Towards Robust Facial Action Units Detection

Jing Yang, Yordan Hristov, Jie Shen, Yiming Lin, Maja Pantic

***Abstract*—Facial action unit (AU) detection plays an important role in performing facial behavioural analysis of raw video inputs. Overall, there are three key factors that contribute towards optimal performance of AU detectors: (1) being able to capture local AU-centered features, (2) exploiting the fact that some action units co-occur with others, and (3) utilizing appearance changes across frames. We briefly review current techniques addressing each factor and discuss the challenges they meet. Given that very few works consider how to effectively and efficiently merge them all into a single framework that can be trained in an end-to-end manner, we propose AUNet, a simple yet strong baseline for landmark-based AU detection. AUNet implements the above-mentioned key factors by (1) using the intermediate layers of a pre-trained face alignment model to act as our AU features space, optimised to satisfy a (2) correlation constraint, derived from the AU labels, as well as a (3) temporal constraint, derived from variations in the contents of consecutive frames in the input videos. The proposed model, with its three key components, remains simple in nature and aligns with the primary AU detection task. Experiments on several benchmarks show that it substantially improves the AU detector's accuracy and achieves a new state-of-the-art AU detection results on popular benchmarks: BP4D and DISFA. Code is available at https://github.com/jingyang2017/AU-Net.**

***Index Terms*—Facial Action Coding System, Action Unit, AU Features, AU Co-occurrences, AU Dynamics**

I. Introduction

FACIAL expressions are a natural and effective way to convey nonverbal information such as emotions, mental states, and sentiments in face-to-face communication. Given its potential applications in human-robot interaction, digital marketing and behaviour science, facial behaviour analysis has attracted a lot of attention from the affective computing and computer vision communities. This has made it necessary to be able to categorise and describe different facial behaviours in a standardised manner. One way to approach this problem is through the Facial Action Coding System (FACS)—a comprehensive anatomy-driven system, developed by Ekman and Friesen [21]. Based on muscle activity, it defines atomic facial movements, called Action Units (AUs). According to FACS, any facial expression can be decomposed into a mixture of AUs, including but not limited to the six universal emotions (*i.e.* anger, disgust, fear, happiness, sadness, and surprise) [20].

In recent years, especially popular have been AU detection methods that are centered around high-capacity neural models with learnable features [34], [56], [61], [62]. Since specific muscle movements are responsible for activating individual AUs, capturing facial appearance in the Region-of-Interest (RoI) area, relevant to these muscles, is crucial for successful AU detection—*e.g.* the mouth region of interest is relevant for smiling, eyebrow region of interest is relevant for frowning, *etc.* Facial landmarks, due to their spatial semantics, are one popular tool used to assign RoI for each AU and therefore avoid distractions from any other irrelevant regions. A typical pipeline is to first identify the RoI for each AU in the align-cropped version of a facial image, and then predict the presence or absence of each AU from features that are extracted from the corresponding RoI [82]. Aligning and cropping are two necessary pre-processing steps that help accelerate feature learning and reduce data variations caused by head rotation or different face sizes.

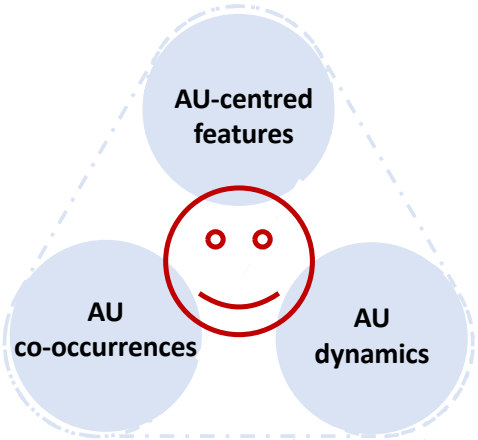


Fig. 1: We summarize three key factors, the intersection of which constitutes a good facial action unit detector—AU-centred features, AU co-occurrences, and AU dynamics.

Apart from their spatially-local nature, another aspect of AUs that is worth accounting for is their correlation caused by how different muscles are innervated by the facial nerves. Scherer *et al.*. [54] have discovered that there are around 7,000 AU combinations across different facial expressions in our daily life. For example, AU1 (inner brow raiser) and AU2 (outer brow raiser) usually appear simultaneously because they are controlled by the Frontalis muscle. Therefore, to improve the performance and robustness of the detection process, the presence of one AU can be exploited to predict another AU more reliably. Some works achieve this by either proposing novel network designs [29], [34] or by incorporating posterior refinement directly on the AU predictions [22], [69].

Lastly, similar to other computer vision tasks (*e.g.* action recognition [70]), performing AU detection on video data leads to better results. Psychological experiments conducted by Bassili *et al.* [5] have suggested that facial activities are more accurately recognized from consecutive frames than from a single static frame. Video fundamentally contains more

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

0000–0000/00$00.00 © 2021 IEEE

information than individual static frames, due to its temporal nature. Therefore it better represents phenomena like the stretching and contrasting of facial muscles involved in the manifestation of a given AU over time. Capturing the change from one state to another is critical to enriching the otherwise static descriptors. This has been realized by complementing the single frame's representation with dynamic movement [75], or capturing temporal dependencies among sequential frames [30], [35], [83].

Based on the above observations, we summarize three crucial factors for having an efficient and robust facial AU detector, as illustrated in Figure 1: **(1)** Influencing the learnt AU features to derive their meaning from their local neighbourhoods (in pixel space) rather than globally from the whole input image. **(2)** Exploiting the naturally-present co-occurrence of different AUs. **(3)** Utilizing temporal dynamics between consecutive frames. We have made the observation that most algorithms and methods in the AU literature usually implement one of these factors. To simultaneously incorporate all three of them in an end-to-end manner, we introduce our strong baseline method for ***Facial AU Detection with Face Alignment*** (AUNet). The paradigm of AUNet is depicted in Figure 4.

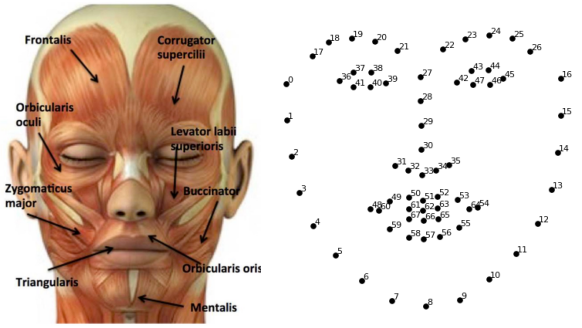
In this work, we make the following **contributions**: **(1)** We identify and look into three key factors from different AU detection methods in the literature that systematically lead to better performance. Consequently, we propose a relatively simple but strong baseline model that incorporates all three factors. **(2)** We propose Facial AU Detection with Face Alignment—an AU detection model that is built on the correspondence between alignment features and AU locations and works in a one-stage manner. We further extend the proposed Facial AU Detection with Face Alignment with a AU-VAE module, to exploit AU co-occurrences, and a AU-TDN module, to capture temporal AU dynamics. **(3)** We perform extensive experiments and ablation studies and show that our proposed approach sets up new state-of-the-art (SOTA) results on the following public AU benchmark datasets: BP4D [81], DISFA [47] and AffWild2 [33]. As a result, we argue that the proposed baseline provides more robust and discriminative features for AU detection. **(4)** Lastly, we include a commentary on the challenges that the Action Unit detection community faces in general, mainly due to the non-trivial nature of the task and the lack of big, balanced and noiseless data sets.

The rest of the paper is structured as follows. Section II introduces the definition of AUs, and also reviews the domain knowledge accumulated in the field of facial AU detection. Section III presents a brief review of recent developments grouped by factors contributing to a successful facial AU detector. Section IV introduces our proposed AUNet comprised of landmark-centered feature extractor, Variational Auto-Encoder based regulator, and customized Temporal Different Net, respectively. Section V is the experiment part including dataset description, performance metric, implementation details, ablate analysis of individual module, and comparison with state-of-the-art competitors. Section VI discusses the challenges met in current AU detection.

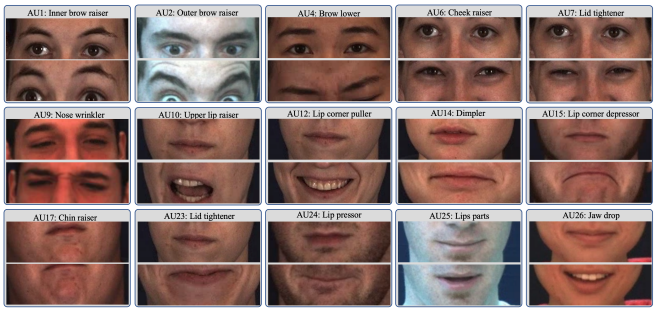
II. Facial Action Coding System

A. AU Label Definition

The presence of a particular AU is decided based on the appearance changes (*i.e.* geometrical shape, texture) caused by facial muscle movement on specific regions of a human face. AUs are objective descriptions for certain facial configurations rather than inferential labels like emotions. Following the definitions in FACS [21] and domain knowledge accumulated by the community [11], [34], [56], [57], we list fifteen AUs, corresponding names, locations, facial muscles, and landmarks in Table I. For context, we additionally provide visualizations for muscles and landmarks on a human face (see Figure 2a), and appearance changes caused by activating individual AU (see Figure 2b). For example, AU2, named outer brow raiser, is controlled by the Frontail muscle and its geometric transformation is reflected by displacements of landmarks (21,22) to 1/2 inner corner distance above outer brown. Often its activation could cause the skin in the center of the forehead to wrinkle horizontally.



(a) Facial muscles [35] (left) and 68 facial landmarks [59] (right) on a facial image.



(b) Facial appearance of fifteen facial AUs. We describe each AU with its neural state and active state.

Fig. 2: Visualizations for facial muscles, landmarks, and appearance changes caused by individual AUs in a 2D image. Refer to Table I for corresponding mapping.

B. AU Combination

AU combination refers to the multiple AUs present at the same time. Considering the anatomical characteristics of faces, there exist strong correlations between different AUs. One facial expression could contain one or several AUs. Co-occurring AUs can be additive or non-additive, Being additive means co-occurring AUs will not change the characteristic of a

TABLE I: Descriptions, corresponding locations, muscles and landmarks for fifteen AUs—a combination of two popular datasets: BP4D [81] and DISFA [47]. The location rule is applicable to an aligned-cropped face. Distance between two inner eye corners is used as the scaled distance. We supplement Table 1 in [56] with muscle names and landmarks [84].

