Biomechanics-guided Facial Action Unit Detection through Force Modeling

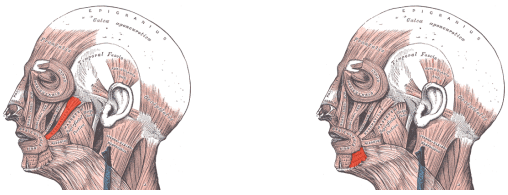
Zijun Cui[1], Chenyi Kuang[1], Tian Gao[2], Kartik Talamadupula[2], Qiang Ji[1]
[1]Rensselaer Polytechnic Institute, [2]IBM Research
ceejkl@gmail.com, {kuangc2, jiq}@rpi.edu, {tgao, krtalamad}@us.ibm.com

Abstract

Existing AU detection algorithms are mainly based on appearance information extracted from 2D images, and well-established facial biomechanics that governs 3D facial skin deformation is rarely considered. In this paper, we propose a biomechanics-guided AU detection approach, where facial muscle activation forces are modelled and are employed to predict AU activation. Specifically, our model consists of two branches: 3D physics branch and 2D image branch. In 3D physics branch, we first derive the Euler-Lagrange equation governing facial deformation. The Euler-Lagrange equation represented as an ordinary differential equation (ODE) is embedded into a differentiable ODE solver. Muscle activation forces together with other physics parameters are firstly regressed, and then are utilized to simulate 3D deformation by solving the ODE. By leveraging facial biomechanics, we obtain physically plausible facial muscle activation forces. 2D image branch compensates 3D physics branch by employing additional appearance information from 2D images. Both estimated forces and appearance features are employed for AU detection. The proposed approach achieves competitive AU detection performance on two benchmark datasets. Furthermore, by leveraging biomechanics, our approach achieves outstanding performance with reduced training data.

1. Introduction

Action unit (AU) describes a local facial behavior, representing the movement of one facial muscle or a group of facial muscles [6]. For example, AU12 (lip corner puller) is corresponding to the muscle *zygomatic major*. AU15 (lip corner depressor) is corresponding to the muscle *depressor anguli oris*. In Figure 1, we visualize the muscles *zygomatic major* and *depressor anguli oris*. Due to the muscle activation, facial changes, in terms of both appearance and skin geometry, can be observed. Action unit detection task is to automatically predict if an AU is activated or not, given a 2D image. The majority of existing AU detection methods perform AU detection based on appearance information extracted from 2D images [4, 16, 26, 46]. These algorithms are mainly data-driven, whose performance highly depends on the quantity and quality of AU annotations. Unfortunately, AU annotations are hard to obtain and prone to errors. Besides, data-driven AU detection algorithms can't generalize well to unseen scenarios beyond training samples.



(a) Zygomaticus major (b) Depressor anguli oris

Figure 1. Visualization of *depressor anguli oris* (shown in (a)) and *orbicularis oris* (shown in (b)). Images are from https://en.wikipedia.org/wiki available under Public Domain.

To perform robust and generalizable AU detection under limited AU annotations, generic knowledge about the anatomical spatial relationships among facial muscles is considered, based on which AU relationships are derived [16, 17, 48, 49]. Facial biomechanics, which defines the dynamic of facial 3D deformation given muscle activation forces and is represented as second-order ODEs, is rarely considered. Facial biomechanics is important since it directly connects the muscle activation to skin deformation through principled physics laws, which is applicable to different subjects and independent of a specific dataset.

In this paper, we propose a biomechanics-guided AU detection approach, where facial muscle activation forces are modelled given 2D images and are employed for AU detection task. Our model consists of two branches: 3D physics branch and 2D image branch. In 3D physics branch, the Euler-Lagrange equation governing 3D deformation is firstly derived, which is represented as an ordinary differential equation (ODE). Muscle activation forces together with other physics parameters are regressed and utilized for physics-based reconstruction by solving the ODE. 2D im-

age branch compensates 3D branch by employing appearance information. Finally, the estimated muscle activation forces together with image features are employed for AU detection. Our contributions lie in three parts:

- We are the first to explicitly integrate facial biomechanics for AU detection task. Particularly, our physics branch explicitly models muscle activation forces. After training, physically plausible and anatomically meaningful forces are employed for AU detection.
- We are the first to introduce a generalized coordinate using facial blendshape basis and derive the Euler-Lagrange equation in the defined generalized coordinate. The Euler-Lagrange equation is then embedded into a differentiable ODE solver for physics-based reconstruction.
- We empirically demonstrate the effectiveness of our proposed approach on two benchmark datasets. Furthermore, our method remains robust under limited AU annotations and is cross-dataset generalizable.

2. Related Works

Facial Action Unit detection: Existing AU algorithms are mainly appearance-based approaches, where deep features are extracted from 2D images for AU prediction tasks [11, 16, 37]. Sophisticated network architectures are considered for extracting better deep features, such as ResNet50 [10] and Inception Network [38]. AU spatial relationships have been widely considered for appearance-based AU detection, most of which are learned from data [11, 20, 33, 36, 37]. Besides, temporal relationships among AUs within consecutive frames have been modelled via a long short-term memory model (LSTM) [2, 9]. Probabilistic models (e.g., dynamic Bayesian network and restricted Boltzmann machine) have also been considered for temporal AU relationship modelling [18, 42–44]. To simultaneously model spatial and temporal dependencies among AUs, a spatio-temporal graph convolutional network has been considered [33]. Nevertheless, these appearance-based AU detection methods suffer from subtle appearance changes. 3D geometric deformation is another important visual cue since it is evident that skin exhibits different facial expressions, activated by the underlying facial muscles. Some works explored the usage of 3D geometric deformation for expression recognition, whereby 3D shape changes are represented by deep features and are combined with appearance-based features [8,22,39]. The methods mentioned above are mainly data-driven and require sufficient AU annotations to perform well.

