Facial Prior Guided Micro-Expression Generation

Yi Zhang, Xinhua Xu, Youjun Zhao, Yuhang Wen, Zixuan Tang, and Mengyuan Liu

***Abstract*— This paper focuses on the facial micro-expression (FME) generation task, which has potential application in enlarging digital FME datasets, thereby alleviating the lack of training data with labels in existing micro-expression datasets. Despite obvious progress in the image animation task, FME generation remains challenging because existing image animation methods can hardly encode subtle and short-term facial motion information. To this end, we present a facial-prior-guided FME generation framework that takes advantage of facial priors for facial motion generation. Specifically, we first estimate the geometric locations of action units (AUs) with detected facial landmarks. We further calculate an adaptive weighted prior (AWP) map, which alleviates the estimation error of AUs while efficiently capturing subtle facial motion patterns. To achieve smooth and realistic synthesis results, we use our proposed facial prior module to guide motion representation and generation modules in mainstream image animation frameworks. Extensive experiments on three benchmark datasets consistently show that our proposed facial prior module can be adopted in image animation frameworks and significantly improve their performance on micro-expression generation. Moreover, we use the generation technique to enlarge existing datasets, thereby improving the performance of general action recognition backbones on the FME recognition task. Our code is available at https://github.com/sysu19351158/FPB-FOMM.**

***Index Terms*— Facial micro-expression, image animation.**

I. Introduction

FACIAL micro-expression (FME) is a brief facial movement that reveals an emotion that a person tries to conceal [1], [2], [3]. Recently, micro-expression has drawn the attention of psychologists and computer scientists. Normal facial expressions, called macro-expressions, usually last for 0.5 to 4 seconds, while the duration of micro-expressions is less than 0.5 seconds [2]. The short duration and low intensity of facial movements make micro-expressions hard to recognize with the naked eye. FMEs can elucidate the relationship between repressed emotions and facial expression [2]. Hence, there are potential applications in various fields, such as criminal investigation, medical treatment, education, and business negotiations [4].

Manuscript received 30 August 2022; revised 2 May 2023 and 6 October 2023; accepted 4 December 2023. Date of publication 27 December 2023; date of current version 4 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62203476 and in part by the Natural Science Foundation of Shenzhen under Grant JCYJ20230807120801002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Raymond Fu. *(Corresponding author: Mengyuan Liu.)*

Yi Zhang, Xinhua Xu, and Mengyuan Liu are with the National Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Shenzhen 518055, China (e-mail: nkliuyifang@gmail.com).

Youjun Zhao, Yuhang Wen, and Zixuan Tang are with the School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510006, China.

Digital Object Identifier 10.1109/TIP.2023.3345177

The lack of training data with labels is believed to be one of the biggest problems in automatic FME analysis [5], [6]. Deep learning methods have achieved fine results in many computer vision fields. However, data-driven neural network methods have not been able to obtain dominant places in FME analysis, and many feature designs and network methods rely heavily on prior knowledge and manual features. A recent study [7] showed that neural network methods, with an adequate amount of supplementary training data, can achieve comparable performance on the FME discriminative task as methods dedicated to FME. This shows that providing more annotated training data for deep neural networks is a promising way of improving deep neural networks in automatic FME analysis tasks.

There are a limited number of experts who are facial action coding system certification holders since the general public has difficulties accessing large-scale FME datasets for practice. Constructing an FME dataset can be very expensive for several reasons. First, FMEs are involuntary expressions that require professionally designed instructions to elicit when constructing an FME database. Second, the duration of FMEs is extremely short and requires multiple high-frame-rate cameras and strict lighting conditions to capture accurately. Third, many experts are needed to label the collected FMEs. Although some datasets [1], [8], [9], [10], [11] have been provided, data shortages remain, not to mention the inconsistent configuration of each existing dataset. In [12], a GAN-based method was proposed to transfer facial movements in micro-expression. However, it could only synthesize a single image on an identical human face, instead of transferring the motion to any other human face or generating a whole video clip.

To solve the above problem, we introduce an FME generation task that can create novel FMEs. Inspired by the image animation task, our task aims to create a novel FME video by driving a target frame with motion information from a driving video. We compare the differences between our task and the traditional image animation task in Fig. 1, where the first two columns are the onset and apex frames of a driving video, and the last column denotes the target frame. As shown in Fig. 1 (a), our task focuses on encoding subtle facial motion information from a driving video. Meanwhile, Fig. 1 (b) shows that the traditional image animation task focuses on encoding macro motion information and handling different camera view problems. Currently, image animation methods [13], [14] have achieved considerable success by transferring macro motion information from driving videos to target frames using a self-supervised learning framework. Without involving strong supervision signals, these methods have difficulty encoding subtle motion information, especially from FME videos. Few works focus on the micro-expression generation task. This

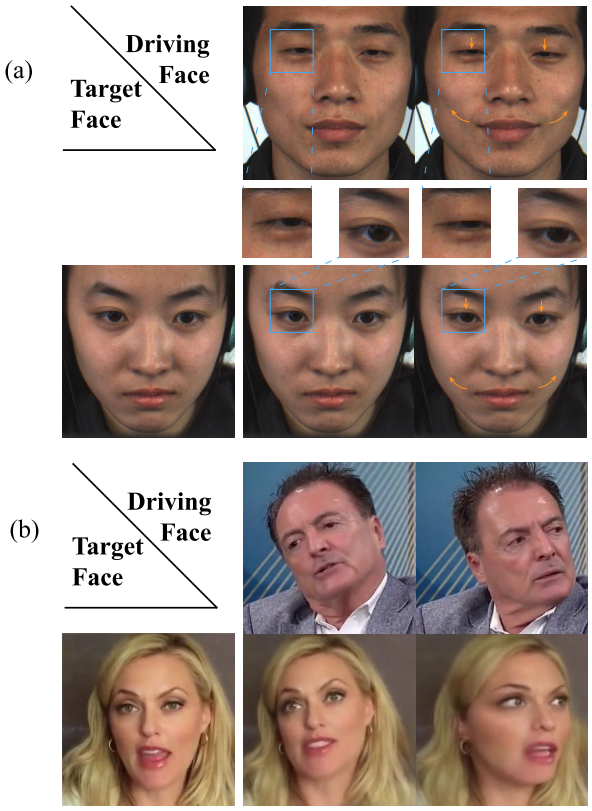


Fig. 1. Comparison between the FME generation task and traditional image animation task. Both tasks aim to transfer the motion information of a driving video to a target frame. However, current image animation methods in the human face domain focus on the macro movement (b), while the FMEs usually occur in local areas with subtle movement (a). The low intensity of the features mean that the motion translation for FMEs is more difficult than that for the normal image animation task. The enlargement of existing datasets using FME generation could benefit the community in various ways, such as improving the performance of FME recognition neural networks.

paper takes advantage of facial prior knowledge to alleviate optimization of the whole self-supervised paradigm for the FME generation task. Our contributions are threefold.

- We introduce image animation to the micro-expression generation task. With the help of the image animation paradigm, we use existing micro-expression datasets to transfer the action information of micro-expressions that are difficult to collect to new target faces and then generate new micro-expression video data. At present, there are few works on this type of task.
- We design a "plug-and-play", learnable facial prior module. This module encodes facial position features related to human emotions and can adaptively learn the features of micro-expression directly through the loss function of image animation. Using this module, we can learn and encode subtle and short-duration micro-expression action features in the human face and generally improve the effectiveness of image animation methods in the micro-expression generation task. This process helps to further improve the method for encoding micro-expression prior knowledge and adaptively using the latest image animation frameworks in the micro-expression generation task.
- We evaluate our proposed method on a cross-dataset benchmark using different scenarios and achieve state-of-the-art performance on this task. Furthermore, we verify the general improvement of the facial prior module for image animation in micro-expression generation in the experiment. Moreover, we use the generation technique to enlarge existing FME datasets, thereby systematically improving the recognition performance of action recognition backbones for the FME recognition task.