AU index	AU name	Location	Muscle name	RoI with landmarks
AU1	Inner brow raiser	1/2 scale above inner brow	Frontalis	21,22
AU2	Outer brow raiser	1/3 scale above outer brow	Frontalis	17,26
AU4	Brow lowerer	1/3 scale below brow center	Corrugator supercilii	midpoint of 21,22
AU6	Cheek raiser	1 scale below eye bottom	Orbicularis oculi	midpoint of 2,14, midpoint of 41,46
AU7	Lid tightener	Eye	Orbicularis oculi	38,43
AU9	Nose wrinkler	1/2 scale above nose bottom	Levator labii superioris	1/2 scale below 39,42
AU10	Upper lip raiser	Upper lip center	Levator labii superioris	50,52
AU12	Lip corner puller	Lip corner	Zygomaticus major	48,54
AU14	Dimpler	Lip corner	Buccinator	48,54
AU15	Lip corner depressor	Lip corner	Triangularis	48,54
AU17	Chin raiser	1/2 scale below lip	Mentalis	1/2 scale below 56,58
AU23	Lip tightener	Lip center	Orbicularis oris	51,57
AU24	Lip presser	Lip center	Orbicularis oris	51,57
AU25	Lips part	Lip center	Orbicularis oculi	51,57
AU26	Jaw drop	1/2 scale below lip.	Mentalis	1/2 scale below 56,58

single AU—each muscle movement is responsible for one AU, and vice versa. Table II lists the co-occurring AUs associated with the prototypic expressions of emotion [20]. For example, happiness can be generated by the simultaneous activations of AU6 and AU12, as we generally raise cheek and pull lip corners when we smile. Statistically, [54] has observed around 7,000 AU combinations across facial expressions in our daily life, making the AU detection task more challenging.

TABLE II: Expression-dependent AU co-occurrences adapted from EMFACS [25]. AUs are from Table I.

Expression	AUs
Happiness	AU6+AU12, AU7+AU12
Anger	AU4+AU7, AU17+AU24,AU23
Fear	AU1+AU2+AU4
Sadness	AU1, AU1+AU4, AU6+AU15
Surprise	AU1+AU2+AU26
Disgust	AU9,AU10

C. AU evolvement

The activation of AU will cause the contraction or relaxation of one or more facial muscles [21]. Facial muscular activities would produce changes in the direction and magnitude of the skin surface motion. In terms of intensity, AU evolution is indeed divided into four segments [21], [68]: neutral, onset, apex, and offset. In the neutral phase, there are no signs of activation of facial muscle movement; In the onset phase, the muscles are stretching and facial appearance changes as the intensity of AU grows stronger; In the apex phase, the intensity of AU reaches the highest level and there are no more changes in facial appearance. In the offset phase, the muscles are relaxing and the facial appearance returns to neutral.

III. Automated Action Unit Detection

Facial AU detection has been studied for decades in the field of artificial intelligence and computer vision, and great progress has been achieved with the rise of convolutional neural networks (CNNs) [29], [35], [56], [61], [69]. AU detection is typically formulated as a multi-label classification problem including three stages [46]: **(1)** data pre-processing, **(2)** feature extraction, and **(3)** AU prediction (see Figure 3).

Data pre-processing usually contains two main steps: face detection and face alignment, which are crucial for learning AU features and predicting the AU existence. Through face detection, all located faces are used to crop the corresponding face regions from the entire input image. Face alignment is then performed by predicting facial landmarks and using them to align the detected faces to a common pre-defined face size and in-plane rotation. The second step is important for establishing spatial correspondence among different face images so that different feature dimensions consistently encode semantics for specific facial parts. Since our work focuses on the remaining two steps, we recommend the recent surveys about face detection and alignment [15], [48] for extensive reading. After the input faces have been aligned and cropped they are fed into a CNN-based network for the feature learning stage. The quality of the learned features is directly determined by the properties of the neural AU detector—*e.g.* whether its network design accounts for the local, subtle appearance changes associated with specific AUs' activations. The final stage is to predict AU occurrence based on the learned features.

In the following parts of this section, we identify three main design factors that are directly related to the performance of the resultant AU detector. These factors are usually incorporated in the second and third stages of the AU detection process as per Figure 3 and are respectively: **(1)** Biasing the feature extraction stage towards learning spatially-local representations. **(2)** Modeling AU co-occurrences. **(3)** Utilising temporal information.

A. Learning Spatially-local Representations

Similar to conventional computer vision approaches, some methods adopt handcrafted features extracted from the input facial images (*e.g.* Gabor wavelets [4], Local Binary Patterns

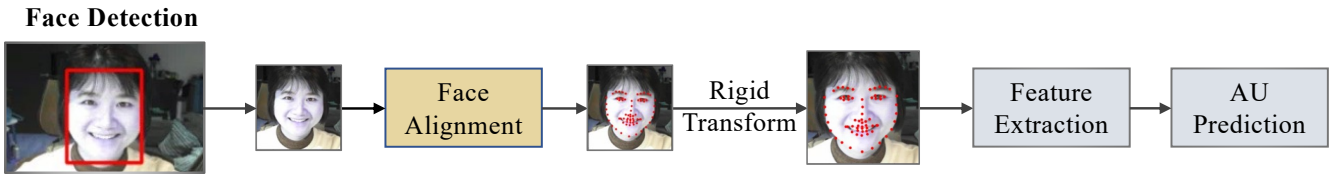


Fig. 3: The pipeline of a generic automatic action unit detection system.

(LBP) [82], Histogram of Oriented Gradients (HOG) [2]) to describe an AU. However, constrained by the representation capabilities of such features (not tuned for a specific task), these methods struggle to tackle challenging situations caused by variations in pose, illumination and occlusion *etc.*. Recently, those difficulties have been alleviated by representations learned through deep neural architectures in a data-driven manner.

Moreover, based on the AU descriptions in Table I, efficient AU feature learning should focus more on the local regions responsible for a particular AU and avoid any noise coming from other unrelated regions. However, due to the high dimensionality of the input image data, automatically locating these regions is difficult. Broadly speaking, there are two ways to solve this problem in the literature—manually partition the input faces and specify which face regions are responsible for predicting which AUs (explicit attention) or elaborate the network design to let it learn where to focus in the input facial images(implicit attention).

To define attention for AUs, a number of existing methods have made attempts. As a very early attempt, Zhao *et al.* [87] disentangle a global image into multiple small blocks, and then independently learn appearance changes within sub-areas with a newly introduced regional network. This work achieved much better performance than alternative approaches trained on the entire images. Motivated by such success, Li *et al.* [36] propose the EAC-Net that builds upon VGG19 [60]. Due to the way it combines enhancing layers and cropping layers this work can be thought of as effectively combining both explicit and implicit attention. Specifically, the enhancing layer is designed to add landmark attention during feature learning on the entire input image because the regions around certain facial landmarks are recognized as AU-related areas. The cropping layer is operated on sparsely distributed regions which are decided by the relationships between landmarks and AUs. Since then, this cropping rule that assigns particular RoI for each AU is widely adopted in the following works [29], [34], [35], [56], [57] In general, patch-wise feature extraction guided by explicit attention between AUs and sub-regions directly helps to avoid interruptions from unrelated regions. However, it has several drawbacks: firstly, it lacks a holistic view of an entire face which is important for detecting the co-occurring AUs; secondly, it requires perceptual consistency of RoI which is sensitive to the positions of facial landmarks.

Instead of leveraging a pre-trained face alignment to determine RoI for each AU, Wu and Ji [73] propose a constrained joint cascade regression framework for simultaneously conducting AU and landmark detection. The aim of this framework is to automatically encode the global dependencies in the feature space between facial actions and landmarks. Following this idea, Shao *et al.* [56], [57] present a multi-task solution under a CNN-based network to exploit the strong correlations between AUs and landmarks in their high-level representations. They further develop an adaptive attention learning module to refine the coarse attention map which is initially specified by facial landmarks [36]. Li *et al.* [37] observe that the ambiguous assignment of RoI for AUs could result in sub-optimal attention for AUs. For example, AU12, AU14, and AU15 have obvious regional overlap around lips. To remedy this, they propose a self-diversified multi-channel attention network to specify per AU's affecting feature. More recently, Yang *et al.* [77] propose a cross-modality attention module that combines the semantic embedding from AU description and visual features from a face image, to learn more useful and discriminative features from more meaningful facial areas. This is the first work to exploit AU semantic description (*i.e.* facial area/position, action, motion direction/intensity, and relation of AUs) to guide the network where to focus. Compared to explicit attention, the implicit solutions are more flexible without a manual assignment step, but cannot accurately remove the unrelated areas, which may produce noisy AU representations.

Overall, using the properties of individual action units—*e.g.* their spatially-local nature—as an inductive bias in the training of neural feature extractors is worthwhile. However, different AUs are not completely independent of each other.

B. AU Co-occurrences Modeling

As noted in Section II, for some facial expressions multiple AUs can be simultaneously active due to the underlying face anatomy. A number of works in the AU literature explicitly account for this phenomenon and as a result consistently achieve better detection performance when compared to methods that treat AUs as independent.

There are different ways to model AU co-occurrences at the feature extraction and AU prediction stages. As Restricted Boltzmann Machines are effective in modeling higher-order dependencies among random variables, Wang *et al.* [72] use them to model label relations. Zhao *et al.* [86] introduce a relational regularizer to force the predictions to satisfy the positive and negative relations resided in AU labels. As an alternative, graph Laplacian matrix is used to define constraints of AU co-occurrences of the output labels in [22]. More recently, Corneanu *et al.* [13] develop a structure reference network, which is composed of stacked fully-connected layers, to infer structure among AUs through iterative message passing on separate predictions from isolated patches. The relationship inference part works as a post-processing step and has proven to be beneficial in improving detection accuracy. However, it

is isolated from the feature extraction and is hence limited to the feature representation.

Alternatively, other works concentrate on modeling semantic correlations among AUs at the feature extraction level. Zhao *et al.* [86] propose a joint patch and multi-label learning framework with consideration of positive and negative AU relations. Wang *et al.* [71] connect features and AU labels with the Restricted Boltzmann Machine to capture both dependencies among features and the salient information for the input features. However, the performance of both is constrained by using hand-crafted features, which can be not discriminative enough for capturing the facial morphology. To automatically explore the underlying relationships between individual local facial regions, Niu *et al.* [51] introduce a local relationship learning module based on Long Short-Term Memory (LSTM) and append it after convolutional layers. To refine spatial attention for learning AU features, Shao *et al.* [58] design a fully-connected Conditional Random Fields (CRFs) to adaptively capture AU relations at the pixel level.