For robust and generalizable AU detection, generic knowledge is considered. The generic knowledge is about static spatial muscle relationships, based on which AU relationships are derived [5, 48]. Facial biomehcanics models dynamic 3D skin deformation given the muscle activation. To the best of our knowledge, facial biomechanics is not considered for AU detection yet. In this paper, we propose a biomechanics-guided AU detection approach to explicitly incorporate the facial biomechanics for AU detection.

Physics-based Facial Motion Modeling: Physics-based facial motion modelling remains an attractive topic in computer graphics field. Given the fact that facial soft tissues are structurally complex and exhibit highly non-linear constitutive behavior, modeling facial motion behavior is challenging. Usually, human face is represented via a volumatric or surface 3D mesh [1, 3, 34]. Finite volume method [1, 3, 41] and finite element method [34] are widely considered for spatial discretization. Newton's second law is the governing differential equation for facial motion [14]. Physics parameters involved in the differential equations reflect real properties of facial tissues and muscles, such as stiffness. Physics parameters have to be carefully specified by domain experts [23] or learned via neural networks from data [13]. To perform forward simulation over time, Euler method is widely used for solving the differential equations given muscle activation. Though anatomically explainable and generalizable, existing physics-based facial motion modeling techniques in computer graphics are very computationally expensive, prohibiting them from being employed for computer vision tasks, such as AU detection. In this work, we move one step forward in bridging realistic facial models to image-based facial recognition tasks, through the proposed biomechanics-guided AU detection approach.

Besides facial motion modeling, different works have been done on combining principled physics laws with existing deep learning techniques [12, 19, 27, 28]. These works, though employ physics laws, focus on synthetic physics systems and rarely consider real applications in computer vision. Furthermore, majority of existing physics-based deep models for computer vision tasks integrate physics laws via a regularization term to ensure physical consistency [40, 45]. In this paper, we propose a biomechanics-guided approach which explicitly encodes facial biomechanics for AU detection task. Particularly, in 3D physics branch, the proposed physics-based decoder simulates 3D skin deformation by solving its governing ordinary differential equations.

3. Proposed Method

We firstly introduce the facial biomechanics. We then introduce the proposed biomechanics-guided AU detection framework. In the end, we introduce the training objective of the proposed model.

3.1. Facial Biomechanics

To model facial biomechanics, we employ a 3D surface mesh for facial skin modeling. The muscles are then implicitly modeled via pre-defined muscle effective areas on the 3D surface mesh. The facial motion is described via the mesh motion. A 3D mesh contains vertices $i = 1, 2, ..., N$, with N being the total number of vertices. Deformation $\boldsymbol{u}_i$ denotes the position of i-th vertex relative to its initial position at rest. An observed skin deformation $\boldsymbol{u} = \{\boldsymbol{u}_i\}_{i=1}^{N}$ corresponds to a meaningful facial expression. The external forces $\boldsymbol{F}^{ext}$ causing this skin deformation are originate from facial muscles underneath the facial skin $\boldsymbol{F}^{mus}$. In the following, we firstly introduce the muscle activation force $\boldsymbol{F}^{mus}$ and how it deforms the facial skin through external forces $\boldsymbol{F}^{ext}$. We then introduce the Euler-Lagrange equation governing the motion of 3D facial skin given external forces $\boldsymbol{F}^{ext}$. Lastly, we introduce the forward dynamic solving the second-order ODE governing the dynamics.

Facial Muscle Activation Force: The facial muscles are innervated by the facial nerve and then deform the skin by contracting. Based on their contraction, the facial muscles devoted to facial expression generation can roughly be split into two types [7]: sheet and sphincter. A sheet muscle consists of muscle fibers arranged in parallel threads. During a contraction, muscle fibers shorten equally to pull the skin towards a fixed origin site of that muscle. For example, *frontalis* belongs to sheet muscle. In a sphincter muscle, muscle fibers form closed curves. During a contraction, sphincter muscle slides to generate different movements, e.g., closure or protrusion, and there is no fixed origin site in sphincter muscles. Both *orbicularis oculi* and *orbicularis oris* can be treated as sphincter muscles. In the end, a contracted muscle exhibits a muscle activation force $\boldsymbol{F}_m^{mus}$, where m denotes m-th facial muscle and $m = 1, 2, ..., M$ with M being the total number of muscles.

Vertices on a skin mesh receive non-zero muscle activation forces as external forces and then move. For each muscle, its effective area includes all the vertices on the 3D mesh receiving forces from that muscle and is defined according to facial muscle anatomy. One vertex can receive activation forces from different muscles. Particularly, we leverage facial blendshapes that semantically correspond to facial muscle activations to label the correspondence between muscles and mesh vertices (i.e., effective areas) and pre-define a distribution matrix $\mathbb{P} \in \mathbb{R}^{N \times M}$. N is the number of vertices and M is the number of muscles. $\mathbb{P}(i, m)$ indicates the probability of m-th muscle introducing external force to i-th vertex. We visualize four major facial muscles and their effective areas in Figure 2.

Given the distribution matrix, muscle forces $\boldsymbol{F}^{mus}$ are distributed to each vertex on the mesh to obtain the external force $\boldsymbol{F}^{ext}$ that each vertex receives due to muscle contraction. The total external force that i-th vertex receives from muscles then becomes $\boldsymbol{F}_i^{ext} = \sum_{m=1}^{M} \mathbb{P}(i, m)\boldsymbol{F}_m^{mus}$. In the end, the external force matrix $\boldsymbol{F}^{ext} \in \mathbb{R}^{N \times 3}$ becomes

