Jituo Li, Guodong Lu, and Juntao Ye. Automatic skinning and animation of skeletal models. *TVC*, 27(6-8):585–594, 2011. 2

[16] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, and Hao Li. Learning formation of physically-based face attributes. In *CVPR*, pages 3407–3416, 2020. 3

[17] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and Iain A. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR*, pages 94–101, 2010. 2, 3, 5, 6

[18] Seyed Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. DISFA: A spontaneous facial action intensity database. *TAC*, 4(2):151–160, 2013. 2, 8

[19] Timo Menzel, Mario Botsch, and Marc Erich Latoschik. Automated blendshape personalization for faithful face animations using commodity smartphones. In *VRST*, pages 22:1–22:9, 2022. 2, 6

[20] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *TAC*, 10(1):18–31, 2019. 2

[21] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *CVPR*, pages 11917–11926, 2019. 8

[22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4172–4182, 2023. 4

[23] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *ICCV*, pages 20630–20640, 2023. 2, 3

[24] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP*, pages 1–5, 2023. 8

[25] Zhiwen Shao, Hengliang Zhu, Junshu Tang, Xuequan Lu, and Lizhuang Ma. Explicit facial expression transfer via fine-grained representations. *TIP*, 30:4610–4621, 2021. 3

[26] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *CVPR*, pages 6267–6276, 2021. 8

[27] Wenfeng Song, Xuan Wang, Yiming Jiang, Shuai Li, Aimin Hao, Xia Hou, and Hong Qin. Expressive 3d facial animation generation based on local-to-global latent diffusion. *TVCG*, 2024. 2, 3, 6, 7, 8

[28] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *SIGGRAPH*, pages 13:1–13:11, 2023. 1, 6, 7, 8

[29] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, pages 6306–6315, 2017. 4

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 4, 6

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 6

[32] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *CVPR*, pages 17979–17989, 2023. 1

[33] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, pages 700–717, 2020. 3

[34] Shangfei Wang, Zhuangqiang Zheng, Shi Yin, Jiajia Yang, and Qiang Ji. A novel dynamic model capturing spatial and temporal patterns for facial expression analysis. *TPAMI*, 42 (9):2082–2095, 2020. 1

[35] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *CVPR*, pages 12780–12790, 2023. 4, 6, 8

[36] Yanfu Yan, Ke Lu, Jian Xue, Pengcheng Gao, and Jiayi Lyu. FEAFA: A well-annotated dataset for facial expression analysis and 3d facial animation. In *ICME*, pages 96–101, 2019. 2

[37] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. In *FGR*, pages 211–216, 2006. 1, 2

[38] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun J. Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *IVC*, 32(10):692–706, 2014. 3

[39] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun J. Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *IVC*, 32(10):692–706, 2014. 1, 2