Recently, graph neural networks have also been utilized for modeling different AUs and their correlations. Li *et al.* [34] appends Gated Graph Neural Network built on a pre-defined knowledge graph after learned AU features. A similar idea could be found in [84] with a more advanced structure—HRNet [63]—as a feature extractor. Instead of borrowing the entire backbone from the image classification task, Yang and Yin [76] split the backbone into shared and AU-specific parts, and then use an AU relationship graph to regularize the latter. Niu *et al.* [50] embed prior knowledge about the relationship between different AUs through a graph convolutional network. This further allows them to use additional useful information from big unlabeled data sets. However, AU relationships vary in different expressions and individuals, which can not be well encoded by a single constant structure. Song *et al.* [62] propose a performance-driven structure by considering dynamic relations among AUs. More recently, Jacob and Stenger [29] introduce self-attention, based on a transformer module [18], into their network design because of its particular efficacy in capturing connections between distant patches. Since AU is encoded in a high dimensional feature vector with high flexibility, supervision alone between AU prediction and AU label meets the challenge in accurately reasoning about AU relationships.

Notably, both the local and co-occurring nature of AUs can be modelled in individual frames. However, facial expressions ultimately have progression through time which is also worth paying attention to.

C. Motion and Temporal Modeling

For a more coherent and reliable AU prediction in consecutive frames of AU evolvement [10], [16], [35], [55], [67], [83], a number of research groups are looking to capture facial appearance changes, evidenced by stretching and contrasting of facial muscles, by modeling the spatial and temporal variations. Guided by the spatially-local mechanism in Section III-A, the learned AU representation is capable of precisely revealing the subtle motion of muscle activity.

Earlier attempts leverage graphical models to achieve this. For example, Tong *et al.* [67] use Dynamic Bayesian Networks (DBNs) to consider two types of conditional dependency at different time steps. The first type is between two states of a single AU while the second type is between two states of different AUs. Change *et al.* [10] employ Hidden Conditional Random Fields (HCRFs) to establish connections between expression predictions and AUs in an image sequence. As an alternative, Hidden Markov Models (HMMs) are utilized in [55] with a new effective non-parametric output probability estimation method.

Recently, as a state-of-the-art temporal fusion network, Long Short Term Memory (LSTM) [26] has been applied in human action recognition [16], as well as facial AU detection. For example, Jaiswal *et al.* [30] input the dynamic appearance features learned by the time-windowed convolutional layers to a Bi-directional LSTM to model the temporal information. Li *et al.* [35] attach LSTM layers on sequential features from individual RoI nets to predict all AUs simultaneously. Zhang *et al.* [83] use LSTM to capture temporal movement with learnable task-related context to capture the temporal label attention.

Another direction is to complement the single frame's representation with its corresponding motion information. For example, [75] compute a motion field for a static frame by calculating the dense optical flow through unsupervised learning. However, flow estimation is easily disturbed by the variation of lighting and is insensitive to subtle movements of facial components.

Although models with temporal information have achieved a higher detection rate, it does suffer from a few shortcomings. For example, the sequential model requires a fixed number of frames as input, which ignore the different transition durations for different AUs and may result in the loss of temporal scale information.

D. Others

Although the AU labeling rule proposed in FACS [21] is generic and universal, the facial appearance changes caused by a facial action vary among different subjects because human faces are person-specific. From the view of feature engineering, to achieve a robust AU detector whose performance is similar on different persons, person-specific features should be eliminated from the final AU representation. The prevailing view of learning an anti person-specific representation is to first pre-define the person-specific features, then regulate the AU features orthogonal to them. The final features are supposed to be generalizable to different identities. For example, Niu *et al.* [51] treat facial landmarks as a kind of person-specific features and then introduce a person-specific shape regularization module which enforces the learned AU features orthogonal to the normalized landmark vectors produced by a facial landmark detector [3]. Zhang *et al.* [85] reckon identity features are person-specific and thus optimize the network in an adversarial manner to learn features effective for AU detection but invariant to the subject. The network is designed under a multi-task framework composed of a face recognition

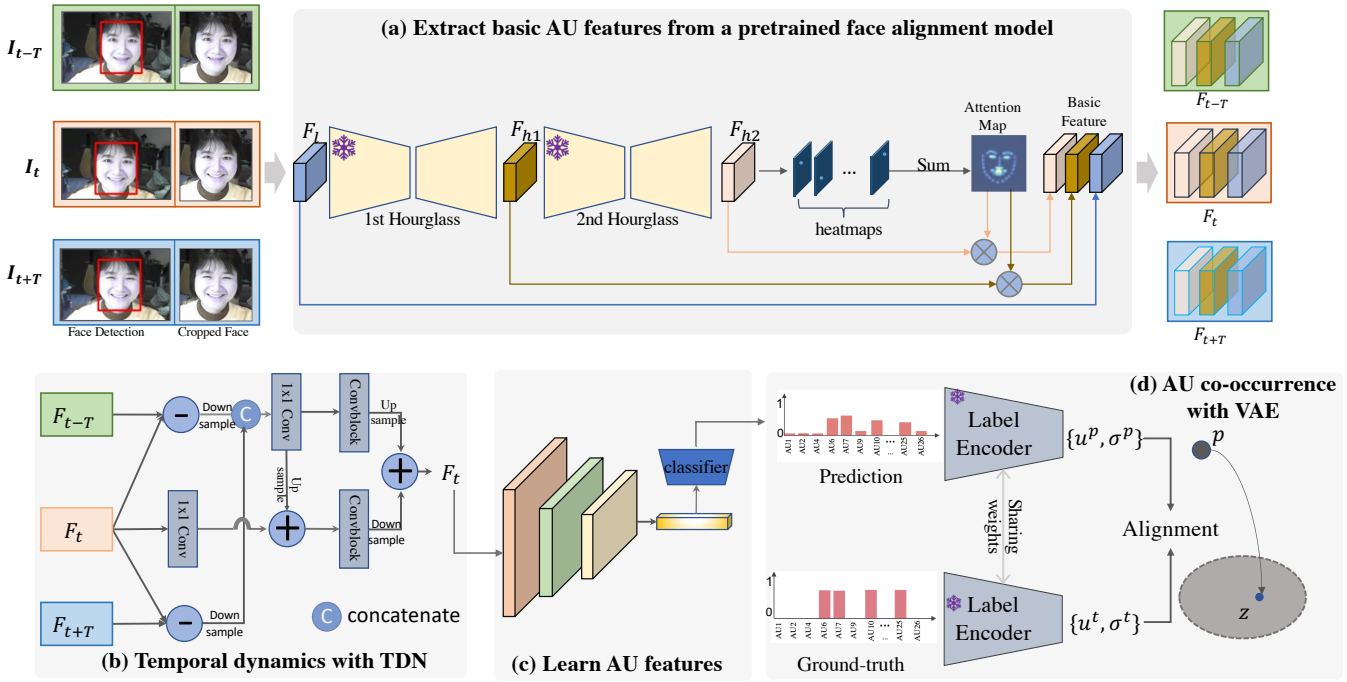


Fig. 4: Schematic overview of proposed Facial AU Detection with Face Alignment(AUNet) method. **(a)** For a face image I_t and its former I_{t-T} and later frames I_{t+T}, we extract per image's representations from a pre-trained frozen face alignment model (FAN), which is built upon a two-hourglass architecture. We obtain basic AU features by concatenating three feature sources from FAN: feature tensor before the 1st hourglass (F_l), feature tensor from the first hourglass (F_{h1}), and feature tensor from the second hourglass (F_{h2}). The latter two are further multiplied with aggregated heat-maps (attention maps) for capturing landmark attention. **(b)** Based on the three frame-based feature tensors ($\boldsymbol{F}_{t-T}$, $\boldsymbol{F}_t$, $\boldsymbol{F}_{t-T}$), we capture their temporal dynamics with a temporal different net module (AU-TDN). TDN works by complementing the static representation $\boldsymbol{F}_t$ with local motion information. '-' is a subtraction operation applied on pairs (*e.g.* $(\boldsymbol{F}_{t-T}, \boldsymbol{F}_t)$, $(\boldsymbol{F}_{t+T}, \boldsymbol{F}_t)$). We put the TDN right after basic features in order to preserve the spatial context which is crucial for AU detection. **(c)** We learn the compact and discriminative AU features via a sequence of convolutional blocks and pass features into the AU classifier to get the AU prediction. **(d)** We use a variational auto-encoder (VAE) module to regularize AU co-occurrences in the training stage. Considering the pretrained label encoder in VAE capturing such characteristic in its latent space, we realize the regularization by aligning the distributions between AU prediction and AU ground-truth. This module is discarded in the inference stage.

task and an AU detection task. A similar idea can be found in *et al.* [64]. In contrast to learning recognition model on AU datasets, Tang *et al.* [64] pre-train a face recognition model on the recognition datasets and utilize it to provide identity features for faces from AU datasets. They have a contrast finding to [51] and claim that facial shape is not person-specific information.

Inspired by the increased accuracy for object classification achieved by self-supervised learning, another line of work focuses on how to design a pretext task to learn a general facial representation for the AU detection task. By doing so, a large amount of unlabeled data can be explored to learn a robust local representation. For example, Chang and Wang [11] divide the facial area into eight parts based on the AU-related appearance changes, then enhance regional representation learning in a knowledge-driven self-supervised framework. Bulat *et al.* [6] first learn a universal face representation on enforcing consistency between the cluster assignments produced for different augmentations of the same face image [9], then fine-tune on different facial analysis tasks including facial AU detection. More recently, Ma *et al.* [44] instead use masked autoencoding as a pre-training paradigm [27].

IV. A Strong Baseline

By combining and utilising the properties of the locality of AU features, AU co-occurrences, and AU evolvement, we set up a strong facial AU detection baseline—AUNet—which obtains competitive performance while remaining parsimonious as an architecture. In this section, we elaborate on the network design, loss functions, and implementation/training details of AUNet.

A. Landmark-centered Features

Facial landmarks represent the semantically salient regions of a human face. Given that AUs are present in local regions around facial landmarks, previous approaches [34], [87] use landmarks to pre-define RoIs for AUs. However, RoI-based methods usually incur more time due to the extra pre-processing steps, are sensitive to errors in the landmark predictions, and might be sub-optimal since not all AUs can be related to a single patch in the input image. To avoid these issues, we rely on intermediate features from a pre-trained facial landmark model as a foundation for our AU representations [66].