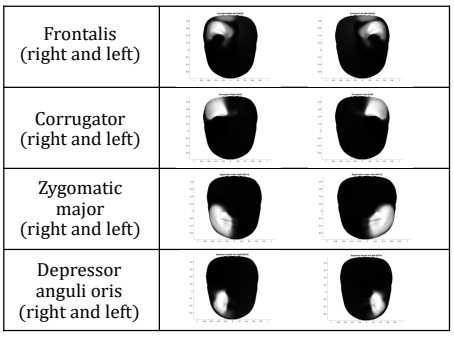


Figure 2. Four major muscles and their effective areas. The bright part indicates the mesh vertices on a mesh corresponding to the muscle. For the muscles that are causing bilateral symmetric forces, e.g., *fontalis*, we consider the effective areas on the right and left sides of the face separately.

$$\boldsymbol{F}^{ext} = \mathbb{P}\boldsymbol{F}^{mus} \tag{1}$$

Facial Skin Motion: The external forces cause the motion of a 3D mesh. For a 3D mesh in a motion, the deformation of each vertex is a function of time, i.e., $\boldsymbol{u}_i(t) = \{u_{i,x}(t), u_{i,y}(t), u_{i,z}(t)\}$ in the Cartesian coordinate. Deformation $\boldsymbol{u}_i(t)$ denotes the position of i-th vertex at time t relative to its initial position at rest. The velocity and acceleration in the Cartesian coordinate of each vertex are $\frac{d\boldsymbol{u}_i(t)}{dt} := \dot{\boldsymbol{u}}_i(t)$ and $\frac{d^2\boldsymbol{u}_i(t)}{dt^2} := \ddot{\boldsymbol{u}}_i(t)$, respectively. We treat a 3D mesh as a spring-mass system, and according to Newton's second law, the second-order ordinary differential equation (ODE) for the motion of i-th vertex is

$$m_i \frac{d^2\boldsymbol{u}_i(t)}{dt^2} := m_i\ddot{\boldsymbol{u}}_i(t) = \boldsymbol{f}_i \tag{2}$$

where m_i is the mass of i-th vertex. $\boldsymbol{f}_i = \{\boldsymbol{F}_i^{ext}, \boldsymbol{f}_i^{elas}\}$ denotes the forces that i-th vertex receives. $\boldsymbol{F}_i^{ext} = \{F_{i,x}^{ext}, F_{i,y}^{ext}, F_{i,z}^{ext}\}$ is the external force applied to i-th vertex which is caused by muscle activation. If there is no external force applied, $\boldsymbol{F}_i^{ext} = \boldsymbol{0}$. $\boldsymbol{f}_i^{elas}$ is the internal elastic force. For every pair of vertices connected through a spring, i.e., (i, i'), they receive the elastic forces of the same magnitude but opposite direction due to the deformation of the spring, i.e., $\boldsymbol{f}_i^{elas} = -\boldsymbol{f}_{i'}^{elas}$, and the elastic force is computed following Hooke's law [30].

To solve the second-order ODE (Eq. 2) in the Cartesian coordinate, we need to solve for $\boldsymbol{u}$ of dimension $N \times 3$. N, representing the total number of vertices of a facial mesh, is usually huge ($N \sim 30,000$). Directly solving for $\boldsymbol{u}$ is hence computational challenging. Instead, to reduce the dimension, we consider the dynamic in a generalized coordinate

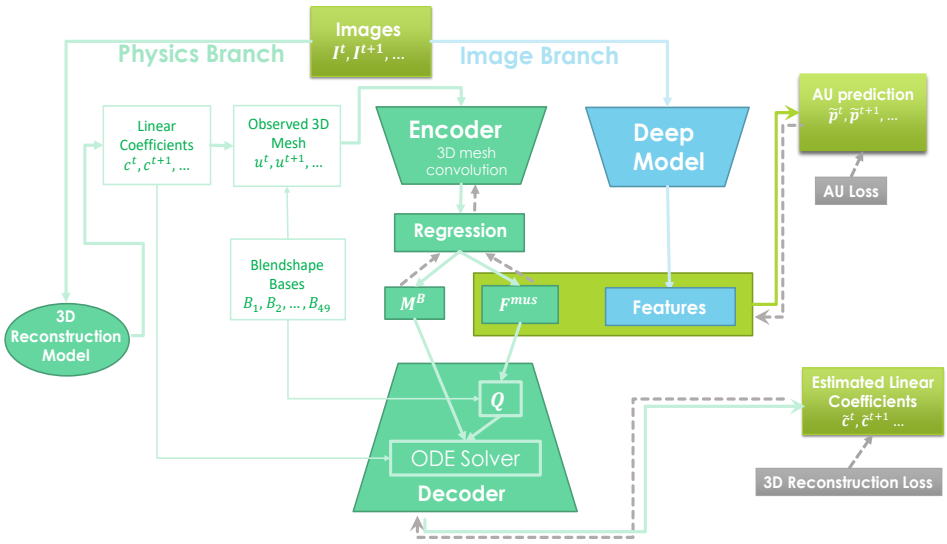


Figure 3. Overview of the proposed physics branch for biomechanics-guided AU detection

with lower dimension. Particularly, we leverage the typical construction of a 3D facial mesh and represent a mesh as a linear combination of 3D facial blendshapes, i.e.,

$$\boldsymbol{u} = c_1 B_1 + c_2 B_2 + \cdots + c_K B_K \tag{3}$$

where $\{B_j\}_{j=1}^K$ are known and K is the total number of basis defined by the reconstruction model. Both $\boldsymbol{u}$ and B_j are of $N \times 3$ dimension. $\boldsymbol{c} = \{c_j\}_{j=1}^K$ are the coefficients specific to each reconstructed mesh $\boldsymbol{u}$. As a 3D facial mesh can be completely specified by a facial blendshape basis $\{B_j\}_{j=1}^K$, the blendshape basis can be used as the generalized coordinate to efficiently capture the facial dynamics. We thus define the motion via coefficients $\boldsymbol{c}$.