Our basic FME generation framework achieved first prize in the MEGC2021 generation track; the competition report can be found in [15]. This paper is substantially expanded in four aspects:

- We extend the facial prior module into a learnable encoding module. This module extracts the position features related to action units in the face and adaptively inputs the position features into the network in the form of keypoints and feature maps. This compensation explicitly provides the position information required in the image animation framework, so the model pays more attention to the motion information in the samples, thereby enhancing the model's attention and transferability for subtle movements. On this basis, the learning module can fit and learn the spatial features and action features in a dataset by relying on the objective function of image animation to further improve the effectiveness of the facial prior module in the micro-expression generation task.
- We verify the generality of the idea of introducing facial prior. With the plug-and-play design of the facial prior module, we have consistently improved the performance of the generative models, including FOMM [13] and MRAA [14]. We succinctly and effectively demonstrate the universality of the core idea by improving vanilla generative models with widely used facial information. We emphasize that this generality will facilitate follow-up researchers to modularly design the prior knowledge they focus on and use more dedicated generative models.
- As a supplement to expensive and inconvenient expert evaluation, we propose two strategies to further evaluate the generation performance of different models in a cost-effective manner. In public evaluation, volunteers are recruited to choose the preferred model in paired videos. In automatic evaluation, the qualities of the generated samples are compared in terms of the recognition results predicted by different models.
- We use the generation technique to enlarge the FME dataset, improving the performance of action recognition backbones in FME recognition. The improvement in various 3D neural networks indicates the effectiveness of using generated data to boost the performance of data-driven methods on the FMER task.

The rest of this paper is organized as follows. Section II reviews previous work on micro-expression analysis and image animation. Section III details the facial prior that we introduce to the image animation framework and how we measure the prior knowledge mathematically. Section IV introduces the experiments and analyzes the results. Section V discusses possible future research directions based on micro-expression generation. Finally, Section VI concludes and summarizes various observations.

II. Related Work

A. Micro-Expression Analysis

Traditional micro-expression analysis, including micro-expression recognition and micro-expression spotting, has attracted increasing attention. Micro-expression recognition aims to assign an emotional label to a well-segmented micro-expression sequence. The first public challenge on micro-expression recognition was held in 2018 [16]. Handcrafted features, including LBP-TOP, HOOF and 3DHOG, are used in baseline methods [17]. Various methods [18], [19], [20], [21], [22], [23] have studied and experimented on micro-expression recognition. Experiments based on single datasets [24] and cross datasets [25], [26], [27] are used for evaluation. Micro-expression spotting aims to detect the onset and offset frames of a long micro-expression sequence. The first public challenge on micro-expression spotting was held in 2019 [28]. Various methods [29], [30], [31], [32] have been proposed to spot FMEs from spontaneous videos. The latest public challenge on micro-expressions [33] involved a new challenge of spotting both macro-expressions and micro-expressions from long videos [32], [34], [35], [36].

Large-scale training data and specific facial features benefit micro-expression analysis. The performance of previous micro-expression methods is limited by the scale of public micro-expression datasets. Some institutions began to construct datasets by conducting trials to collect micro-expressions from a wide range of subjects. The datasets include the CASME dataset [37], CASME II dataset [1], CAS(ME)^2 dataset [11], CAS(ME)^3 dataset [38], SAMM dataset [8] and SMIC dataset [9].

The facial action coding system (FACS) [39] is a comprehensive, anatomically based system for describing all visually discernible facial movements. FACS breaks down facial expressions into individual muscle movements called action units (AUs) [40]. AUs have been widely used in facial expression research, especially in the macro-expression domain [41], [42]. By contrast, there is limited research on AU detection for micro-expressions [43], [44]. AU face regions are studied [45], [46], [47] to detect subtle micro-expression movement for recognition. Samples from most existing datasets are labeled with the emotion category and activated AU information, which we can leverage. Another feature of existing datasets is that the subjects are seated in front of a monitor to record the full-frontal faces of the participants. Moreover, the collected raw videos are carefully processed by professionals to rule out irrelevant facial movements and reserve ME-related motions. Considering these features of existing datasets, we introduce another useful tool called Dlib [48]. Dlib is a library widely used for face detection, facial keypoint prediction and face embedding [49], [50]. With 68 landmarks predicted by Dlib [48], faces can be accurately detected and aligned. The strict preprocessing of dataset construction [37] ensures that Dlib is sufficiently robust as a feature extractor for the ME analysis in our method [51]. In recent years, research based on deep learning methods has taken high-level features as input, such as optical flow [52], while others prefer to feed their networks unprocessed videos to automatically extract facial features [26].

However, due to the expensive labeling issue, the micro-expression recognition task remains difficult due to limited data. Facial micro-expression recognition (FMER) is essentially an action recognition task that is also considered a difficult task in micro-expression analysis. Before deep learning methods were introduced, LBP-TOP [53] and optical flow [29] were applied to this task. Shallow networks have also been introduced to combat the limited data [54], [55], [56]. Several articles [57], [58], [59], [60], [61] apply specifically designed deep neural networks to the FMER task. There are also other deep learning techniques being introduced to the task, such as adversarial training [12] and neural architecture search (NAS) [62]. A recent study [7] reports that with additional training data, the performance of general networks could match that of networks designed specifically for FMER. In our work, we propose to use an improved image animation technique to generate novel micro-expression data, thereby improving the performance of general action recognition backbones [63] on the FMER task.

B. Image Animation

Image animation aims to take a source image and a driving video to generate a video following the motion of the driving video while preserving the identity of the source image. In the facial image animation task, strong prior knowledge must be considered. Several methods [64], [65], [66], [67] extract facial features such as landmarks and boundaries before face synthesis or animation. Our work is similar since we use dlib [48] and AUs [40] to predict landmarks as prior knowledge of micro-expressions. Other image animation methods do not require strong prior knowledge as explicit keypoints or landmark labels in a particular domain since they use self-supervised or unsupervised learning methods to extract features. Some research [68], [69] focuses on extracting object landmarks to represent structural motion in an unsupervised manner. Then, with the representation of critical object information, an image can be animated based on the structure of keypoints or landmarks. Bansal et al. [70] proposed Recycle-GAN as an unsupervised data-driven approach for video retargeting. Wiles et al. [71] proposed X2Face training with a large collection of data to control face generation in a fully self-supervised manner. Siarohin et al. [72] proposed Monkey-Net in an unsupervised way trained to extract object keypoints and animate images. Similar work includes FOMM [13] and MRAA [14], which also utilize self-supervised methods to animate images by keypoints extracted from different domains of data. Since micro-expressions have unique characteristics such as short duration and low density, self-supervised methods can hardly obtain related motion features and semantic information. Hence, we introduce dlib and AUs to acquire facial landmarks to locate the occurring area and represent the intensity of the micro-expression as strong prior knowledge.

III. Proposed Method

Our method applies facial prior knowledge to enhance the performance of FME generation. First, we design the facial prior module to represent AUs with facial landmarks as the