Concretely, we leverage the stacked-hourglass-based FAN [7], [49] model which has previously been explored as a feature extractor for face recognition [78] and face emotion recognition [66]. FAN is trained for landmarks localization with heat-maps (Gaussian circle peaking at key points) as supervision on a large corpus of facial data, covering a variety of head poses. Unlike models trained on ImageNet for a classifications task which are generic (*i.e.* not specific to faces) and coarse, FAN features **(1)** can capture fine-grained aspects of input faces — a direct consequence of the face alignment task; **(2)** are robust to appearance variations from pose and illumination, as the model is pre-trained on a large variety of facial poses; **(3)** closely align with AU detection — having a good localization of the AU region correlates with higher AU accuracy.

As illustrated in Figure 4, we derive our AU features by combining intermediate features (F_l, F_{h1}, F_{h2}) and heat-maps (F_{hm}) from the pre-trained FAN. Specifically, we first aggregate heat-maps to a single plane (attention map), then multiply it with the penultimate layer's outputs of two hourglasses (F_{h1}, F_{h2}), and finally concatenate these features with the low-level features (F_l) to obtain F as basic features.

To learn compact and discriminative AU features, we feed $\boldsymbol{F}_t$ to a sequence of convolutional layers, followed by max-pooling (see Figure 4(c)). Being consistent with FAN [7], we utilize its multi-scale residual block ($\sim 0.4M$ parameters) as convblock (see Figure 5) in each convolution layer, which has proven to be superior to the vanilla basic block in capturing multi-scale feature representations. After that, we vectorize the representation by average pooling and transfer it to the AU classifier comprised of a hidden layer to reduce dimension, a ReLU activation, and a linear layer. Finally, we take the output of the AU classifier as AU predictions.

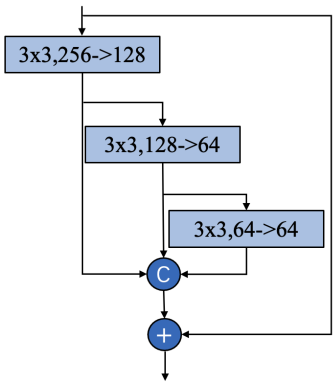


Fig. 5: Multi-scale residual block in FAN [7].

Following previous works [13], [34], [36], [56], we formulate the AU detection task as an imbalanced multi-label classification problem and optimize the learnable weights with a weighted binary cross-entropy loss:

$$\mathcal{L}_{cls} = -\sum_{i}^{n_{au}} w_i[y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)], \quad (1)$$

where $\hat{y}_i \in \hat{\mathbf{y}}$ denotes the prediction of i-th AU; $y_i \in \mathbf{y}$ is the ground-truth label of i-th AU (Here, "1" denotes presence and "0" denotes absence); n_{au} is number of AUs; w_i is used to weigh i-th AU and is calculated on its frequency of occurrence (r_i): $w_i = (1/r_i)/\Sigma_{k=1}^{n_{au}}(1/r_k)$.

B. Modeling AU Co-occurrence

To account for the co-occurring AUs, we leverage a Variational Auto-Encoder (VAE) [32]. By design, a VAE reconstructs an input signal via a low-dimensional latent representation which is optimised to contain useful information about the input. It has previously been shown to work well for learning shape priors in the context of a segmentation task after being pre-trained on segmentation masks [42], [52].

Here, we use the VAE architecture as a label regularizer and let it penalize deviation of the AU predictions from the ground-truth labels but in the feature space of a pre-trained VAE. The raionale is that the VAE feature space is better at capturing label co-occurences. To this end, the VAE is *first* optimized to auto-encode the ground-truth AU labels (reconstruction loss $\mathcal{L}_{bce}$) when constraining latent code to follow a pre-defined Normal prior distribution (KL divergence loss $\mathcal{L}_{kl}$). We formulate the overall objective as

$$\begin{aligned} \mathcal{L}_{vae} &= \mathcal{L}_{bce}(\mathbf{y}, \mathrm{D}(\mathrm{E}(\widetilde{\mathbf{y}}))) + \lambda \mathcal{L}_{kl}, \quad \text{where} \\ \mathcal{L}_{bce} &= -\sum_{i}^{n_{au}}[y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)]; \\ \mathcal{L}_{kl} &= KL(\mathcal{N}(\mu, \Sigma)||\mathcal{N}(\mathbf{0}, \mathbf{I})), \end{aligned} \quad (2)$$

where $y_i \in \mathbf{y}$, $\hat{y}_i \in \mathrm{D}(\mathrm{E}(\widetilde{\mathbf{y}}))$, μ, Σ are the output of encoder E, and $\widetilde{\mathbf{y}}$, E, D indicate the corrupted label, encoder, and decoder of VAE, respectively. λ is weight for balancing the two losses. To improve robustness, we add random noise from the uniform distribution (-0.2,0.2) on $\mathbf{y}$ to get $\widetilde{\mathbf{y}}$.

Once the VAE is trained we freeze its encoder E and *then* use it to steer the classification output of the AU prediction network. The latter is achieved by minimizing the divergence between the AU prediction and the ground-truth AU labels in the latent space of E:

$$\mathcal{L}_{cons} = ||\mu_p - \mu_g)||^2 + ||\Sigma_p - \Sigma_g||^2 \quad (3)$$

where $\{\mu_p, \Sigma_p\}, \{\mu_g, \Sigma_g\}$ denote the outputs from E when fed with AU predictions and AU labels, respectively.

E complements the AU co-occurrences by minimizing the consistency loss $\mathcal{L}_{cons}$ that measures how well the AU predictions correspond with the ground-truth AU labels in the latent space of the E. We incorporate E into AU-FAN and obtain **AU-VAE** by optimizing following objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \gamma \mathcal{L}_{cons}, \quad (4)$$

where γ is a hyper-parameter that weighs the contribution of the consistency loss, relative to the classification loss. Setting γ to 0 induces AU-FAN in Section IV-A. γ is to guide the optimization of the network satisfying the prior regularization over the AU labels. Since VAE module is used to consider the co-occurring AUs, each training set owns its individual VAE module.

C. Temporal Fusion

Temporal Different Net (TDN) [70] is a simple and principled temporal modeling framework [70] which is designed to capture multi-scale temporal information for efficient action recognition. In order to leverage the motion of facial textures in a sequence of static frames, we instantiate a TDN with AU-FAN to capture the change of facial muscles through time due to its high flexibility.

Given frame I_t, its former frame I_{t-T} and later frame I_{t+T} at interval T, we create a tuple of face images for the same person $I = \{I_{t-T}, I_t, I_{t+T}\}$. Appearance changes between two facial images of the same person in a video are always caused by two main factors: AU motion and pose motion [38]. The former motion is subtle and local. Thus, to better capture AU motion and alleviate the disturbance caused by changes in head pose, we align images in I based on five landmarks (*i.e.* left eye, right eye, nose, left mouth corner, and right mouth corner) of I_t and augment them with same transformations before feeding them into FAN for extracting basic AU features. Meanwhile, we put the TDN right after the basic feature extraction module in order to preserve the spatial context which is important for AU detection. As with TDN [70], we take its low-resolution architecture design as we share the similar sparse signal observation that AU evolvement causes high response around specific facial muscles while exhibiting small differences for most areas. Based on this, we design the detailed structure of TDN (in Figure 4(b)) as follows: **(1)** apply average pooling on feature differences $\{F_t - F_{t-T}, F_t - F_{t+T}\}$; **(2)** extract motion features with convolution layers (1×1 Conv, Convblock); **(3)** add the motion features with static features.

V. Results

In this section, we first introduce datasets, then explain evaluation metrics, and finally conduct the experimental analysis. We design a network (AUNet) that takes a tuple of facial frames as input and predicts the presence of AUs of the in-between frame in an end-to-end manner with the landmark-centred features (Section IV-A), variational auto-encoder module (Section IV-B), and frame difference module (Section IV-C). Accordingly, we divide the analysis part into three subsections to examine the good properties of landmark-centred features, perform ablate studies on sub-modules, and finally compare the proposed method with the state-of-the-art works.

A. Datasets

BP4D [81] contains 41 participants (23 females and 18 males) involved in eight spontaneous expression sessions. Totally, 328 videos with $\sim 140,000$ frames were recorded and annotated with 12 AUs (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, and AU24). Each AU has a binary annotation with "0" representing the absence of AU and vice versa. We follow the convention of [29], [56], [57] to split the data and conduct subject-exclusive 3-fold cross-validation.

DISFA [47] involves 27 participants (12 females, 15 males). Spontaneous facial expressions of participants from two views were recorded when they were asked to watch videos. Over 100,000 frames were annotated with 6-level intensities $\{0,1,2,3,4,5\}$ of different AUs. "0" indicates the absence of AU while "5" shows its most expressive degree. Similar to the settings in [56], [61], we extracted frames captured by the left camera, used threshold "2" to binarize the intensities, and conducted subject-exclusive 3-fold cross-validation on 8 frequent AUs (AU1, AU2, AU4, AU6, AU9, AU12, AU25, and AU26). Considering the serious AU imbalance issue in DISFA, we fine-tuned the pre-trained weights on BP4D to alleviate model overfitting [29], [56].

AffWild2 [33] consists of 56 videos from 63 subjects (32 males, 31 females). Different from BP4D and DISFA collected in constrained laboratory conditions, AffWild2 was captured in the wild and exhibited wide variability in illumination, occlusions, backgrounds, *etc.*. Similar to BP4D, AffWild2 provides binary annotations for eight AUs (AU1, AU2, AU4, AU6, AU12, AU15, AU20, AU25). Following the official guidance from [33], we used $\sim 227,000$ images as the training set and the rest of $\sim 67,0000$ as the validation set.

B. Performance metric

To be comparable to the existing works, we adopt F1-score [31] as a performance metric, which is widely used in the multi-label classification task. F1-score is calculated by

$$\mathrm{F1} = \frac{2 * \boldsymbol{p} * \boldsymbol{r}}{\boldsymbol{p} + \boldsymbol{r}}, \tag{5}$$

where $\boldsymbol{p}$ and $\boldsymbol{r}$ denote the precision and recall of a class, respectively. Comparisons are done on individual AU and their average (abbreviated as "Avg").