To derive the motion law w.r.t. $\boldsymbol{c}$ in the specified generalized coordinate, we leverage the Euler-Lagrange equation. For $\boldsymbol{u}_i = \boldsymbol{u}_i(q_1, q_2, ..., q_k)$, $\boldsymbol{q} = \{q_1, q_2, ..., q_k\}$ defines a generalized coordinate. Euler-Lagrange equation defines the motion in the generalized coordinate as $M(\boldsymbol{q})\ddot{q} + C(q, \dot{q}) = \boldsymbol{Q}$, where $M(\boldsymbol{q})$ is the generalized mass and $\boldsymbol{Q}$ is the generalized force. $C(q, \dot{q}) = \dot{M}\dot{q} - \frac{1}{2}\dot{q}^T(\frac{\partial M}{\partial q})^T\dot{q}$. Defining the generalizaed coordinate $\boldsymbol{q}$ using the Blendshape basis, we derive the Euler-Lagrange equation in the specified generalized coordinate as

$$M^B\ddot{\boldsymbol{c}} + C(\boldsymbol{c}, \dot{\boldsymbol{c}}) = \boldsymbol{Q} \tag{4}$$

where $M^B = \frac{1}{2}\sum_i m_i^B$ is the generalized mass, and $m_i^B = m_i\hat{B}(i)$ where $\hat{B}(i) = B(i)^T B(i)$ and $B(i) = [B_1(i), B_2(i), ..., B_j(i), ..., B_K(i)] \in R^{3\times K}$. $\boldsymbol{Q} = \sum_i Q_i$ is the generalized force with $Q_{ij} = B_j^T(i)\boldsymbol{f}_i$. $C(\boldsymbol{c}, \dot{\boldsymbol{c}}) = \frac{dM^B}{dt}\dot{\boldsymbol{c}} - \frac{1}{2}\dot{\boldsymbol{c}}^T(\frac{\partial M^B}{\partial \boldsymbol{c}})^T\dot{\boldsymbol{c}}$. Since M^B is not a function of time t, nor a function of $\boldsymbol{c}$, $C(\boldsymbol{c}, \dot{\boldsymbol{c}}) = 0$. In the end, the Euler-Lagrange equation for a whole mesh in the specified generalized coordinate is

$$M^B\ddot{\boldsymbol{c}} = \boldsymbol{Q} \tag{5}$$

$\boldsymbol{Q} = \sum_i Q_i$ only contains external force caused by muscle contraction as $Q_{ij} = B_j^T(i)\boldsymbol{f}_i^{ext}$, because the summed elastic force for each pair of vertices becomes zero. Detailed derivations are in Appendix A and the verification of the derived dynamic law is in Appendix B.

Forward Dynamic with an ODE Solver: Given the derived dynamic law in the generalized coordinate defined by the blendshape basis (Eq. 5), we perform forward dynamic by solving the second-order ODE. We consider a standard Euler integration method for solving the ODE as

$$\begin{aligned} \boldsymbol{c}(t+1) &= \boldsymbol{c}(t) + \dot{\boldsymbol{c}}(t) \\ \dot{\boldsymbol{c}}(t+1) &= \dot{\boldsymbol{c}}(t) + \ddot{\boldsymbol{c}}(t) \end{aligned} \tag{6}$$

where $\ddot{\boldsymbol{c}}(t) = (M^B)^{-1}\boldsymbol{Q}(t)$ according to the dynamic law. With the estimated linear coefficients $\boldsymbol{c}(t+1)$, we can obtain the corresponding mesh $\boldsymbol{u}(t+1)$ following Eq. 3.

3.2. Biomechanics-based AU Detection

In this section, we introduce the architecture of our model which consists of two branches: 3D physics branch and 2D image branch. An overview of the proposed physics branch is shown in Figure 3. Both two branches take a 2D video as an input. 3D physics branch models muscle activation forces with geometric deformations leveraging biomechanics. 2D image branch employs standard image-based AU detection model and estimates 2D appearance information. Both 3D forces and 2D appearance information are used for facial AU detection, whereby 2D appearance information serves as a compensation of 3D forces.

3.2.1 3D Physics Branch

3D physics branch employs an encoder-decoder architecture, and biomechanics is embedded by incorporating a differentiable ODE solver into the decoder. 3D physics-based reconstruction is then performed by solving the ODE given external forces together with physics parameters, which are regressed given 3D deformations. Given an input image I, a 3D mesh $\boldsymbol{u}$ is obtained using the 3D reconstruction [15].

Encoder with 3D mesh convolution: The encoder E extracts 3D geometric features given 3D deformation $\boldsymbol{u}$ as

$$\boldsymbol{z}^{3d} = E(\boldsymbol{u}; \Phi) \tag{7}$$

The encoder is realized via a 3D mesh convolution neural network [29] mapping a deformed mesh $\boldsymbol{u}$ to the latent variables space. Φ denotes trainable parameters within the encoder that are unknown and to be learned during training.

Physics parameter estimation through regression: The extracted 3D geometric features $\boldsymbol{z}^{3d}$ are employed to estimate physically meaningful parameters through a regression neural network $\mathcal{F}_{reg}$ as,

$$\boldsymbol{z}^{phys} = \mathcal{F}_{reg}(\boldsymbol{z}^{3d}; \Theta) \tag{8}$$

with $\boldsymbol{z}^{phys} = \{M^B, \boldsymbol{F}^{mus}\}$. $\mathcal{F}_{reg}$ is realized through a multi-layer perception and Θ denotes trainable parameters.