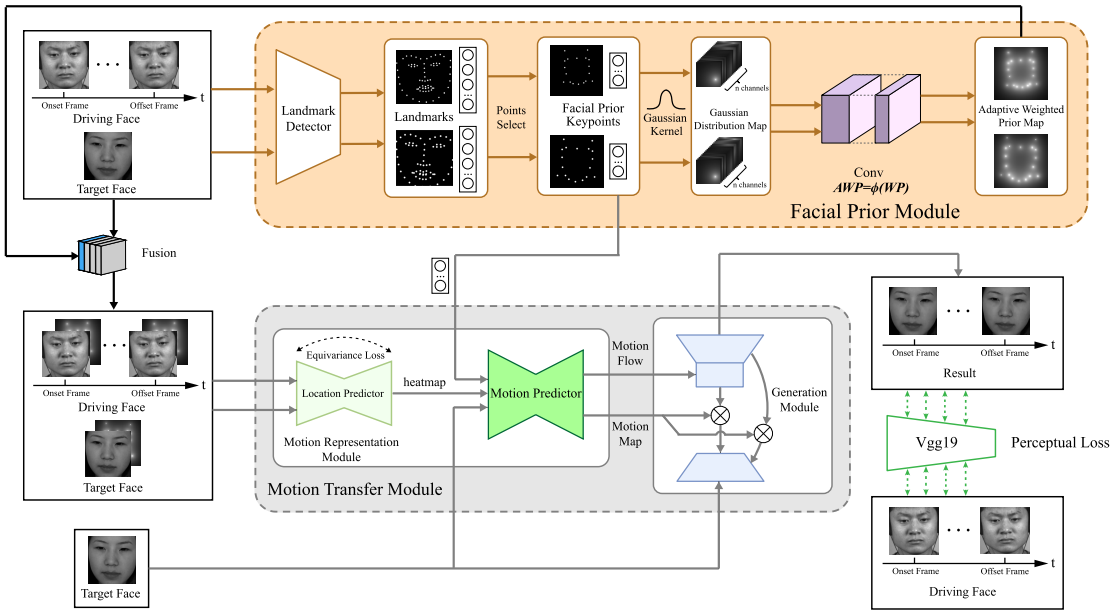


Fig. 2. Illustration of our framework that takes target faces and driving facial micro-expression (FME) videos as inputs and can generate novel FME videos. Our framework contains two main modules, namely, the facial prior module and motion transfer module. We use the image animation method to model the motion transfer problem; therefore, the motion transfer module can be further split into the motion representation module and generation module. In the facial prior module, facial prior keypoints are calculated based on facial landmarks, which can be extracted by a landmark detector. To alleviate the estimation error of facial prior keypoints, we formulate an n-channel Gaussian distribution map with a Gaussian kernel. To enhance the flexibility of localizing facial subregions, we modify the Gaussian distribution map as an adaptive weighted prior (AWP) map using a convolution layer. With the AWP map, the facial prior module can become learnable through the loss function. The target face and each face in the driving video are fused with the AWP map to serve as the extended version of the inputs for the motion representation module. As the AWP map directly encodes facial keypoint information, the extended version of inputs will facilitate the location predictor in the motion representation module to locate keypoints, which leads the model to pay more attention to facial movement instead of keypoint locations. By combining the heatmap output of the location predictor, the target face and facial prior keypoints from our facial prior module, the motion predictor in the motion representation module can generate the motion flow and motion map, which are then used as inputs for the generation module to drive the target face to a novel FME video sequence. Perceptual loss is used to minimize differences between generated and driving videos. Equivariance loss is also used to constrain the location predictor.

facial prior. Second, the extracted prior assists deep motion retargeting by indicating where micro-expressions appear. Finally, the generation module generates photorealistic FME frames by considering motion information with facial prior knowledge. Fig. 2 shows the architecture of FME generation used in this paper, in which the motion representation module and generation module are ideal and can be concretized by different algorithms.

A. FME Generation With Image Animation

We model the motion transfer process using image animation, which aims to generate novel videos. It tackles the problem of transferring motion information from a video to an image [72]. Image animation can be used in FME generation to generate novel FMEs. Generally, image animation can be divided into two subtasks, namely, motion prediction and object reconstruction. The motion prediction module extracts motion information from ordered sequences [70] such as videos. Then, the object reconstruction module applies deep learning methods such as generative adversarial networks (GANs) [73] or variant autoencoders [74] to reconstruct images. In recent years, state-of-the-art image animation methods strive to improve the generalizability of the framework [13], [14], [72], animating various categories of motion with appropriate training data. This is achieved by extracting motion information via self-supervised learning methods, which guarantees the frameworks can learn without prior knowledge.

However, self-supervised learning does not consider semantic information from a specific domain. Features of FMEs make mainstream animation frameworks often fail to work. Experiments also indicate that in FME generation, it is difficult for most self-supervised approaches to learn the motion representation of FMEs [15], [75]. Intuitively, the ability of humans to spot FMEs can be significantly improved if prior knowledge of FMEs is provided [76]. Prior knowledge of FMEs gained in psychology research, namely, the facial prior, can be considered as compensation for self-supervised learning. Our proposed method is motivated by this observation.

B. Facial Prior

Facial keypoints and the facial prior map indicate facial areas where FMEs might occur according to FACS. The basic idea is to utilize the facial prior of each frame instead of using a self-supervised approach. The ideal facial prior is expected to provide a map from micro-expressions to facial landmark locations.

1) Facial Keypoints: Facial keypoints indicate the AU-related areas in MEs. To locate the AU-related area, we use facial landmarks to estimate the centers of the activated AU region of MEs. We introduce facial keypoints to locate the areas of the activated AUs and represent the semantic

information of each group area. We distribute keypoints into groups, with each keypoint located and controlled one-to-one in the group area. Since not all keypoints located in the facial landmarks appear at a very close distance from facial landmarks and can be calculated by them specifically, keypoints can be formulated as a linear combination of landmarks [77].

As mentioned above, not all actions occur in the facial landmarks, which are sometimes difficult to locate by machine learning models due to the lack of edge information. Since facial landmarks are specific to each individual, we can calculate certain keypoints with facial landmarks. Because the distance between the camera and the person's face causes variation in the landmark distance at the pixel level, a reference must be employed to diminish variability between individuals. We introduce the distance between the inner corner point of each eye as a reference, called eye-distance. We first use a landmark detector to locate n facial landmark points. With eye-distance as a reference, we then calculate the keypoint locations by means of the linear combination of different landmarks, which differ from face to face, through the facial landmark detector. Given n facial landmarks predicted by a landmark detector with the coordinates (x_p, y_p) of every single landmark m, keypoint p can be formulated as:

$$\begin{bmatrix} x_p \\ y_p \end{bmatrix} = \sum_{i=1}^{n} k_i \begin{bmatrix} x_i \\ y_i \end{bmatrix} \tag{1}$$

where (x_p, y_p) is the coordinate of keypoint p and k_i denotes the weight factor of the contribution of landmark i to p. Detailed information, such as the number of keypoints and the centers of selected AUs, will be specified in Section *IV-E*.

2) Adaptive Weighted Prior Map: Due to possible estimation errors, we introduce an adaptive weighted prior map into the facial prior to normalize the motion information from all estimated keypoints and assess the uncertainty arising from estimates. It is reasonable to assess the importance of different areas according to their distance to the chosen keypoints [15]. To evaluate the importance mathematically, we formulate the per-pixel uncertainty associated with the estimated keypoints. For a chosen keypoint p with coordinates (x_p, y_p) and a certain pixel i in an image with coordinates (x_i, y_i), we calculate the Euclidean distance between the two points. Assuming that this distance-related uncertainty attributed to one keypoint has a Gaussian distribution over the entire image, we give the formulation for calculating the correlation of pixel i with keypoint p:

$$c_{ip} = e^{-\frac{(x_i - x_p)^2 + (y_i - y_p)^2}{\sigma^2}} \tag{2}$$

where σ is the variance of the assumed Gaussian distribution, which is set to 0.01 manually by experience [13]. For each keypoint, we can calculate the uncertainty distribution over the whole image as $S_p(x_i, y_i) = \sum_{i \in \mathbb{R}^{H \times W}} c_{ip}$.

In [15], the uncertainty distribution is normalized simply by summation $S = \frac{1}{m} \sum_{p=1}^{m} S_p$, which we refer to as the equal weighted prior (EWP) map. Adding all the uncertainty distribution maps might be mathematically concise, but this method lacks interpretability, and the model effect might be limited due to the hasty manual settings of the summation method. Therefore, a convolutional neural network is used to normalize the distance-related uncertainty, which can be formulated as:

$$S = \Psi_w(S^m) \tag{3}$$

in which w is the weight parameter of the neural network and S^m is the result of the concatenation of contribution maps from m keypoints, which has m channels. $\Psi(\cdot)$ is the applied activation function. Since the parameters are optimized according to the training data, the network takes the motion information from actual videos into consideration, thereby alleviating error from the estimation. The result of this normalization of the distance-related uncertainty is our proposed adaptive weighted prior (AWP) map.

Our proposed facial prior module is summarized in Algorithm 1. Sets of points calculated according to prior knowledge help the motion representation framework proceed with image animation and avoid the shortcomings of self-supervised learning. We believe this paradigm can help professionals make better use of different kinds of motion representation models and improve the quality of generated FMEs by providing a new perspective using the facial prior.