C. Implementation

The training and testing processes are performed on the open-source PyTorch framework. All experiments can be done on a single 1080Ti GPU. For the latent dimension in the VAE module, we set $k = 8$ for the BP4D database, $k = 6$ for the DISFA database and $k = 5$ for the AffWild2 database. The VAE module is trained by Adam optimizer for 100 epochs with a learning rate of 0.001 and without weight decay. λ is selected from $\{0.1,0.2,0.3\}$. γ is set to 1. The encoder from the last epoch is taken as a regulator.

The face images are cropped and resized to 256×256 pixels with the rectangles provided by RetinaFace [14]. The geometric augmentations include flipping, rotation (+-15 degree), centre shift (+-10 pixel), and scaling (+-0.25). The content augmentations are implemented with public library albumentations [8], including Gaussian Blur, ImageCompression, and CoarseDropout. The face alignment network is a two-hourglass based FAN [7]. During the training phase, we employ the AdamW optimizer, with an initial learning rate of 1e-4, degrading every 4 epochs by 30% and ending at 12 epochs. The batch size is set as 64. For better performance, hyper-parameters are determined by a grid search.

D. Good properties of AU-FAN

Our premise for an efficient facial AU detector hinges on two propositions: **(1)** semantic facial representations are superior to the global representation learned on raw images, and **(2)** a pre-trained face alignment network provides rich semantic (*i.e.* landmark) information for the AU detection.

Baselines To elaborate the good properties of features from the face alignment model, we compare AU-FAN with vanilla ResNet18 [28], JÂANet [57], and AU-FP [40]. Specially, ResNet18 [28] is trained on raw images and it has similar computation complexity to AU-FAN; JÂANet [57] is built upon a multi-task framework composed of the face alignment task and AU detection task; and AU-FP [40] is learned on proxy features provided by a pretrained face parsing model. For JÂANet, we use the released models for evaluation. For the rest, we use the same training settings as AU-FAN.

Setting To test the generalization ability of different models, we employ BP4D [81] as the training data and evaluate models on datasets from three diverse sources: **(1)** the test split of BP4D [81], **(2)** the DISFA [47] dataset captured in the spontaneous environment, and **(3)** the AffWild2 [33] database captured in the unconstrained environment.

TABLE III: Evaluating facial AU detection models on BP4D. Metric:F1-score (%). Training set: BP4D.

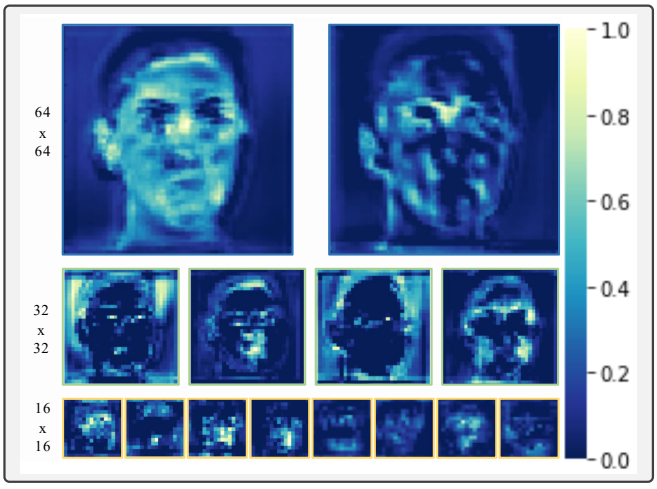
Methods	ResNet18 [28]	JÂANet [57]	AU-FP [40]	**AU-FAN**
AU1	46.2	53.8	53.5	**54.2**
AU2	37.3	**47.8**	43.5	44.9
AU4	50.4	58.2	58.5	**61.6**
AU6	74.9	**78.5**	77.9	76.8
AU7	74.9	**78.6**	76.2	76.6
AU10	80.4	82.7	82.9	**83.6**
AU12	84.3	88.2	87.4	**88.8**
AU14	58.3	63.7	63.5	**63.9**
AU15	34.4	43.3	47.5	**52.3**
AU17	58.6	61.8	63.3	**65.7**
AU23	36.5	45.6	**48.8**	48.5
AU24	43.3	**49.9**	48.6	48.0
Avg	56.6	62.4	62.6	**63.8**

TABLE IV: Evaluating facial AU detection models on DISFA. Metric:F1-score (%). Training set: BP4D.

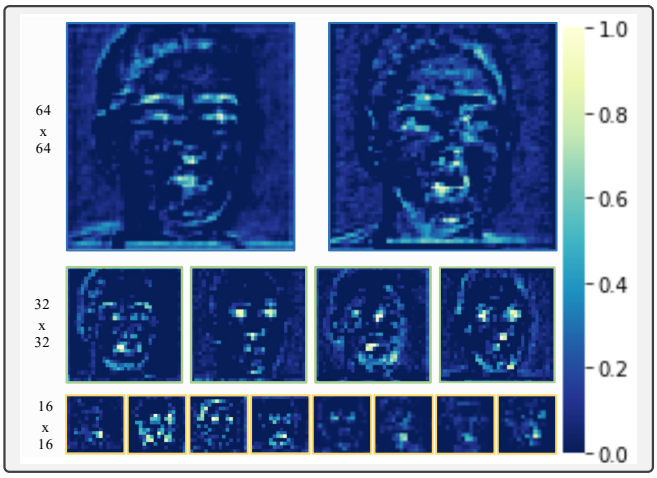
Methods	ResNet18 [28]	JÂANet [57]	AU-FP [40]	**AU-FAN**
AU1	8.5	19.2	21.9	**28.7**
AU2	8.5	16.1	**18.5**	17.2
AU4	13.6	28.5	41.9	**51.2**
AU6	16.5	30.6	**37.7**	36.8
AU12	26.6	35.4	**46.5**	45.8
Avg	14.8	26.0	33.3	**35.9**

TABLE V: Evaluating facial AU detection models on AffWild2. Metric:F1 score (%). Training set: BP4D.

Methods	ResNet18 [28]	JÂANet [57]	AU-FP [40]	**AU-FAN**
AU1	19.8	**42.3**	41.9	39.2
AU2	1.8	2.0.	2.4	**3.4**
AU4	11.8	16.8	19.4	**28.8**
AU6	12.3	**18.9**	18.7	15.1
AU12	14.5	19.5	21.0	**25.4**
AU15	0.1	0.9	14.9	**24.1**
Avg	10.0	16.7	19.7	**22.6**



(a) ResNet18



(b) AU-FAN

Fig. 6: Visualization of feature maps from ResNet18(a) and AU-FAN(b). ResNet18 is confused of where to focus on the facial imagery while AU-FAN resolutely emphasizes the facial landmark-related locations across different levels.

Results The results are compared in Table III, Table IV and Table V. We make the following observations:

Firstly, JÂANet, AU-FAN, AU-FP surpass ResNet18 in both within-database (BP4D) and cross-database (DISFA and AffWild2) evaluations. This suggests that attention-based middle-level features (*e.g.* facial landmarks and regions) are more representative than descriptors learned on the raw image. Further, to analyse of behaviors of AU-FAN and ResNet18, we visualize their feature maps at three resolutions (64×64, 32×32, 16×16). It is evident by Figure 6 that the neural activations of ResNet18 are distributed across the entire face without specific concentration across all levels. By contrast, AU-FAN is capable of perceiving the landmark-centred regions at all times and capable of alleviating the disturbance from unrelated regions. Besides, we do another experiment by replacing FAN in AU-FAN with ResNet18 backbone pre-trained on ImageNet classification task. Compared with ResNet18 trained from scratch, this variant increases the F1-score from 56.6% to 61.5% on BP4D. Overall, these results

confirm the effectiveness of the pre-trained features for AU detection and the extra advantages brought by face-related representations.

Secondly, AU-FAN is superior to JÂANet, especially in the cross-database evaluation. Although JÂANet and AU-FAN both rely on the facial landmarks for semantic attention, the alignment branch in JÂANet was trained on mere AU datasets while the counterpart in AU-FAN was optimized in more challenging face alignment datasets [53], [59], [80], [88]. Due to the complexity and difficulty of AU detection task, it is hard to surpass the predecessors in all AUs (see Table XIII and Table XIII). The overall performance of AU-FAN is superior to the sophisticatedly-designed JÂANet in the majority of AUs. Compared to JÂANet composed of AU attention refinement and local AU feature learning to specify region-of-interest of each AU, AU-FAN uses a global representation for all AUs' detection, which could cause difficulty in capturing weaker appearance changes from AU2 and AU6. Different from other AUs, AU2 (outer brow raiser) and AU6 (cheek raiser) highly co-relate with other AUs in occurrences (see adjacent matrix in Figure 8). The strong correlation is caused by the shared muscle (AU1 and AU2) and common expression (AU6 and AU12). Compared to AU1 (inner brow raiser) that causes obvious wrinkles in forehead and obvious shape changes of eyebrow, changes caused by AU2 are weaker. Compared to AU12 (lip corner puller) that causes lip shape changes around landmark 48 and 54, AU6 is less landmark specific (Table I). Such a situation can be mitigated by integrating the property of AU co-occurrences into network design so that the presence of one AU can be exploited to predict another AU more reliably. This hypothesis is evident in Section V-E that regulating AU-FAN prediction with label prior (AU-VAE) significantly increases their performance.

Thirdly, AU-FAN yields consistently better performance than AU-FP. AU-FP focuses on facial parts (*e.g.* eyes, nose, mouth) via pixel-wise classification while AU-FAN concentrates on facial landmarks. The latter directly encodes facial shape and expression appearance. The superior performance of AU-FAN suggests the overall performance advantage and network robustness despite its simplicity.

In a nutshell, this section shows that attention features are beneficial for describing AUs and those captured in a pre-trained face alignment are especially discriminative and representative for describing AUs. In the following experiments, we hence use AU-FAN as the default design, unless stated otherwise.

Failure cases By no means do we claim that the proposed method AU-FAN solves the real-world AU detection problem. Since AU-FAN is built upon a pre-trained FAN model, its performance is influenced by FAN. As a result, if FAN fails to provide reliable landmarks in situations like severe occlusion, low resolution, and extremely large pose, AU-FAN may give incorrect AU predictions (see Figure 7).

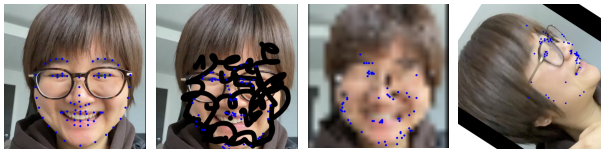


Fig. 7: Examples of failure cases. AU12 is correctly predicted when landmarks around mouth are reliably predicted. However, when FAN can not provide reliable landmark detection in situations like severe occlusion, low resolution and extremely large pose, AU-FAN mistakenly predict AU12 as deactivation.