Decoder with a differentiable ODE solver: Given the muscle activation forces, we first map the muscle activation forces to the generalized forces $\boldsymbol{Q} = \sum_i \boldsymbol{Q}_i$ with

$$\boldsymbol{Q}_i = B^T(i)\boldsymbol{f}_i^{ext} = B^T(i)(\sum_{m=1}^{M} \mathbb{P}(i,m)\boldsymbol{F}_m^{mus}) \tag{9}$$

and $B^T(i) = [B_1(i); B_2(i); ..., B_j(i); ..., B_K(i)] \in R^{K\times 3}$. Given $\{M^B, Q\}$, instead of following a typical data-driven decoder, we customize the decoder via a differentiable ODE solver, through which we integrate the dynamic laws into the decoder. Particularly, the decoder estimates coefficients $\tilde{c}$ by solving the ODE (Eq. 5) as

$$\tilde{c} = \text{ODESolver}[\frac{d^2\boldsymbol{c}(t)}{dt^2} = (M^B)^{-1}\boldsymbol{Q}] \tag{10}$$

We employ the standard Euler integration method as our ODE solver for predicting $\tilde{c}$ as we introduced previously (Eq. 6). The initial velocity is defined as $\frac{c(\Delta t)-c(0)}{\Delta t}$. There is no learnable parameters in the decoder. Through the integrated dynamic law in the decoder, we perform physics-based reconstruction, ensuring the estimated physics parameters are meaningful.

3.2.2 2D Image Branch

We employ ResNet50 [10] for extracting image features. Taking an image I as input, image feature $\boldsymbol{z}^{2d}$ is obtained

$$\boldsymbol{z}^{2d} = \mathcal{F}(I) \tag{11}$$

$\mathcal{F}(\cdot)$ is learned off the shelf and is fixed during training. Our proposed framework is not limited to the ResNet50 and can be straightforwardly applied to any image-based backbone.

3.2.3 AU Prediction

We employ physics parameters including the estimated muscle activation forces $\boldsymbol{F}^{mus}$ and generalized positions $\boldsymbol{c}$, together with image features for predicting AU activation $\boldsymbol{y} = \{y_r\}_{r=1}^R$. R is the total number of AUs to be predicted. $y_r = \{0, 1\}$ where $y_r = 1$ indicates the activation of r-th AU. The estimated forces and generalized positions have specific physics meaning and semantically corresponding to AU activation. Given physics-based parameters and image features, AU activation is predicted as

$$p(y_r = 1) = \sigma(\mathcal{H}([\boldsymbol{F}^{mus}||\boldsymbol{c}||\boldsymbol{z}^{2d}])) \tag{12}$$

where $\mathcal{H}$ is a shallow fully-connected neural network. σ is the sigmoid function and "$[\cdot||\cdot]$" denotes concatenation.

3.3. Training Objectives

We first introduce two loss terms: 3D reconstruction loss and AU detection loss. We then introduce two force regularization terms. Lastly, we summarize the total training loss.

3D Reconstruction Loss is employed for training the 3D physics branch. Specifically, given a sequence of observable 3D meshes $\{\boldsymbol{u}_t\}_{t=1}^T$ where T indicates the length of the time sequence. The reconstruction loss is defined as

$$L_{3d} = \frac{1}{T}\sum_{t=1}^{T} ||c(t) - \tilde{c}(t)||_2^2 \tag{13}$$

where $c(t)$ denotes the ground truth generalized position at time t and $\tilde{c}(t)$ denotes the estimated one.

AU Detection Loss: For a sequence with time length T, GT AU labels are $\{\boldsymbol{y}_t^{GT}\}_{t=1}^T$. For each $\boldsymbol{y}_t^{GT}$, we have $\boldsymbol{y}_t^{GT} = \{y_{t,r}^{GT}\}_{r=1}^R$ with r indexing r-th AU and R is the total number of AUs. The AU loss is defined based on AU prediction error through a cross-entropy, i.e.,

$$L_{au} = \frac{1}{TR}\sum_{t=1}^{T}\sum_{r=1}^{R} y_{t,r}^{GT}\log(p_{t,r}) + (1-y_{t,r}^{GT})\log(1-p_{t,r}) \tag{14}$$

where $p_{t,r} = p(y_{t,r} = 1)$ is the predicted occurrence probability of r-th AU at time t.

Force regularizations: To avoid noisy forces, we add L1 norm to the estimated forces as

$$L_{clean} = \frac{1}{M}\sum_{j=1}^{M} |F_j^{mus}|_1 \tag{15}$$

By adding L_{clearn}, only the forces that are sufficiently significant will be kept. Since forces can't be changed suddenly, the estimated forces from two adjacent time steps

Table 1. Comparison to Baseline Methods

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg
3D Geometry	35.0	40.7	49.7	69.0	73.0	78.7	78.4	59.8	10.6	30.8	25.7	13.3	47.0
3D Force	20.5	27.2	49.2	72.0	76.9	76.9	75.3	59.8	12.2	44.4	27.4	15.2	46.4
3D Position	52.9	47.9	56.9	76.3	80.2	80.9	82.3	58.5	18.7	40.1	32.2	20.4	53.9
3D Physics	**58.6**	48.4	58.1	75.3	78.1	80.5	82.7	58.6	29.4	46.6	39.1	30.9	57.3
2D Image	46.3	38.1	52.7	74.4	74.5	82.6	83.9	59.8	**48.1**	55.8	42.9	**45.2**	58.7
Physics + Image	57.4	**52.6**	**64.6**	**79.3**	**81.5**	**82.7**	**85.6**	**67.9**	47.3	**58.0**	**47.0**	44.9	**64.1**

should be similar. We thus introduce the second regularization term to ensure the smoothness of the estimated forces:

$$L_{smooth} = \frac{1}{T}\sum_{t=2}^{T} ||F^{mus}(t) - F^{mus}(t-1)||_2^2 \quad (16)$$

Their effectiveness are empirically shown in Appendix C.
Total Training Loss in the end is defined as

$$L = \lambda_{3d}L_{3d} + \lambda_{au}L_{au} + \lambda_c L_{clean} + \lambda_s L_{smooth} \quad (17)$$

where λ_{3d}, λ_{au}, and λ_c, and λ_s are hyper-parameters balancing the importance of different terms and are to be tuned.

4. Experiments

To evaluate the performance, we first compare the proposed model against different baseline models. We then compare against the state-of-the-art AU detection methods. We further study the effectiveness of the integrated biomechanics through a data efficiency evaluation and a cross-dataset evaluation. Lastly, we provide qualitative evaluation to understand the estimated muscle activation forces.