Algorithm 1 Facial Prior Module

Input: Facial micro-expression video frames $\left\{F^{(i)} \in \mathbb{R}^{C \times W \times H}\right\}_{(i=1)}^{N}$

Output: Facial keypoints $\boldsymbol{P} = \left\{P^{(i)} \in \mathbb{R}^{1 \times 2}\right\}_{(i=1)}^{M}$, FME video frames with facial prior $F_{prior} = \left\{F_{prior}^{(i)} \in \mathbb{R}^{(C+1) \times W \times H}\right\}$

Initialize: Set the facial keypoints matrices $\boldsymbol{P}$, facial landmarks matrices $\boldsymbol{L} = \left\{L^{(1)}, L^{(2)}, ..., L^{(N)}\right\}_{(i=1)}^{N}$ to zero matrices, $\boldsymbol{F} = \left\{F^{(1)}, F^{(2)}, ..., F^{(N)}\right\}_{(i=1)}^{N}$, relationship matrix between landmarks and keypoints $\boldsymbol{R} = \left\{R^{(1)}, R^{(2)}, ..., R^{(M)}\right\}$ according to facial prior knowledge. Set $\mathcal{D}(\cdot)$ as facial landmark detector, $\mathcal{G}(\mu, \sigma)$ as Gaussian kernel with μ as mean and σ as variance, $\Psi_w(\cdot)$ as the normalization function with parameters w.

function FACIAL KEYPOINTS($\boldsymbol{F}$)
  for $i \in [1, N]$ **do**
    $L^{(i)} = \mathcal{D}[F^{(i)}]$
    for $k \in [1, M]$ **do** ▷ M facial keypoints
      $P_k^{(i)}(x_k^i, y_k^i) = R^{(k)} \cdot L^{(i)}$ ▷ Refer to Eq. (1)
    end for
  end for
end function

function ADAPTIVE WEIGHTED PRIOR MAP($\boldsymbol{F}$, $\boldsymbol{P}$)
  for $i \in [1, N]$ **do**
    for $p \in [1, M]$ **do**
      $S_p^{(i)M} = \mathcal{G}(P_p^{(i)}, \sigma)$
    end for
    $S^{(i)} = \Psi_w(S^{(i)M})$
    $F_{prior}^{(i)} = \text{concat}\ (F^{(i)}, S^{(i)})$
  end for
end function

C. Implementation and Optimization

In this subsection, we illustrate our approach to combining the facial prior with state-of-the-art techniques of image

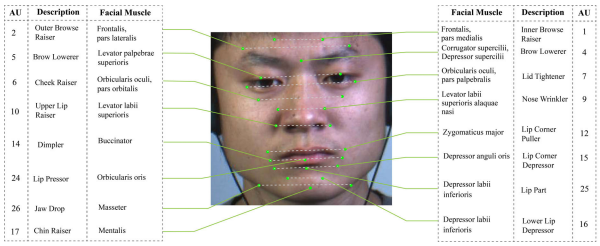


Fig. 3. Relationship of facial muscles, action description and AU. The AU area can be integrated with the correlation of branch muscle movement and positional proximity. For example, the areas of AU1 and AU2 can be integrated into one group since they are controlled by the same motor branch of *Frontalis*.

animation by introducing facial prior into two motion representation frameworks that have yielded excellent results in macro movement generation tasks.

1) Facial Prior: To encode the facial prior into the image animation framework, we investigate the corresponding knowledge from ME research. Specifically, we study the spatial location feature of AUs, which is directly related to the selection of keypoint locations. For starters, we study the relationship between facial muscles and AUs. Facial muscles have significant spatial location information and provide a reference for choosing keypoints. Then, we utilize a Voronoi diagram to assist in choosing the locations of keypoints.

Facial muscle movement is an essential cue to facial expressions and is strongly related to AUs [78]. Fig. 3 shows the AUs and their facial positions along with the related muscles. Micro-expressions can be roughly located with the position of AU-related areas. Since an AU can be formulated with one or more muscle movements, we distribute the AU-related areas based on the facial muscle orientation. The area with the same branch muscle controlled can be integrated.

To integrate spatially neighboring areas of facial muscles or AUs, we use a Voronoi diagram to divide the facial area into several parts according to facial keypoints. In a Voronoi diagram, every point in a given polygon region is closer to its keypoint than to any other. By choosing appropriate AU-oriented keypoint locations, each area of the Voronoi diagram conforms to the distribution of AUs on the human face.

In summary, we investigate the spatial location feature of AUs through facial muscles and introduce a Voronoi diagram to assist in choosing facial keypoints. In some mainstream FME databases [1], [37], [79], FME video samples are provided along with AU labels, which would also help us make better use of AU patterns in the generation process. The selection of keypoints and an illustration of the Voronoi diagram are detailed in Section *IV-B*.

2) FOMM: The first-order motion model [13] has two modules: *motion estimation* and *image generation*. The first step in the motion estimation module is to estimate coarse motion from the target frame to the driving frame. In the motion estimation module, the self-supervised learning method is adopted for application to diverse scenarios involving macro movements. With the dense motion field and occlusion mask eventually extracted in the motion estimation module, the image generation module renders an image of the target object moving, such as the one in the driving video.

We have made the following modifications to FOMM. The keypoint representation acts as a bottleneck leading to a compact motion representation, as Siarohin et al. argued [13]. In vanilla FOMM, an encoder–decoder network learned in a self-supervised manner serves as a keypoint detector to predict K keypoints, indicating K critical parts for motion (K is set manually). However, Fan et al. [75] found that the keypoints predicted by vanilla FOMM are heavily overlapping and argued that subtle variations in FMEs make it difficult for an unsupervised learner to learn the keypoints well. In our proposed adaptive weighted prior-based first-order motion model, the unsatisfying detected keypoints are replaced with facial prior keypoints, which contain more prior knowledge and semantic information. Moreover, our proposed facial prior map is fused by concatenating each frame as an additional channel, providing strong distance-related importance knowledge.

The facial prior map calculates the local importance of the human face. Based on it, our previous work [15] improved the performance of the motion representation framework. However, one limitation is that the contributions of different keypoints are equally weighed by the FPM. To address this problem, an adaptive weighted method is adopted, using a convolutional layer with learnable weights that can be optimized with the training loss function. Technically, the learnable convolutional layer can weigh each facial prior keypoint differently to approximate the prior distribution in a given dataset.

3) Modified MRAA: Compared to keypoints in FOMM, MRAA [14] focuses on semantically relevant regions and tries to disentangle shape and pose in the region space.

Research [14] shows that MRAA surpassed FOMM in many macro movement image animation tasks. However, for FME generation, MRAA has drawbacks compared with FOMM. First, MRAA focuses on regions to extract the motion information more easily; however, for FMEs, local motion is more important yet difficult to extract from driving videos. Second, the disentanglement of shape and pose might make the overall training of the framework uncoordinated, resulting in unsatisfactory facial distortion.

The adaptive facial prior can provide semantic information that could be fed into the region-prediction module of MRAA. Therefore, in our modified version of MRAA, the adaptive weighted prior map is taken as the provided region of interest, similar to our modifications to FOMM.

4) Training Loss: Our proposed frameworks are trained and optimized using a multiresolution perceptual loss and an equivariance constraint loss, which are used in both vanilla FOMM and MRAA.

Perceptual loss [80] has been applied to various generation tasks [13], [14], [81], [82]. It compares the high-level features extracted from two domains. Assume X as the input face and $\hat{X}$ as the reconstructed result of input; then, the perceptual loss L_P can be defined as:

$$L_P = \sum_l \alpha_l \left|\left| E_l(X) - E_l\left(\hat{X}\right)\right|\right| \tag{4}$$

where $\alpha_l > 0, l = 1, 2, 3, \ldots$ are scalars. $E_l(\cdot)$ is the feature extracted by VGG-19 pretrained for ImageNet classification.

Recently, equivariance constraint loss has been reported to generate new image animation videos [13], [14]. The equivariance constraint introduced in [68] and [69] limits the movement of keypoints. In [13], it was extended to additionally include a constraint on the predicted Jacobians from the motion estimator in the pipeline. The equivariance constraint loss is used following the structure of FOMM [13] and MRAA [14] and can be defined as

$$L_E = |A^k_{X\leftarrow R} - \tilde{A}A^k_{\tilde{X}\leftarrow R}| \tag{5}$$

where $\tilde{X}$ represents image X transformed by $\tilde{A}$ and $\tilde{A}$ is a geometric transform matrix. The perceptual loss and equivariance constraint loss have equal loss weights in the experiments [13].

For the adaptive weighted facial prior-based first-order motion model, a large number of facial video sequences including various micro-expressions are used for training the whole network. Each frame in the videos has its own corresponding keypoints and FPM, which can be generated before training. By combining content information from the target image with motion information from the original driving video, as well as their keypoints and FPM, we obtain the corresponding reconstructed image. With perceptual loss calculated between the input driving face and the generated face, the proposed network can be optimized in an effective manner. Additionally, the equivariance constraint loss mentioned above is introduced into Jacobians in the location prediction module so that the model can predict consistent Jacobians with respect to given keypoints. To this end, the final loss is the weighted sum of these two losses. During testing, this framework generates different FMEs on the given target face image according to different input driving videos.

For modified MRAA, similar losses are applied as those used in vanilla MRAA. Since we have provided the region information for the framework, the original two-stage training process is simplified to one stage, and the same scenario as that in modified FOMM is employed.