E. Ablation Studies

Setting To make the best of the manual annotations in Section V-A and obtain a model able to predict fifteen AUs, we train AU-FAN on the mixture of datasets. We follow the subject-exclusive principle to divide image samples into training and testing splits. The training split has $\sim 401,000$ samples: $\sim 100,000$ from BP4D, $\sim 83,000$ from DISFA, and $\sim 227,000$ from AffWild2. The testing split has three parts: $\sim 46,000$ from BP4D, $\sim 43,000$ from DISFA, and $\sim 67,000$ from AffWild2.

Competitors We mainly compare four variants:AU-FAN, AU-VAE, AU-TDN and AUNet (see Table VI). More specifically, AU-FAN is a baseline network; AU-VAE trained AU-FAN with VAE module proposed in Section IV-B. Each database owns its respective VAE module; AU-TDN accommodated AU-FAN with AU-TDN in Section IV-C by setting interval T to 10; AUNet trained AU-FAN with AU-VAE in Section IV-B and AU-TDN in Section IV-C.

TABLE VI: Description of four variants.

Methods	FAN feature	VAE	TDN
AU-FAN	✓	✗	✗
AU-VAE	✓	✓	✗
AU-TDN	✓	✗	✓
AUNet	✓	✓	✓

TABLE VII: Evaluation of ablate models on BP4D. Metric:F1 score (%). Training set: mixture of BP4D, DISFA and AffWild2.

Methods	AU-FAN	AU-VAE	AU-TDN	**AUNet**
AU1	54.9	54.9	56.5	**59.8**
AU2	46.6	**53.0**	50.2	47.3
AU4	61.0	61.1	63.1	**63.9**
AU6	82.9	83.8	82.3	**83.9**
AU7	74.5	74.6	74.2	**76.8**
AU10	86.2	**88.2**	87.5	87.5
AU12	87.5	88.3	88.8	**90.5**
AU14	50.7	63.2	**64.4**	61.2
AU15	40.0	44.8	**48.5**	46.8
AU17	64.6	66.6	**67.2**	66.1
AU23	47.9	37.8	50.4	**50.9**
AU24	47.6	43.4	45.8	**51.3**
Avg	62.7	63.3	64.9	**65.5**

Results According to the results in Table VII, Table VIII and Table IX, we analyse the following factors.

Regularization effect of AU-VAE AU-VAE yields the consistent improvements over AU-FAN over all three databases. Statistically, AU-VAE increases the average F1-score by 0.6%

TABLE VIII: Evaluation of ablate models on DISFA. Metric:F1 score (%). Training set: mixture of BP4D, DISFA and AffWild2.

Methods	AU-FAN	AU-VAE	AU-TDN	**AUNet**
AU1	69.2	68.4	72.4	**75.9**
AU2	69.0	70.9	71.5	**72.1**
AU4	67.6	76.8	74.9	**77.3**
AU6	53.3	**59.7**	55.8	59.5
AU9	**70.3**	62.6	49.6	54.0
AU12	71.2	**76.9**	75.4	74.9
AU25	81.7	86.0	85.7	**91.9**
AU26	51.7	**56.4**	55.0	54.5
Avg	66.7	69.8	67.5	**70.0**

TABLE IX: Evaluation of ablate models on AffWild2. Metric:F1 score (%). Training set: mixture of BP4D, DISFA and AffWild2.

Methods	AU-FAN	AU-VAE	AU-TDN	**AUNet**
AU1	74.8	74.7	76.2	**76.6**
AU2	0.0	0.0	0.0	0.0
AU4	65.9	69.9	63.1	**72.3**
AU6	12.0	16.0	10.7	**31.5**
AU12	45.4	**48.2**	46.5	42.8
AU15	59.0	82.4	66.8	**85.8**
AU25	0.0	0.0	0.0	0.0
Avg	36.7	41.6	38.4	**44.1**

in BP4D, 3.1% in DISFA and 4.9% in AffWild2, respectively. Notably, since the VAE module is omitted in the inference stage, its performance gains are achieved without incurring any computational complexity in deployment.

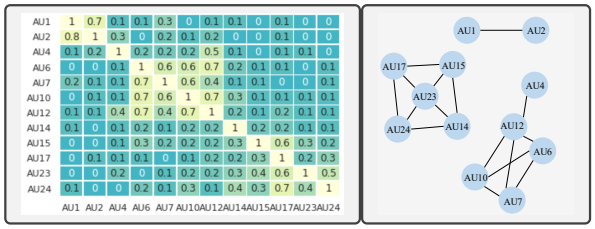
The generic objective of AU-VAE is to encourage the predictions to mimic the dependencies of AUs. It is hence insightful to evaluate this mimicry quality. To do so, we utilize the L2 distance between adjacency matrices [50] and dependency F1-score [65] to measure the faithfulness. Results in Table X show AU-VAE achieves better performance on both metrics. These findings reveal that the learned correlations by VAE are more faithful to annotations and help to improve performance. Figure 8 contains the adjacency matrix and AU linkage graph of AU-VAE on BP4D.

TABLE X: Performance differences between AU-VAE and AU-FAN in terms of L2 distance (lower the better) and dependency F1-score (higher the better).

Metrics	BP4D	DISFA	AffWild2
L2 distance(↓)	2.2%	6.5%	2.1%
Dependency F1-score(↑)	1.5%	3.1%	6.3%

We also compare the aligning output of the encoder (latent space) and aligning output of the decoder (label space). We found that aligning label space deteriorates the performance (65.0% vs. 65.5% in BP4D). This verifies the effectiveness of aligning latent space. Since aligning latent space only requires activating encode only, it also saves training time.

What is learned in TDN Compared with AU-FAN, AU-TDN increased the average F1-score by 2.2%, 0.6% 1.8% in BP4D, DISFA, AffWild2, respectively. The majority of AUs benefit from capturing appearance variations in BP4D except for AU6, AU7 and AU24. It indicates that the static descriptors provide rich evidence for AU6, AU7 and AU24 while the geometric variations from frame difference may provide fuzzy support. For example, AU6 wrinkles the skin around the eye corner and raises the cheeks, which makes displacement difficult to perceive this change. In contrast, these transient differences are captured well in facial textures like wrinkles, bulges and furrows in a static image. To visually demonstrate the interpretation of AU-TDN, we illustrate some feature maps learned by the static and dynamic branches of AU-TDN in Figure 9. The static representations imply a clear facial structure. By contrast, the dynamic representations roughly indicate where is affected by movement, in which the general appearances (e.g. face shapes, contours) are averaged out. For example, in AU2, the dynamic branch draws the network's attention to the forehead with obvious wrinkles when the eyebrow is raised. We also complement the temporal information by integrating FlowNet [19] to AU-FAN. The comparison is in Table XI. The results show both will enhance the single-frame baseline. In comparison, TDN is lightweight in model size and can achieve better performance.



(a) Adjacency matrix (b) Graph

Fig. 8: Adjacency matrices and graph of predictions of AU-VAE on BP4D. Each entry (i, j) is computed as the coefficient correlation between the i-th AU and the j-th AU.

TABLE XI: Comparison between adding different temporal modeling methods to AU-FAN on BP4D.

Method	Model size(M)	F1-score(%)
+TDN	2.66	64.9
+FlowNet	187.75	64.5

In Figure 10, we visualize the predictions on a sample video sequence from the BP4D. When compared with the outputs from the AU-FAN, AU-TDN generates predictions that better overlap with the ground-truth annotations with fewer breaks.

Furthermore, We create AUNet by combining AU-FAN, AU-VAE, AU-TDN, which obtains another 0.6%, 0.2% and 2.5% average F1-score gains for BP4D, DISFA and AffWild2 respectively. This implies that AU-VAE and AU-TDN are complementary in the proposed approach.

Significance anlysis We conduct paired t-tests on predictions between AU-FAN and its improved variants (AU-VAE, AU-TDN and AUNet) on BP4D in Table XII. The p values returned are lower than threshold 0.05 which means the null hypothesis is rejected. We argue argue that performance of AU-VAE,

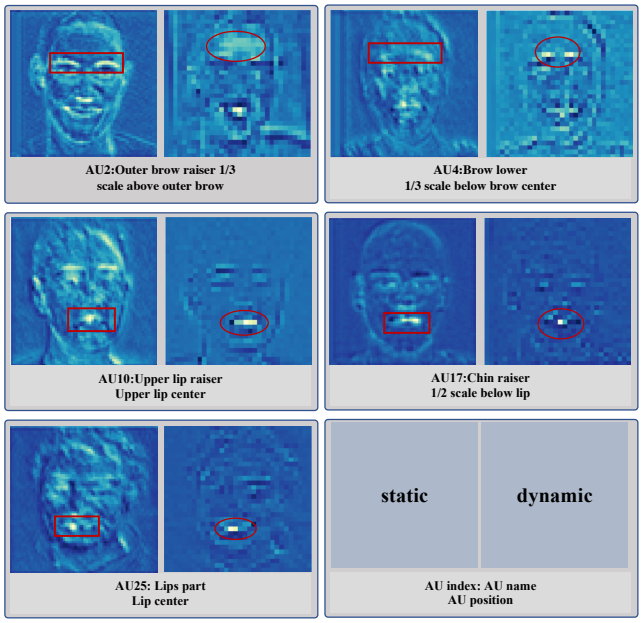


Fig. 9: Visualization of feature maps in AU-TDN module. Each AU group contains a static frame representation, describing the edges and contours of the face and the frame-difference representation, describing activation from AU movement.

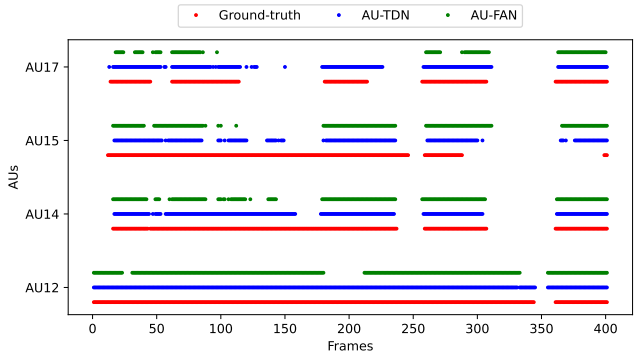


Fig. 10: Visualization of model predictions for different time steps (x-axis) for various actions (y-axis) from a BP4D video. We compare AU-FAN(green), AU-TDN(blue), and the ground-truth (red). The predictions of AU-TDN are more continuous than AU-FAN.