Datasets: We evaluate the proposed method on two benchmark datasets: BP4D [47] and DISFA [21]. Following previous works [16, 37], 12 AUs are used for evaluation on BP4D. For DISFA, we evaluate on 8 AUs and AU with intensity values greater than 2 are annotated as activated. We perform three-fold cross-subject evaluation, and report average performance. More details are in Appendix C.

Evaluation Metrics: F1 score is considered to evaluate the accuracy of AU detection. F1 score is calculated as $\frac{2p \cdot r}{p+r}$, where p is the precision and r is the recall.

Implementation Details: Given a 2D image, the corresponding 3D surface mesh is obtained from the AU specific 3D reconstruction method [15] because its blendshape basis is specific to facial AUs. To the best of our knowledge, this is the first work using 3D blendshapes for AU detection task. Other PCA-based 3D reconstruction models are not shown to be applicable to AU detection task and we thus don't consider. In image branch, the ResNet50 is pre-trained on ImageNet. More details are in Appendix C.

4.1. Comparison to Baseline Methods

We compare the proposed method to baselines using different types of information for AU detection including geometry, force, position in 3D. Physics branch employs both force and position for AU detection. Image branch employs 2D images for AU detection. We perform the evaluation on BP4D as shown in Table 1. In physics branch, combining force and position for AU prediction achieves average F1 score 57.3%, outperforming using force or position alone (46.4% and 53.9%, respectively). Force describes the deformation between two adjacent frames while position describes the current status in the generalized coordinate. In addition, leveraging sufficient physically meaningful parameters (i.e., force and position), the physics branch outperforms geometry based model by 10.3%.

Comparing physics branch and image branch, we can see that they achieve comparable performance on average, with F1 score being 57.3% and 58.7% respectively. Physics based AU detection significantly outperforms image based AU detection on some AUs that are hard to detected through appearance. For example, for AU1, physics based method achieves 58.6% while image based method only achieves 46.3%. These results further demonstrate that physics branch and image branch provides complementary information for AU detection. Finally, by combing physics branch and image branch, we achieve the best performance with average F1 score 64.1%.

4.2. Comparison to the State-of-the-art Methods

We compare the proposed biomechanics-guided AU detection method against the state-of-the-art methods including: MLCR [24], ARL [32], CMS [31], SRERL [16] and LP-Ne [25], , HMP-PS [37], UGN-B [35] and FAU [11]. We report the results on BP4D and DISFA datasets as shown in Table 2 and Table 3, respectively. On BP4D, we achieve competitive performance compared to the SOTA methods, outperforming most of them. Particularly for some AUs that are hard to be detection from images, our approach performs better. For example, for AU1, our approach achieves F1 score 57.4%, which is 5.7% higher than FAU. FAU employs transformers for image-based AU detection, which is a more sophisticated neural architecture

Table 2. The F1 score (in %) for the detection of 12 AUs on the BP4D dataset. The best results are indicated using bold.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg.
ARL [32]	45.8	39.8	55.1	75.7	77.2	82.3	86.6	58.8	47.6	62.1	47.4	55.4	61.1
CMS [31]	49.1	44.1	50.3	79.2	74.7	80.9	**88.3**	63.9	44.4	60.3	41.4	51.2	60.6
SRERL [16]	46.9	45.3	55.6	77.1	78.4	83.5	87.6	60.6	**52.2**	**63.9**	47.1	53.3	62.9
LP-Net [24]	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
HMP-PS [37]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	**49.9**	54.5	63.4
UGN-B [35]	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
FAU [11]	51.7	49.3	61.0	77.8	79.5	**82.9**	86.3	67.6	51.9	63.0	43.7	**56.3**	**64.2**
Ours	**57.4**	**52.6**	**64.6**	**79.3**	**81.5**	82.7	85.6	**67.9**	47.3	58.0	47.0	44.9	64.1

Table 3. The F1 score (in %) for the recognition of 8 AUs on the DISFA dataset. The best results are indicated using bold.

Methods	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Avg.
ARL [32]	43.9	42.1	63.6	41.8	40.0	**76.2**	**95.2**	**66.8**	58.7
CMS [31]	40.2	44.3	53.2	57.1	50.3	73.5	81.1	59.7	57.4
SRERL [16]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
LP-Net [24]	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
HMP-PS [37]	38.0	45.9	65.2	50.9	**50.8**	76.0	93.3	67.6	61.0
UGN-B [33]	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
FAU [11]	**46.1**	**48.6**	**72.8**	**56.7**	50.0	72.1	90.8	55.4	**61.5**
Ours	41.5	44.9	60.3	51.5	50.3	70.4	91.3	55.3	58.2

compared to ResNet50, the backbone of our image branch. HMP-PS leveraged hybrid messages capturing AU relationships for AU detection. However, the AU relationships are extracted in an abstract level and are not sufficient to capture the principled relationships. Our approach, instead, by leveraging biomechanics represented as the Euler-Lagrange equation, outperforms HMP-PS by 0.7%. Besides, our proposed framework is not limited to a specific image-based AU model in the image branch. Consistent performance improvement is expected with different image-based AU models (e.g., FAU). On DIAFA dataset, our approach achieves reasonable performance, and outperforms some of the SOTA methods, e.g. CMS, SRERL, and LP-Net. The AU-specific 3D reconstruction model [15] used for generating 3D meshes is not employed for DISFA yet. Hence, the performance of the physics branch of our method is limited due to the low quality of the 3D meshes.

4.3. Effectiveness of Incorporated Biomechanics

We further demonstrate that leveraging the biomechanics can help improve the data efficiency and generalization, with two additional evaluations as shown in the following.