IV. Experiment

In this section, we conduct comprehensive experiments to verify the effectiveness of the proposed algorithm in the generation of FMEs. First, we introduce the FME datasets used in our experiment, along with our implementation. Second, we present the generation results and visualize the different frames to demonstrate the improvement achieved when using the facial prior module. Third, we assess the effectiveness of our proposed facial prior module under two scenarios: 1) We measure the effectiveness of the equal weighted prior module using expert evaluation. In this section, we compare the proposed module with all current methods. 2) We measure the effectiveness of our proposed adaptive weighted prior module through public evaluation. In this section, we perform a significance test of the experimental data to support the conclusions drawn from this experiment. Moreover, we compare the recognition results inferred by a set of recognition models in the automatic evaluation strategy. Finally, we conduct a recognition experiment with action recognition backbones to explore the possibility of using generated data to improve recognition performance. The detailed settings are described in each experiment.

TABLE I

Details of CASME II, SAMM and SMIC-HS Datasets. Some Statistics Are From the Literatures [1], [8], and [9], [55]

	CASME II [1]	SAMM [8]	SMIC-HS [9]
Subjects	24	28	16
Videos	145	133	164
Raw Resolution (pixels)	640×480	960×650	640×480
Cropped Resolution (pixels)	256×256	256×256	256×256
Frame Rate (fps)	200	200	100
Average Number of Frames	70	73	34
Average Video Duration (s)	0.35	0.36	0.34
Micro-Movements	247	159	164
Video Type	Color	Grayscale	Color
Spontaneous/Posed	Spontaneous	Spontaneous	Spontaneous
FACS Coded	Yes	Yes	No
Emotion Inducer	Video Episodes	Tailored Video Stimuli	Video Clips
Ethnicities	1	13	3
Emotion Classes	5	7	3

TABLE II

Selected Driving Videos for Evaluation

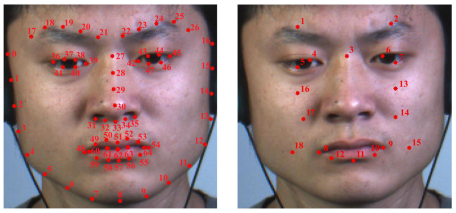
	CASME II [1]	SAMM [8]	SMIC-HS [9]
Positive	EP01_01f	022_3_3	s3_po_05
Negative	EP19_06f	018_3_1	s11_ne_02
Surprise	EP01_13	007_7_1	s20_sur_01

A. Datasets

Our method is trained using three public facial micro-expression datasets: CASME II [1], SAMM [8], and SMIC-HS [9]. Models were trained on the combined dataset, excluding the videos chosen for evaluation. Details of these three datasets are presented in Table I. Note that the sample videos of these datasets have labels of micro-expressions. In CASME II [1], sample videos are classified into five categories of micro-expressions: happiness, disgust, repression, surprise, others. In SAMM [8], sample videos are labeled with seven categories of micro-expressions: contempt, disgust, fear, anger, sadness, happiness, surprise. In SMIC-HS [9], micro-expression clips are classified into three categories: positive, negative, surprise.

For evaluation, our framework generates FMEs from driving sample videos on the template target face. We evaluate our proposed AWP on the cross-database MEGC2021 Generation Challenge benchmark. Specifically, we use nine driving samples and two template faces to generate eighteen videos for evaluation. Selected driving videos are chosen from CASME II [1], SAMM [8] and SMIC-HS [9], which are listed in Table II. The template faces are chosen from CASME I [37] and SMIC-VIS [9] to avoid possible information leakage. To maintain the same emotion categories of generated videos, we refer to the dataset with the simplest categories among the three datasets, which is SMIC-HS [9]. The mentioned categories (“positive”,“negative”,“surprise”) are from the configuration of this dataset.

In summary, the generative models will be trained on the appointed micro-expression datasets. At the same time, the face samples used to be driven will not appear in the training set, which could provide justification for the generalization of the model and prove the effectiveness of the model on unknown neutral faces.



(a) (b)

Fig. 4. (a) The locations of 68 landmarks. (b) The locations of 18 keypoints.

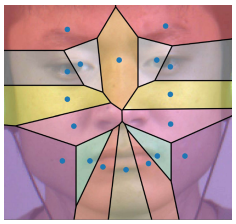


Fig. 5. Control area of each keypoint with colors associated with same AUs.

B. Implementation

We use dlib [48] to locate 68 facial landmark points, as shown in Fig. 4(a). We evaluate the importance of each landmark and use AU to represent the occurrence of FMEs. Eighteen points are selected to locate the micro-expressions, as shwon in color in Fig. 4(b). A total of 12 of 18 points are selected from 68 facial landmarks predicted with dlib, and the remaining 6 points are linear combinations of different landmarks. The points 1-12 selected from the 68 facial landmarks are 19, 24, 27, 38, 41, 43, 46, 48, 54, 55, 57 and 59. With eye-distance as a reference, points 13 to 18 are calculated from 68 facial landmarks and are formulated as:

$$\begin{bmatrix} p_{13} \\ p_{14} \\ p_{15} \\ p_{16} \\ p_{17} \\ p_{18} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} p_{42} \\ p_{42} \\ p_{42} \\ p_{39} \\ p_{39} \\ p_{39} \end{bmatrix} - \frac{1}{2}\begin{bmatrix} p_{39} \\ p_{39} \\ p_{39} \\ p_{42} \\ p_{42} \\ p_{42} \end{bmatrix} + \begin{bmatrix} p_{29} \\ p_{35} \\ p_{54} \\ p_{29} \\ p_{35} \\ p_{54} \end{bmatrix} \tag{6}$$

where p_n denotes the coordinates (x_n, y_n) of point n. Eye-distance is employed with p_{39} and p_{42} as the inner corner point of each eye. Each point is a linear combination of 68 landmarks, with 3 landmarks contributing to the location of a single point. Fig. 5 shows the control region of 18 keypoints parted by the Voronoi diagram. AU areas are divided into regions, where each region controls one or multiple AUs.

We convert color video to grayscale video and then use face detection and facial area extraction as a preprocessing step. Each grayscale image from the video is cropped to a resolution of 256×256 pixels and self-concatenated to obtain 3 channel images. We convert all videos to grayscale to make all the samples in the same format before feeding them into the network since images in SAMM are all grayscale with 3 channels while the other two datasets are color ones. Details about the training process are as follows. Keypoints selected in our experiments are depicted in Fig. 4(b). Additionally, synthesized facial prior maps S_m were normalized before entering the motion prediction module. The training process was terminated after 5000 iterations.

C. Ablation Study

Ablation studies were conducted on two frameworks, FOMM and MRAA, to verify the effectiveness and robustness when introducing the facial prior into video generation frameworks. Fig. 6 presents the generation results and visualizations of the interframe difference between the current frame and the given target face.

1) FOMM vs. FOMM With Equal Weighted Prior Map vs. FOMM With Adaptive Weighted Prior Map: Considering only FOMM (II) and its modifications ($IV\&V$) in Fig.6, we find that the adaptive weighted method helps the motion representation module highlight facial motion and reduces noise. Examples include the right angulus oris shifting in cases 1&2, cheek raising in cases 3&4, and eye blinking in cases 5&6. With the adaptive weighted method, these movements dominate other subtle motions in the period in which they occur, which they would not have without this approach, as shown in the interframe-difference pseudo-colored visualizations. Additionally, the adaptive weighted method prevents premature and redundant movements before the apex frame, which is evident in cases 1−6. This means that generated FMEs arise and disappear in a smoother manner.

2) MRAA vs. MRAA With Adaptive Weighted Prior Map: Considering only MRAA (I) and its modification (III) in Fig.6, we can see that MRAA with adaptive weighted prior map better captures subtle motionr, while MRAA produces comparatively unsatisfying results, with many incorrect motions that do not convey the same emotion as the driving sequences do.

3) MRAA With Adaptive Weighted Prior Map vs. FOMM With Adaptive Weighted Prior Map: We compared the modifications of MRAA and FOMM that yielded the best results, that is, III and V, in Fig.6. Visually, FOMM with adaptive weighted prior map outperforms MRAA with adaptive weighted prior map since FOMM with AWP animates lifelike facial expressions even with motions that are difficult to learn, such as the right angulus oris shifting in cases 1&2.

D. Evaluation Metrics

Two evaluation scenarios, i.e., expert evaluation and public evaluation, are adopted to assess the performance of our approach.

Specifically in expert evaluation, each generated video is subjectively evaluated based on the quality and AUs by experts who are FACS certification [39] holders. The facial region is divided into upper and lower parts and evaluated separately with a score of 0-3. By separating the face into two parts, evaluations can take into account partial facial movements that may occur. Furthermore, the judges can provide a score of 0-3 in a category called *noise* to evaluate the overall quality of the generated video. The noise score decreases if the generated video has background artifacts. Therefore, the maximum available score for each video is 9.