AU-TDN and AUNet are significantly different and improved compared to the baseline performance of AU-FAN.

TABLE XII: Paired t-Tests between AU-FAN and its variants on BP4D.

Method	t-statistic	p-value
AU-VAE	-26.78	6.33×10^{-158}
AU-TDN	-26.94	7.62×10^{-160}
AUNet	-26.40	1.51×10^{-153}

F. Comparison with State-of-the-Art Methods

Baselines To verify the effectiveness of proposed approach, we compare AUNet, and its single frame baseline AU-FAN with state-of-the-art works including the methods focusing on attention or regional features: DRML [87] and JÂAANet [57], methods exploiting the AU correlations: SRERL [34], HMP-PS [62], and MEGraph [ResNet50] [43], methods leveraging dynamic change: AR-LSTM [45] and SAT-Net [39]. The results for the aforementioned methods are taken from the papers [39], [61], [62]. To show the generality of proposed modules, we instantiate ResNet18 [28] with AU-VAE and AU-TDN, dubbed as ResNet18-A, and have them in comparison.

Results Table XIII and Table XIV show the results of different methods on the BP4D [81] database and DISFA dataset [47], respectively. The comparisons demonstrate that our AUNet are able to yield a new SOTA performance on both datasets in terms of average F1-score. At the same time, AUNet achieves the best or second-best detection performance for 6/12 in BP4D and the 6/8 in DISFA.

Besides, AU-FAN alone achieves comparable performance with SOTA methods, 0.4% average F1-score higher than the latest HMP-PS [62] in BP4D, and comparable to JÂAANet [57] in DISFA, which indicates that our feature representation extracted from the face alignment model is informative and effective. In addition, VAE and TDN are easily accommodated into ResNet18 [28], boosting up ResNet18's average F1-score by $\sim 3\%$ in BP4D and $\sim 5\%$ in DISFA. This shows that the proposed VAE and TDN modules are flexible to network structures.

Complexity discussion Compared to existing works, AUNet is less complex and it has far fewer parameters. In terms of the network structure, SRERL [34] contains VGG19, a cropping module and graph convolution network, JÂAANet [57] combines hierarchical and multi-scale region learning, face alignment and global feature learning and adaptive attention learning, and MEGraph [43] is trained in a two-stage manner to optimize assigning AU-specific feature, generating a facial graph, learning specific AU representation modeling, then modeling AU relationship.

Most works compared did not disclose their model size. To the best of our knowledge, JÂAANet has $\sim 25M$ parameters with $\sim 8.38G$ flops, ResNet18 has $\sim 11.38M$ parameters with $\sim 2.35G$ flops, and MEGraph has $\sim 31.67M$ parameters with $\sim 10.29G$ flops. Except for the face alignment module, $\sim 12.11M$ parameters with $\sim 13.97G$ flops, the trainable parameters of AUNet is $\sim 2.66M$ with $\sim 3.84G$ flops and that of AU-FAN is $\sim 1.86M$ with $\sim 3.02G$ flops. Both the model size and flops of our proposed models are comparable to ResNet18, but they have significantly superior performance.

VI. Challenges

FACS [21] is a powerful means for measuring facial activity. However, annotating AU data sets based on the proposed AU labeling rules is expensive, time-consuming and error-prone. It takes more than 100 hours to train an expertise AU annotator, and it costs 30 minutes or more to manually code 1 AU for a one-minute video [17], [21]. Because of the labor-intensive annotating process, the available data sets for training AU detectors (see Table XV) are constrained by the number of

TABLE XIII: F1-score (%) on BP4D. Comparison with SOTA. Bold numbers indicate the best performance; Underline numbers indicate the second best.

Methods	DRML	JÂANet	SRERL	HMP-PS	MEGraph	AR-LSTM	SAT-Net	ResNet18	ResNet18-A	AU-FAN	**AUNet**
Ref	[87]	[57]	[34]	[62]	[43]	[45]	[39]	[28]	Ours	Ours	Ours
Year	2016	2021	2019	2021	2022	2019	2021	2016	This	This	This
AU1	36.4	53.8	46.9	53.1	53.7	48.0	54.1	46.2	52.1	54.2	**58.0**
AU2	41.8	47.8	45.3	46.1	46.9	43.2	**49.5**	37.3	42.5	44.9	48.2
AU4	43.0	48.2	55.6	56.0	59.0	53.1	58.3	50.4	54.6	61.5	**62.4**
AU6	55.0	**78.5**	77.1	76.5	**78.5**	76.9	77.7	74.9	76.6	76.8	76.4
AU7	67.0	75.8	78.4	76.9	**80.0**	78.4	77.7	74.9	73.7	76.6	77.5
AU10	66.3	82.7	83.5	82.1	**84.4**	82.8	83.6	80.4	80.4	83.6	83.4
AU12	65.8	88.2	87.6	86.4	87.8	87.9	86.5	84.3	85.2	**88.8**	88.5
AU14	54.1	63.7	63.9	64.8	67.3	**67.7**	63.2	58.3	60.3	63.9	63.3
AU15	33.2	43.3	52.2	51.5	**52.5**	45.6	49.1	34.4	36.7	52.3	52.0
AU17	48.0	61.8	63.9	63.0	63.2	63.4	61.8	58.6	63.3	**65.7**	65.5
AU23	31.7	45.6	47.1	49.9	50.6	47.9	48.7	36.5	34.0	48.5	**52.1**
AU24	30.0	49.9	53.3	54.5	52.4	**56.4**	49.3	43.3	53.5	48.0	52.3
Avg	48.3	62.4	62.1	63.4	64.7	62.6	63.3	56.6	59.4	63.8	**65.0**

TABLE XIV: F1-score (%) on DISFA. Comparison with SOTA. Bold numbers indicate the best performance; Underline numbers indicate the second best.

Methods	DRML	JÂANet	SRERL	HMP-PS	MEGraph	AR-LSTM	SAT-Net	ResNet18	ResNet18-A	AU-FAN	**AUNet**
Ref	[87]	[57]	[34]	[62]	[43]	[45]	[39]	[28]	Ours	Ours	Ours
Year	2016	2021	2019	2021	2022	2019	2021	2016	This	This	This
AU1	17.3	**62.4**	45.7	38.0	54.6	26.9	41.2	4104	52.4	59.3	60.3
AU2	17.7	**60.7**	47.8	45.9	47.1	24.4	33.1	47.8	46.5	55.3	59.1
AU4	37.4	67.1	59.6	65.2	**72.9**	58.6	63.0	43.4	67.0	69.4	69.8
AU6	29.0	41.1	47.1	50.9	54.0	49.7	**56.4**	43.1	43.6	49.0	48.4
AU9	10.7	45.1	45.6	50.8	**55.7**	34.2	43.0	37.2	38.5	45.9	53.0
AU12	37.7	73.5	73.5	76.0	76.7	71.3	73.1	73.8	71.3	77.0	**79.7**
AU25	38.5	90.9	84.3	93.3	91.1	83.4	82.9	82.4	87.3	91.8	**93.5**
AU26	20.1	67.4	43.6	**67.6**	53.0	51.4	60.6	61.8	67.1	60.0	64.7
Avg	26.7	63.5	55.9	61.0	63.1	50.0	56.7	53.8	59.2	63.5	**66.1**

coded AUs, samples, and subjects, thus causing the over-fitting problem in model learning.

Besides, the reliability of annotation is attenuated because of the ambiguous nature of AUs as well as the subjective difference. Some AUs are obvious and easy to annotate (*e.g.* AU4, AU9) while others are subtle and hard to annotate (*e.g.* AU7, AU23, AU24). A lack of assessment of inter-rater reliability for existing data sets further potentially aggravate any inaccurate annotations. As a result, models trained on data sets with inconsistent annotations often have generalisation and deployment issues when it comes to real-world applications.

TABLE XV: Comparison among existing AU data sets.

Data set	#Subjects	#Faces	#AUs	Video	In the wild
BP4D [81]	41	140K	12	Y	N
DISFA [47]	27	100K	12	Y	N
FERA2015 [24]	85	350K	14	Y	N
EmotioNet [23]	-	960K [1]	11	N	Y
Affwild2 [33]	63	390K	8	Y	Y

Another limitation of current spontaneous data sets is the AU label imbalance. One reason explaining such phenomenon is the practicalities of data acquisition For example, when collecting BP4D [81], participants kept neutral expressions the most of time when they were watching videos. Thus, the number of inactive samples is much larger than that of active ones for some AUs. Apart from this, annotating positive expressions (*e.g.* smile) is easier than capturing appearance evidence for negative expressions such as disgust, and anger. Therefore, AUs associated with common expressions occur more frequently than those with less common expressions. Accordingly, the trained model performs worse on the less common AUs without enough data for learning good representations.

To augment the training set without intensive manual annotation, synthesizing new data [41] or using unlabeled data [50], [74] are two straightforward solutions. Built upon the development of generative models, self-supervised learning, semi-supervised learning, and weakly supervised learning, these methods have achieved comparable or superior performance to the supervised baselines in public benchmarks. However, how these models generalize to real-world scenarios is still unclear. We also argue some potential risks would be brought by the extra data because these methods did not consider facial attributes (*e.g.* ethical background, age ranges, subjects) in creating the extra data.

VII. A STATEMENT ON ETHICS

Generalisation capabilities Given that the FACS [21] rules used for annotating AUs (see TableI) are identity-invariant, any AU detector trained on an annotated dataset according to FACS is also supposed to work in an identity-invariant

[1]Note that these are not manually annotated but automatically labeled by the method described in [23]

fashion. However, since AU labeling is labor-intensive and time-consuming, the available AU datasets are small in size and not diverse in the number of identities they contain. Thus, one can't really guarantee that the trained methods are fair with respect to certain protected attributes — age, gender, and ethnicity. We have reviewed some works that propose ways for enhancing the generalization capabilities of AU detection models, in Section III-D. The core idea is to extract AU features orthogonal to identity features.