Data Efficiency Evaluation: To show that, by leveraging the anatomically meaningful forces, we can perform AU detection with reduced data dependency, we evaluate the AU detection performance under different amount of AU labels. Particularly, we consider 50%, 20%, 5% amount of training data for AU detection. We perform AU detection by using AU detection loss plus reconstruction loss. In comparison, we report the AU detection with AU detection loss only. Results are shown in Table. 4 As shown, as we decrease the training data, leveraging AU detection loss only can't provide good AU detection performance anymore due to the lack of samples. On the other hand, by leveraging the 3D reconstruction, the extracted geometric features are ensured to be physically meaningful, and thus are able to provide reasonably well AU detection performance. For example, with only 50% training data, leveraging reconstruction loss, we

Table 4. Data Efficiency Evaluation

AU labels	AU	AU + Reconstruction
50%	39.0	**46.2**
20%	34.7	**44.3**
5%	27.9	**33.0**

obtain average F1 score 46.2%, which is 7.2% higher than the AU detection performance using AU loss only. These results further show that by leveraging the biomechanics, our AU detection approach is less data dependent.

Cross-dataset Evaluation: We demonstrate the generalization ability through a cross-dataset evaluation. Particularly, we train the physics branch on BP4D dataset and test it on DISFA dataset. We consider the performance on AUs that overlapped between BP4D and DISFA. In comparison, we report the baseline data-driven model, which is trained with AU loss only. Results are shown in Table 5. As shown,

Table 5. Cross-dataset Evaluation

Method	data-driven	physics-based
Avg.	32.1	**38.8**

the physics branch based on biomechanics generalizes better, with 6.7% improvement compared to the data-driven model. These results show that biomechanics helps estimate forces which apply to different subjects cross different datasets, leading to improved generalization.

4.4. Understanding the Estimated Forces

To verify that the estimated muscle activation forces are physically meaningful, we perform additional evaluation of the estimated forces. Directly performing quantitative evaluation of the estimated muscle activation forces is challenging due to the lack of ground truth forces in the benchmark datasets. Instead, we perform surrogate qualitative evaluations to better understand the estimated forces.

Correlation between Force and AU Activation: We first visualize the magnitude of the estimated forces and the corresponding AU activation, as shown in Figure 4. As shown,

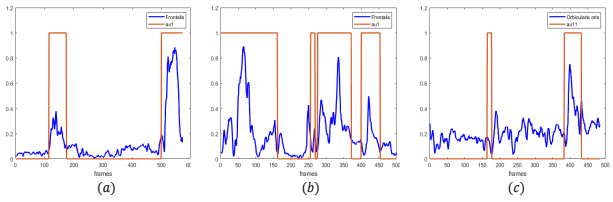


Figure 4. Correlation between estimated forces in magnitude (blue curves) and AU labels (red curves): (a) AU1 and *Fontalis* of subject F001 (T1); (b) AU1 and *Fontalis* of subject F008 (T5); (c) AU11 and *Orbicularis Oris* of subject F018 (T1).

the magnitude of the estimated forces is positively correlated to the activation of AUs indicated by the GT AU labels. Particularly, in Figure 4 (a), we observe two peaks of the magnitude of the estimated force of *Fontalis* are corresponding to the periods of AU1 being activated.

Force Visualization: We in addition visualize the estimated forces distributed on the 3D mesh, as shown in Figure 5. We

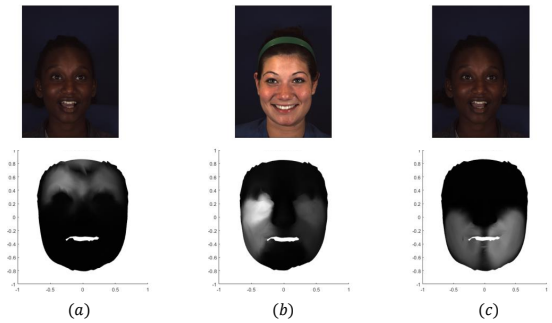


Figure 5. Force visualization: (a) for *Fontalis* (subject F001, 2469th frame of T1); (b) for *Orbucularis Oris* (subject F008, 413th frame of T8); (c) for *Zygomatic Major* (subject F001, 2471th frame of T1).

plot the magnitude of estimated muscle forces. On the top row, we display the original images, and on the bottom row, we display the corresponding estimated forces. The brighter the area, the larger the magnitude of the force. In Fig. 5 (a), AU1 and AU2 are activated. Correspondingly, we observe the significant active force for the corresponding muscle (i.e., *Fontalis*). Similarly, in Fig. 5 (b), AU6 is activated, and the estimated force of the muscle *Orbucularis Oris* is observed. In Fig. 5 (c), we observe the activated AU12 and the corresponding force for muscle *Zygomatic Major*.

5. Conclusion

In this paper, we proposed a biomechanics-guided AU detection approach, where facial muscle activation forces are modeled and are employed for AU detection. Specifically, forces are modelled under an encoder-decoder framework, where decoder performs physics-based reconstruction. By leveraging facial biomechanics, physically plausible and anatomically meaningful forces are estimated, given 2D videos. Both the estimated muscle activation forces and image features are employed for AU detection. We empirically demonstrate the effectiveness of the proposed method by comparing to state-of-the-art AU detection methods. Furthermore, we demonstrate that, by leveraging facial motion mechanism, our model performs robust AU detection under limited AU labels and is cross-dataset generalizable.