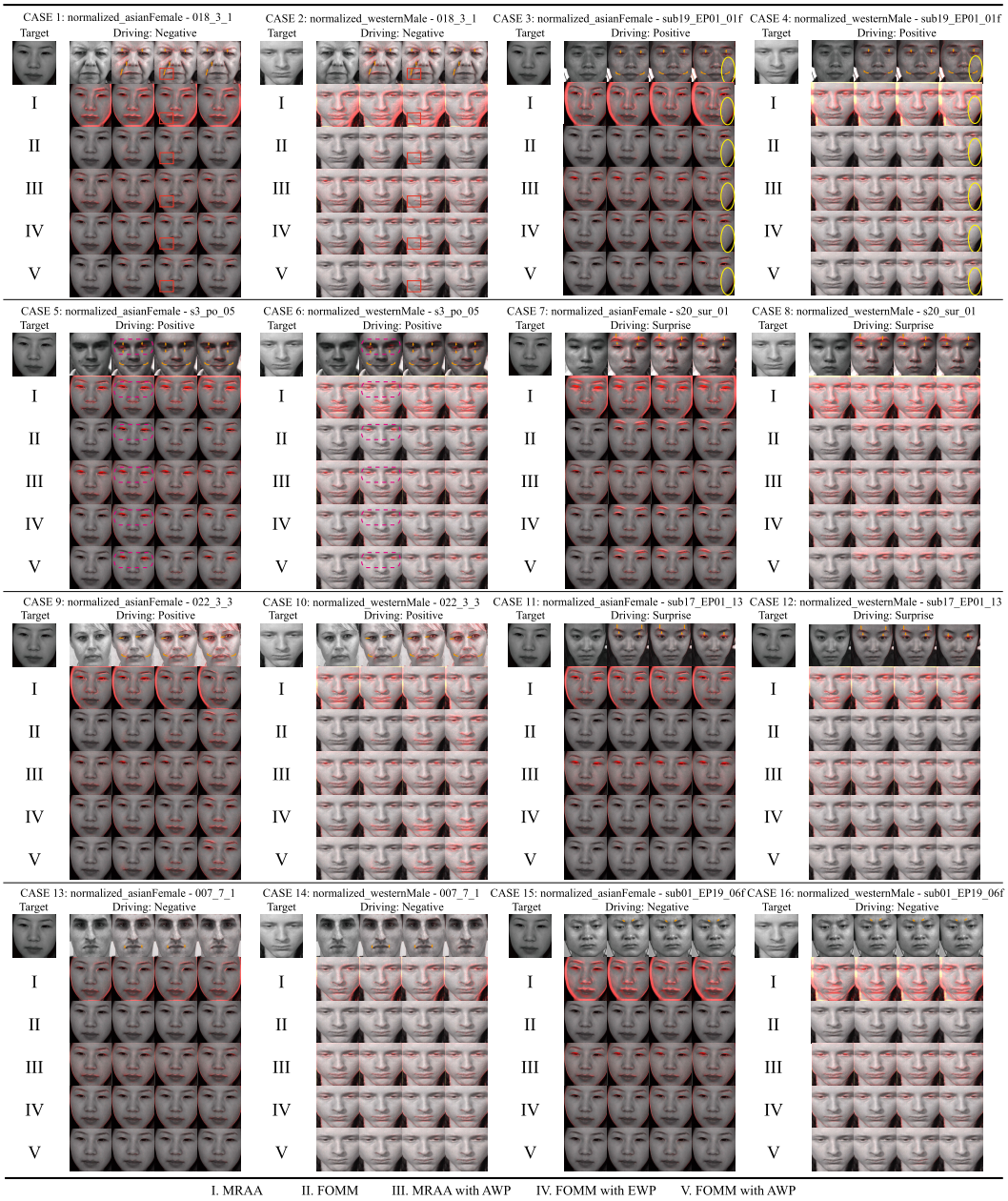


Fig. 6. Comparison of MRAA, FOMM, MRAA with adaptive weighted prior (AWP) map, FOMM with equal weighted prior (EWP) map and FOMM with adaptive weighted prior (AWP) map. In each case, the target face and driving sequence are presented in Row 1. The results are presented in Rows 2, 3, 4, 5 and 6. All the images are covered with pseudo colored frame difference between the current frame and the target face. For better visualization, we sort the presented cases by the noticeableness of the FME. Markers indicate some subtle movements, such as angulus oris shifting (red rectangle), eye blinking (magenta dotted capsule), cheek raising (yellow ellipse) and many others (orange arrow).

The following details the score categories:

1) Score 0: Completely Incorrect
2) Score 1: Poor
3) Score 2: Good
4) Score 3: Excellent

As for public evaluation, a total of 31 students were recruited as volunteers with payment (mean age 20.68, standard deviation(SD) = 1.35, including 20 males and 11 females). Given each compared video pair, the participants were required to choose the better of two given FMEs generated by two different models. The comparison pairs were formed according to the degree of improvement step by step (e.g., FOMM v.s. FOMM with EWP, FOMM with EWP v.s. FOMM with AWP). This design was used for three reasons: 1) the step-by-step comparison can illustrate the effectiveness of each step of improvement; 2) this design can compare different kinds of methods based on the transitivity of comparison; and 3) the step-by-step comparison reduces the size of comparisons for the untrained volunteers. Since the subjects are not professionally trained, grading ME video samples for too long might be tedious for them and affect their judgment. Therefore, we provided only a subset of the possible pairings. Furthermore, we presented the results of two models modified with the proposed AWP and asked them to choose

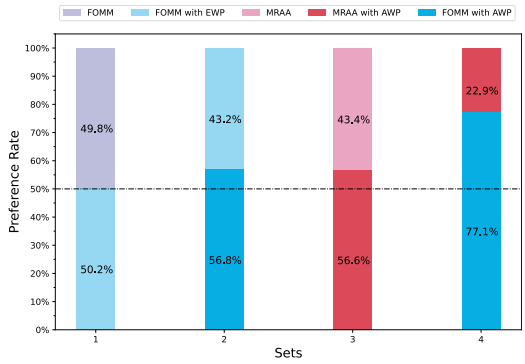


Fig. 7. Average voting rate results between pairwise methods using public evaluation. The voting rate represents the subject's confidence in the preference model in a set of public evaluation. The higher the voting rate is, the higher the subjects' confidence in the preference model. Since the first set of data failed the statistical test, we use slashes to denote it. Please note that *EWP* is the abbreviation of *equal weighted prior*, and *AWP* is the abbreviation of *adaptive weighted prior*. Best viewed in color.

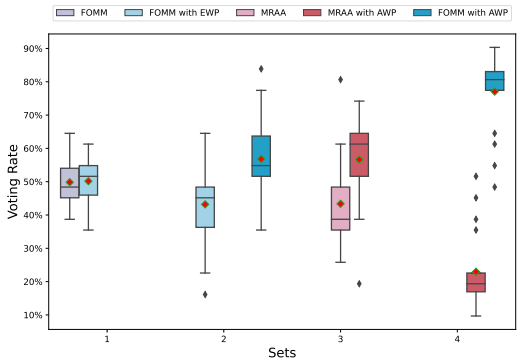


Fig. 8. Boxplot of the voting rate of each set of data in public evaluation. In the boxplot, the red dot denotes the mean and the line in the box denotes the median. There is a large gap in the mean values of the last three sets that passed the statistical test. There are also significant differences in the distributions of voting rate among different samples in each set. These results yield further conclusions. Please note that *EWP* is the abbreviation of *equal weighted prior*, and *AWP* is the abbreviation of *adaptive weighted prior*. Best viewed in color.

the best model to provide a reference for future research. That is, each participant graded 72 video pairs in total.