Performance measurement Normally AU datasets are annotated by professionally-trained annotators under the guidance of manual FACS [21] rules. The ground-truth label is a vector composed of "1" denoting the presence of an AU and "0" representing the absence of a particular AU. For performance measurement, two experimental protocols are commonly used: subject-independent and cross-datasets evaluations. In subject-independent evaluation we split the entire dataset into training and testing splits in a subject-independent manner. The model trained on the training set is evaluated on the test split to show its generalizability to different identities. We can further divide the datasets K times and perform K-fold cross-validation. Alternatively, in a cross-dataset evaluation regime, one dataset is used to train the model and another dataset is to test it. For example, in Section V, we train on BP4D, then test on DISFA and AffWild2. Accuracy and F1-score are two popular metrics for evaluating the AU detection models. Based on recent works, F1-score is more widely used. Cross-dataset is better than single-dataset theoretically but in practice, the datasets might be too different - *e.g.* the large pose variations in AffWild2. Therefore if any models are evaluated cross-dataset we should only compare the results relatively speaking —*i.e.* which model transfers best from the training dataset to the test set, rather than looking at the absolute results - F1-score and accuracy

Privacy and Ethical considerations Since AUs are annotated on human faces which contain identity information, most AU datasets are for research and some [1], [12] are not publicly available due to privacy issues. AU datasets represent a potentially valuable trove of automatic facial behavior analysis. In the deep learning era, all models are data-hungry but many labeled AU datasets in individual institutes can not be shared with other researchers. Developing methods to allow other researchers to benefit from these data without having direct access to them would greatly benefit the community. Federated Learning [79] could have great potential in this direction.

Technology misuse Since facial AU detection systems are designed to analyse nonverbal information in communication, they should be deployed and monitored under the necessary regulations to avoid issues like the invasion of privacy. For example, if this technology is provided in a way where personal and identity-specific information is revealed or leaked, it can be used to trace these subjects and their behavior patterns for targeting political or other aims. These issues should be considered when building upon this work to applications.

VIII. Conclusion

Automating facial AU detection facilitates the development of facial behavior analysis. In this work, we summarize three key factors comprising a good facial AU detector: **(1)** being able to capture local AU-centered features, **(2)** being able to exploit the fact that some action units co-occur with others, and **(3)** being able to utilize appearance changes across frames. Accordingly, we propose AUNet, a simple yet strong baseline for landmark-based AU detection by **(1)** using the intermediate layers of a pre-trained face alignment model to act as our AU features space and by being optimised to satisfy the **(2)** AU correlation constraint, derived from the AU labels, as well as a **(3)** temporal constraint, derived from variations in the contents of consecutive frames in the input videos. We further conduct an empirical analysis of how each sub-module increases model performance. We observe that features from AU-FAN are AU-focused, AU-VAE enables detect co-occurring AUs, and AU-TDN enriches static representation. Experiments on BP4D and DISFA show that the proposed AUNet while having a simpler pipeline and lower computation overhead, has achieved state-of-the-art results. We hope our work will offer some insights into the future of machine detecting facial AUs.

References

[1] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker, et al. From joyous to clinically depressed: Mood detection using spontaneous speech. In *FLAIRS Conference*, 2012.

[2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[3] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshop*, 2013.

[4] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[5] J. N. Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 1979.

[6] A. Bulat, S. Cheng, J. Yang, A. Garbett, E. Sanchez, and G. Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. In *European Conference on Computer Vision*, 2022.

[7] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[8] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020.

[9] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 2020.

[10] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[11] Y. Chang and S. Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[12] J. F. Cohn and M. A. Sayette. Spontaneous facial expression in a small group can be automatically measured: An initial demonstration. *Behavior Research Methods*, 2010.

[13] C. Corneanu, M. Madadi, and S. Escalera. Deep structure inference network for facial action unit recognition. In *European Conference on Computer Vision*, 2018.

[14] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

[15] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal on Computer Vision*, 2019.

[16] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[17] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 1999.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[19] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[20] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 1971.

[21] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[22] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *IEEE International Conference on Computer Vision*, 2015.

[23] F. B.-Q. C. et al. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[24] V. M. et al. Fera 2015-second facial expression recognition and analysis challenge. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[25] W. V. Friesen, P. Ekman, et al. Emfacs-7: Emotional facial action coding system. 1983.

[26] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 2005.

[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[29] G. M. Jacob and B. Stenger. Facial action unit detection with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[30] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[31] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, 2013.

[32] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[33] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2020.

[34] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[35] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[36] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017.

[37] X. Li, Z. Li, H. Yang, G. Zhao, and L. Yin. Your "attention" deserves attention: A self-diversified multi-channel attention for facial action analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021.

[38] Y. Li, J. Zeng, S. Shan, and X. Chen. Self-supervised representation learning from videos for facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[39] Z. Li, Z. Zhang, and L. Yin. Sat-net: Self-attention and temporal fusion for facial action unit detection. In *International Conference on Pattern Recognition (ICPR)*, 2021.

[40] Y. Lin, J. Shen, Y. Wang, and M. Pantic. Fp-age: Leveraging face parsing attention for facial age estimation in the wild. *arXiv*, 2021.

[41] Z. Liu, D. Liu, and Y. Wu. Region based adversarial synthesis of facial action units. In *International Conference on Multimedia Modeling*, 2020.

[42] B. Luo, J. Shen, S. Cheng, Y. Wang, and M. Pantic. Shape constrained network for eye segmentation in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

[43] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[44] B. Ma, R. An, W. Zhang, Y. Ding, Z. Zhao, R. Zhang, T. Lv, C. Fan, and Z. Hu. Facial action unit detection and intensity estimation from self-supervised representation. *arXiv preprint arXiv:2210.15878*, 2022.

[45] C. Ma, L. Chen, and J. Yong. Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. *neurocomputing*, 2019.

[46] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 2017.

[47] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE transactions on affective computing*, 2013.

[48] S. Minaee, P. Luo, Z. Lin, and K. Bowyer. Going deeper into face detection: A survey. *arXiv preprint arXiv:2103.14983*, 2021.

[49] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, 2016.

[50] X. Niu, H. Han, S. Shan, and X. Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. In *Advances on Neural Information Processing Systems*, 2019.

[51] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[52] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya. Learning and incorporating shape models for semantic segmentation. In *International conference on medical image computing and computer-assisted intervention*, 2017.

[53] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshop*, 2013.

[54] K. R. Scherer. *Handbook of methods in nonverbal behavior research*. Cambridge University Press, 1985.

[55] L. Shang and K.-P. Chan. Nonparametric discriminant hmm and application to facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[56] Z. Shao, Z. Liu, J. Cai, and L. Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *European Conference on Computer Vision*, 2018.

[57] Z. Shao, Z. Liu, J. Cai, and L. Ma. Jaa-net: joint facial action unit detection and face alignment via adaptive attention. *International Journal on Computer Vision*, 2021.

[58] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 2019.

[59] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *IEEE International Conference on Computer Vision Workshop*, 2015.

[60] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[61] T. Song, L. Chen, W. Zheng, and Q. Ji. Uncertain graph neural networks for facial action unit detection. In *AAAI Conference on Artificial Intelligence*, 2021.

[62] T. Song, Z. Cui, W. Zheng, and Q. Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[63] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[64] Y. Tang, W. Zeng, D. Zhao, and H. Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *ICCV*, 2021.

[65] P. Tirupattur, K. Duarte, Y. S. Rawat, and M. Shah. Modeling multi-label action dependencies for temporal action localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[66] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021.
[67] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 2007.
[68] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2011.
[69] R. Walecki, V. Pavlovic, B. Schuller, M. Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *IEEE International Conference on Computer Vision*, 2017.
[70] L. Wang, Z. Tong, B. Ji, and G. Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
[71] S. Wang, S. Wu, G. Peng, and Q. Ji. Capturing feature and label relations simultaneously for multiple facial action unit recognition. *IEEE transactions on affective computing*, 2017.
[72] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *IEEE International Conference on Computer Vision*, 2013.
[73] Y. Wu and Q. Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[74] J. Yan, J. Wang, Q. Li, C. Wang, and S. Pu. Weakly supervised regional and temporal learning for facial action unit recognition. *IEEE Transactions on Multimedia*, 2022.
[75] H. Yang and L. Yin. Learning temporal information from a single image for au detection. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2019.
[76] H. Yang and L. Yin. Re-net: A relation embedded deep model for au occurrence and intensity estimation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
[77] H. Yang, L. Yin, Y. Zhou, and J. Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
[78] J. Yang, A. Bulat, and G. Tzimiropoulos. Fan-face: a simple orthogonal improvement to deep face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
[79] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2019.
[80] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.
[81] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 2014.
[82] Y. Zhang, W. Dong, B.-G. Hu, and Q. Ji. Classifier learning with prior probabilities for facial action unit recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[83] Y. Zhang, H. Jiang, B. Wu, Y. Fan, and Q. Ji. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In *IEEE International Conference on Computer Vision*, 2019.
[84] Z. Zhang, T. Wang, and L. Yin. Region of interest based graph convolution: A heatmap regression approach for action unit detection. In *Proceedings of ACM International Conference on Multimedia*, 2020.
[85] Z. Zhang, S. Zhai, L. Yin, et al. Identity-based adversarial training of deep cnns for facial action unit recognition. In *British Machine Vision Conference*, 2018.
[86] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[87] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[88] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

IX. Biography Section

Jing Yang received the Ph.D. from University of Nottingham. Her research interest is deep face analysis.

Yordan Hristov is a research scientist at Meta AI London. He received his BSc with Honours in Computer Science from the University of Edinburgh in 2016 and defended his PhD in Machine Learning and Robotics from the University of Edinburgh in 2020. His research interests include deep generative models for facial data, disentangled representations and semi-supervised learning.

Jie Shen is a research scientist at Meta AI and an honorary research fellow at the Department of Computing at Imperial College London. He received his B.Eng. in electronic engineering from Zhejiang University in 2005, and his MSc in advanced computing and Ph.D. from Imperial College London in 2008 and 2014. His research interests include facial analysis, computer vision, affective computing, and social robots. He is a member of the IEEE.

Yiming Lin is a research scientist at Meta. He received his Ph.D. degree from Imperial College London in 2021. His research interests include face tracking, face parsing, face recognition and facial attribute analysis. He is a member of IEEE.

Maja Pantic is a professor in affective and behavioural computing in the Department of Computing at Imperial College London, UK. She was the Research Director of Samsung AI Centre, Cambridge, UK from 2018 to 2020 and is currently an AI Scientific Research Lead at Meta London. She currently serves as an associate editor for both the IEEE Transactions on Pattern Analysis and Machine Intelligence and the IEEE Transactions on Affective Computing. She has received various awards for her work on automatic analysis of human behaviour,including the Roger Needham Award 2011. She is a fellow of the UK's Royal Academy of Engineering, the IEEE, and the IAPR.