Acknowledgments

This work is supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

References

[1] Michael Bao, Matthew Cong, Stephane Grabli, and Ronald Fedkiw. High-quality face capture using anatomical muscles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[2] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017. 2

[3] Matthew Cong, Michael Bao, Jane L E, Kiran S Bhat, and Ronald Fedkiw. Fully automatic generation of anatomical face simulation models. In *Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 175–183, 2015. 2

[4] C. Corneanu, M. Madadi, and S. Escalera. Deep structure inference network for facial action unit recognition. In *ECCV*, 2019. 1

[5] Zijun Cui, Tengfei Song, Yuru Wang, and Qiang Ji. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[6] Paul Ekman, Wallace V. Friesen, and J. C. Hager. Facial action coding system. *A Human Face, Salt Lake City, UT*, 2002. 1

[7] Marco Fratarcangeli. *Computational Models for Animating 3D Virtual Faces*. PhD thesis, Linköping University Electronic Press, 2013. 3

[8] Boqing Gong, Yueming Wang, Jianzhuang Liu, and Xiaoou Tang. Automatic facial expression recognition on a single 3d face by exploring shape deformation. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 569–572, 2009. 2

[9] Jun He, Dongliang Li, Bin Yang, Siming Cao, Bo Sun, and Lejun Yu. Multi view facial action unit detection based on cnn and blstm-rnn. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 848–853. IEEE, 2017. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5

[11] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2, 6, 7

[12] Weiqi Ji, Weilun Qiu, Zhiyu Shi, Shaowu Pan, and Sili Deng. Stiff-pinn: Physics-informed neural network for stiff chemical kinetics. *The Journal of Physical Chemistry A*, 125(36):8098–8106, 2021. 2

[13] Petr Kadleček and Ladislav Kavan. Building accurate physics-based face models from data. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2(2):1–16, 2019. 2

[14] Jungmin Kim, Min Gyu Choi, and Young J Kim. Real-time muscle-based facial animation using shell elements and force decomposition. In *Symposium on Interactive 3D Graphics and Games*, pages 1–9, 2020. 2

[15] Chenyi Kuang, Zijun Cui, Jeffrey Kephart, and Qiang Ji. Au-aware 3d face reconstruction through personalized au-specific blendshape learning. In *European Conference on Computer Vision (ECCV)*, 2022. 5, 6, 7

[16] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI*, 2019. 1, 2, 6, 7

[17] Yongqiang Li, Jixu Chen, Yongping Zhao, and Qiang Ji. Data-free prior model for facial action unit recognition. *IEEE Transactions on affective computing*, 4(2):127–141, 2013. 1

[18] Yongqiang Li, Shangfei Wang, Yongping Zhao, and Qiang Ji. Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on image processing*, 2013. 2

[19] Minliang Liu, Liang Liang, and Wei Sun. A generic physics-informed neural network-based constitutive model for soft biological tissues. *Computer methods in applied mechanics and engineering*, 372:113402, 2020. 2

[20] Zhilei Liu, Jiahui Dong, Cuicui Zhang, Longbiao Wang, and Jianwu Dang. Relation modeling with graph convolutional networks for facial action unit detection. In *International Conference on Multimedia Modeling*, pages 489–501. Springer, 2020. 2

[21] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 6

[22] Iordanis Mpiperis, Sotiris Malassiotis, and Michael G Strintzis. Bilinear models for 3-d face and facial expression recognition. *IEEE Transactions on Information Forensics and Security*, 3(3):498–511, 2008. 2

[23] Andrew Nealen, Matthias Müller, Richard Keiser, Eddy Boxerman, and Mark Carlson. Physically based deformable models in computer graphics. In *Computer graphics forum*, volume 25, pages 809–836. Wiley Online Library, 2006. 2

[24] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. In *Advances in Neural Information Processing Systems*, pages 907–917, 2019. 6, 7

[25] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2019. 6

[26] X. Niu, H. Han, S. Yang, and S. Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *CVPR*, 2019. 1

[27] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. 2

[28] Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: A navier-stokes informed deep learning framework for assimilating flow visualization data. *arXiv preprint arXiv:1808.04327*, 2018. 2

[29] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 704–720, 2018. 5

[30] Jan Rychlewski. On hooke's law. *Journal of Applied Mathematics and Mechanics*, 48(3):303–314, 1984. 3

[31] Nishant Sankaran, Deen Dayal Mohan, Srirangaraj Setlur, Venugopal Govindaraju, and Dennis Fedorishin. Representation learning through cross-modality supervision. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 6, 7

[32] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 2019. 6, 7

[33] Zhiwen Shao, Lixin Zou, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Spatio-temporal relation and attention learning for facial action unit detection. *arXiv preprint arXiv:2001.01168*, 2020. 2, 7

[34] Eftychios Sifakis, Igor Neverov, and Ronald Fedkiw. Automatic determination of facial muscle activations from sparse motion capture marker data. In *ACM SIGGRAPH 2005 Papers*, pages 417–425. 2005. 2

[35] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence(AAAI)*. 6, 7

[36] Tengfei Song, Zijun Cui, Yuru Wang, Wenming Zheng, and Qiang Ji. Dynamic probabilistic graph convolution for facial action unit intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2021. 2

[37] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6267–6276, 2021. 2, 6, 7

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2

[39] Bilal Taha, Munawar Hayat, Stefano Berretti, Dimitrios Hatzinakos, and Naoufel Werghi. Learned 3d shape representations using fused geometrically augmented images: Application to facial expression and action unit detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2900–2916, 2020. 2

[40] Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. *arXiv preprint arXiv:2102.13156*, 2021. 2

[41] J. Teran, S. Blemker, V. Ng Thow Hing, and R. Fedkiw. Finite volume methods for the simulation of skeletal muscle. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association, 2003. 2

[42] Yan Tong and Qiang Ji. Learning bayesian networks with qualitative constraints. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[43] Z. Wang, Y. Li, S.Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *ICCV*, 2013. 2

[44] Ziheng Wang, Shangfei Wang, and Qiang Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3422–3429, 2013. 2

[45] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11532–11541, 2021. 2

[46] Huiyuan Yang, Taoyue Wang, and Lijun Yin. Adaptive multimodal fusion for facial action units recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2982–2990, 2020. 1

[47] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 6

[48] Yong Zhang, Weiming Dong, Baogang Hu, and Qiang Ji. Classifier learning with prior probabilities for facial action unit recognition. In *CVPR*, 2018. 1, 2

[49] Yong Zhang, Weiming Dong, Baogang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *CVPR*, 2018. 1