Through this test, the total vote count of each video is obtained for further analysis. Therefore, we propose two metrics to measure the vote result of public evaluation: preference rate and voting rate. Preference rate indicates the proportion of preferred samples in the comparison while voting rate represents the average confidence in the model during the comparison. Specifically, the preference rate of model A when compared with model B is formulated as:

$$\mathbf{PrefRate}_{(A|B)} = \frac{\sum_{i=1}^{M} \mathbf{1}_{(VC_i^{(A|B)} > VC_i^{(B|A)})}}{M}, \tag{7}$$

the voting rate of model A when compared with model B is formulated as:

$$\mathbf{VotingRate}_{(A|B)} = \frac{\sum_{i=1}^{M} VC_i^{(A|B)}}{M \times N}, \tag{8}$$

where $VC_i^{(A|B)}$ is the vote count of model A in sample i when compared with model B. $\mathbf{1}_{(VC_i^{(A|B)} > VC_i^{(B|A)})}$ is a value that changes based on the comparison result: it is equal to 1 when $(VC_i^{(A|B)} > VC_i^{(B|A)})$ and 0 otherwise. M is the total number of compared video pairs, and N is the total number of participants. When comparing models A and B in sample i, $VC_i^{(A|B)} + VC_i^{(B|A)} = N$. Furthermore, when N is odd, $\mathbf{PrefRate}_{(A|B)} + \mathbf{PrefRate}_{(B|A)} = 1$ and $\mathbf{VotingRate}_{(A|B)} + \mathbf{VotingRate}_{(B|A)} = 1$.

E. Expert Evaluation

The expert evaluation results in Table III are the challenge results from the MEGC2021 generation track. This is the first year the challenge was held. More detailed figures are available at https://megc2021.github.io/GeneResultevaluation.html.

As shown in Table III, the FOMM with our proposed EWP outperforms the other existing micro-expression generation methods and achieved first place in the MEGC2021 generation track.

Generally, expert evaluation can only provide evaluation results for the considered methods, and it is not convenient to conduct exploratory work, such as ablation experiments. Luckily, the DMT-FMEG [75] included in this challenge utilized vanilla FOMM and trained the model in the FME dataset. Since the expert evaluation scores are more authoritative and rare, this result provides a powerful research reference to further explore the effect of the facial prior module in improving state-of-the-art methods for image animation in micro-expression.

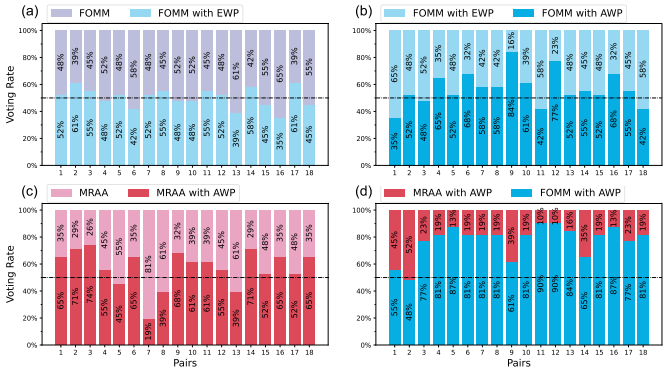


Fig. 9. Detailed comparison results between pairwise methods using public evaluation. Given a pair of FME videos generated by different methods, subjects were asked to vote for the one that seems more realistic. Voting rate indicates the sum of votes from all subjects for each testing sample. To ensure fair comparisons, we use the same samples as in the MEGC2021 Generation Challenge. The voting rates for all samples are shown: a higher score indicates better performance. *EWP* is the abbreviation of *equal weighted prior*, and *AWP* is the abbreviation of *adaptive weighted prior*. Best viewed in color.

F. Public Evaluation

The detailed results of the public evaluation are shown in Fig. 9. As mentioned above, we proposed two metrics to assess the experimental data. Due to the limited number of experimental samples in our public evaluation, sampling errors may be introduced, resulting in incorrect interpretation of the results. To this end, we assess the statistical significance of our results to confirm the conclusions.

To begin, we perform statistical modeling of our public evaluation. We conducted four sets of comparative experiments,

TABLE III

OVERALL EXPERT EVALUATION OF EXISTING STATE-OF-THE-ART METHODS. THE METHODS AND RESULTS ARE FROM THE FIRST FME GENERATION CHALLENGE IN MEGC2021

Methods	Overall	Normalized			
		Expert1	Expert2	Expert3	Overall
ID:3311	173	85/140	51/107	37/76	1.57062
FAMGAN [85]	236	104/140	66/107	66/76	2.228101
FOMM [76]	303	140/140	107/107	56/76	2.736842
FOMM with EWP	**316**	**139/140**	**101/107**	**76/76**	**2.936782**

TABLE IV

SIGNIFICANCE TEST RESULTS OF EACH SET OF PUBLIC EVALUATION. H_0 IS THE EXPECTED MEAN OF THE VOTES THAT IMPROVED METHOD M_a RECEIVES $\mu < 15.5$. IF THE TEST STATISTIC OF THE SET $Z > 1.65$, THEN WE REJECT HYPOTHESIS H_0 AND CONSIDER M_a TO BE BETTER IN THE PUBLIC EVALUATION, I.E., THIS SET PASSES THE SIGNIFICANCE TEST

	Set 1	Set 2	Set 3	Set 4
M_a	FOMM with EWP	FOMM with AWP	MRAA with AWP	FOMM with AWP
M_b	FOMM	FOMM with EWP	MRAA	MRAA with AWP
H_0	$\mu < 15.5$			
Z	0.11	**2.42**	**2.06**	**9.92**

each consisting of 31 participants choosing the better model among 18 pairwise videos according to the generation quality. We refer to the better model as M_a and the worse one as M_b. A pairwise sample comparing M_a and M_b receives 31 independent subject votes. Suppose the probability of voting for M_a for each participant in a sample is p, and the number of votes received by M_a in a pairwise video comparison is a random variable X; then, X follows a binomial distribution, that is, $X \sim B(31, p)$. According to the central limit theorem (CLT), when the sample number n in the binomial distribution is sufficiently large (greater than 20), a binomial distribution can be approximated using a normal distribution. Therefore, for $X \sim B(31, p)$, we can use the normal distribution $X \sim N(31p, 31p(1-p))$ as an approximation. That is, for a video pair generated by M_a and M_b, its 31 evaluations by the public can be regarded as following a normal distribution $X \sim N(31p, 31p(1-p))$. Then, the significance of the conclusions can be tested using a normal distribution significance test.

Since we assume that M_a is better than M_b in the evaluation, it is reasonable to set the null hypothesis H_0 as follows: fewer than $\frac{31}{2}$ participants will vote for M_a in a sample. Assuming X represents the number of votes that M_a receives and that its expected mean is μ, H_0 can be formulated as $H_0 : \mu < 15.5$. Moreover, the alternative hypothesis can be formulated as $H_1 : \mu \geq 15.5$. Since there are 18 pairs of videos generated by M_a and M_b being voted on by the public, the process can be abstracted as 18 sampling tests, and we can obtain statistics regarding the 18 samples of X. On the basis of this statistic, we can calculate the mean $\bar{X}$ and standard deviation σ of the random variable. Therefore, the test statistic Z can be calculated as follows:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

where $n = 18$ is the number of samples. Since the alternative hypothesis is $H_1 : \mu \geq 15.5$, a right-tailed test should be used. Assuming the commonly used significance level of 0.05, the test critical value is 1.65 according to the distribution function of the standard normal distribution. That is, if the test statistic $Z > 1.65$, then we can reject hypothesis H_0 and conclude that the results are statistically significant.

On the basis of the data collected via public evaluation, we conducted a significance test on four sets of method comparison experiments; the results are shown in Table IV, where sets 2, 3, and 4 have a test statistic that exceeds 1.65. Therefore, the following analysis draws conclusions from only these sets in the public evaluation, while the conclusion from set 1 is considered nonsignificant, as shown in Fig. 7.

To better illustrate the analysis of the public evaluation, we present boxplots of the collected data. Fig. 8 shows a large gap in the average voting rate of several groups of data that have passed the significance test, and there is a large gap in the voting rate distribution. In other words, the significance test can be used to explicitly preprocess the data collected from the public evaluation to ensure that the conclusions we draw are reliable.

G. Result Analysis

In this section, we summarize the experimental data collected from the expert evaluation and public evaluation described above. The conclusions verify the effectiveness of our method.

First, we introduce expert evaluation. Expert evaluation is an evaluation of existing methods by qualified experts according to preset conditions. Since MEGC2021 was the first competition, according to the expert evaluation provided in Table III, FOMM with EWP achieved the best results among all methods in the competition. Furthermore, another group of methods that participated in the competition, DMT-FMEG [75], used the vanilla FOMM model trained on micro-expression datasets. From the perspective of expert review, FOMM with EWP has better micro-expression generation performance than does the vanilla FOMM method. Overall, FOMM with EWP scored the highest in the first micro-expression generation competition and surpassed the FOMM in the expert evaluation.

Next, we present the public evaluation, which is convenient for conducting ablation experiments to measure the degree of model improvement. However, since these data come from

TABLE V

PREFERENCE RATE RESULT OF THE PUBLIC EVALUATION. THE PREFERENCE RATE INDICATES THE PROPORTION OF PREFERRED SAMPLES A METHOD GENERATED IN A PAIRWISE COMPARISON WITH ANOTHER METHOD. NOTE THAT SINCE THE FIRST SET OF PUBLIC EVALUATION DID NOT PASS THE SIGNIFICANCE TEST, WE DO NOT DRAW ANY CONCLUSIONS FROM IT

FOMM vs FOMM with EWP		FOMM with EWP vs FOMM with AWP		MRAA vs MRAA with AWP		MRAA vs MRAA with AWP	
FOMM	8/18	FOMM with EWP	4/18	MRAA	4/18	MRAA with AWP	1/18
FOMM with EWP	10/18	FOMM with AWP	**14/18**	MRAA with AWP	**14/18**	FOMM with AWP	**17/18**

public opinion, we must perform a statistical test on the data collected during this process. According to the first set of data in the public evaluation (FOMM vs. FOMM with EWP), the difference between these methods might not be considered significant, so we will not draw any conclusions. The remaining data groups passed the significance test, and we analyzed them in turn.

For the second set of public evaluation data (FOMM with EWP vs. FOMM with AWP), the FOMM with AWP method produced more preferred videos (i.e., a preference rate greater than 0.5 was obtained, as shown in Table V). Furthermore, public subjects were, on average, more than half as confident (i.e., more than half the voting rate, as shown in Fig. 7) in this approach compared to FOMM with EWP. Therefore, FOMM with AWP outperforms FOMM with EWP in the FME generation task. Furthermore, introducing a learning layer in the facial prior module is beneficial to encoding the prior knowledge of the human face, thereby improving the performance of the model on the FME generation task.

Similarly, we use the facial prior module to improve MRAA. Through the preference rate and voting rate indicators of the public evaluation, we can see that the MRAA with AWP generate more preferred videos for public subjects and leads to higher model confidence, which means that AWP improves MRAA's performance on FME generation tasks. The improvement effects of the MRAA with AWP and FOMM with AWP models indicate that it is feasible to introduce a facial prior module into the image animation framework to encode the facial prior, thereby improving the generation effect of FMEs. To provide a reference for follow-up research, we conducted the public evaluation experiment on these two improved models. The results show that the improved FOMM with AWP model is better for generating micro-expressions.

Through the above analysis of the experimental data of the expert evaluation and public evaluation, we can see that the introduction of a learnable facial prior module generally improves the effectiveness of image animation methods in micro-expression generation. Specifically, in the expert evaluation, the FOMM with EWP outperforms the vanilla FOMM.The second set of data of the public evaluation illustrates introducing a learnable module makes the FOMM with AWP superior to the EWP method, which means that it also outperforms the vanilla FOMM method. Moreover, the public evaluation indicates that the MRAA with the learnable AWP outperforms the vanilla MRAA method.

Furthermore, by comparing FOMM to FOMM with EWP through expert evaluation and public evaluation, we can see that the evaluation process still requires expert input when the improvement effect is relatively subtle and specialized. On the other hand, there are more pronounced gaps in the expert evaluation when the public perceives certain effects to be significant. Therefore, we recommend statistical tests after each public evaluation to ensure the significance of the conclusions.

In summary, the two parts of human evaluation, including expert evaluation and public evaluation, conjointly prove the superior effectiveness of the proposed method to perform FME transfer to an unknown neutral face.

H. Automatic Evaluation

Both expert evaluation and public evaluation require the involvement of humans, which could be time-consuming and lack objectivity; an automatic evaluation strategy could effectively address these shortcomings.

Here, we evaluate the generation quality by comparing the recognition results for real samples and generated samples. This specific scenario applies a performance requirement to the recognition method. Notably, a recent study [7] showed that general action recognition backbones could achieve comparable state-of-the-art performance to that of the models dedicated to FME when provided an adequate amount of training data; therefore, we evaluate the samples with three general action recognition models using $F1_{macro}$ and accuracy as metrics and take the average as the final result. For reference, we also report the recognition results for real samples, i.e., the driving videos in the generation experiment. The training strategy of the three discriminative models is augmented by generated data, which will be specified in the next section.

As shown in Table VI, the recognition results of the real samples are all in the leading position, which provides a good reference for the remaining models. The samples generated by both FOMM and MRAA have better recognition scores when they are equipped with AWP, which indicates that they both generate more authentic data with the assistance of AWP. Notably, the recognition score on the dataset generated by FOMM with AWP is significantly higher than that generated by MRAA with AWP, which echoes the result of the significance test in the public evaluation.

The automatic evaluation protocol has the following advantages: 1) The reproducibility of the protocol using general action recognition backbones is better. 2) Averaging multiple results from different backbones instead of using only one model can reduce accidental results caused by potentially limited model performance. 3) Due to the introduced training data enlarged by generative models, the performance of the general action recognition backbones reaches applicable results. This can be seen from the fact that all three sets of networks achieved consistent conclusions, which also aligns with the expert/public evaluation.

TABLE VI

RESULTS OF AUTOMATIC EVALUATION. THE FIRST COLUMN SHOWS THE SOURCE OF DATA BEING TESTED. SCORES OF THE BETTER MODELS WITH OR WITHOUT AWP ARE SHOWN IN BOLD, WHILE THE BEST OF ALL MODELS IS UNDERLINED

Dataset	$F1_{Macro}$					Accuracy				
	MC3 [64]	R(2+1)D [64]	R3D [64]	Avg.	ΔAvg.	MC3 [64]	R(2+1)D [64]	R3D [64]	Avg.	ΔAvg.
Synthetic dataset generated by FOMM	0.389	0.265	0.481	0.378	-	0.444	0.389	0.500	0.444	-
Synthetic dataset generated by FOMM with AWP	0.389	0.452	0.372	**0.405**	+0.026	0.444	0.500	0.444	**0.463**	+0.019
Synthetic dataset generated by MRAA	0.121	0.167	0.167	0.152	-	0.222	0.333	0.333	0.296	-
Synthetic dataset generated by MRAA with AWP	0.145	0.333	0.372	**0.283**	+0.132	0.278	0.444	0.389	**0.370**	+0.074
Real	0.750	0.656	0.533	0.646	-	0.778	0.667	0.556	0.667	-

In summary, by comparing the recognition results for real samples and generated data, we can evaluate the generation quality automatically. The performance difference among models reflected by this evaluation strategy also matches the human evaluation.

I. Augmented Recognition

A recent study [7] reports that the performance of large networks could be greatly improved by considering more data. Moreover, general discriminative models can match the performance of models specifically designed for the FMER task. However, the samples are collected in real scenarios and are restricted by the expensive labeling issue. With the micro-expression generation technique, we can expand the size of the training set by a factor of ten or more, thereby substantially expanding the sample diversity of the training set. We conduct recognition experiments with the training dataset enlarged by the FME generation technique, which we refer to as augmented recognition.

In the augmented recognition experiment, we use the dataset of MEGC2019 recognition [28], which is a composite of the three mainstream micro-expression datasets, namely, CASME2, SAMM, and SMIC-HS. The dataset consists of three classes: positive, negative, and surprise. We apply subject-independent conduction in the augmented recognition, to ensure that no subject simultaneously appears in both the training and testing sets. The ratio of the training set to the testing set is 7:3.

To avoid data leakage, the generation model is trained with only the separated training set. We use the first frame of every training video clip as the source frame, and the remainder of each video clip as the driving video. Every sample in the ME dataset has a relatively neutral first frame. The generation strategy indicates that we could enlarge a training set with n samples to n^2, in theory.

Another practical problem in augmented recognition is the proportion of generated data in the training process. We calculate the loss from real samples and generated data separately, then balance their weights with a parameter. Suppose the loss from real samples is L_{real} and the loss from generated data is L_{gen}; therefore, the total loss can be formulated as:

$$L_{total} = L_{real} + \lambda L_{gen}$$

where λ is the balance parameter that can be adjusted manually. With the balance parameter, we can easily set the proportion of generated data in the training process to prevent it from overwhelming the effect of real samples.

TABLE VII

RESULTS OF AUGMENTED RECOGNITION. R.D. IS SHORT FOR REAL DATA, WHILE G.D. STANDS FOR GENERATED DATA. THE RATIO BETWEEN R.D. AND G.D. INDICATES THE MAGNITUDE OF THE GENERATED DATA ENROLLED IN THE TRAINING PROCESS COMPARED TO THE REAL DATA. RESULTS OF THE BEST CONFIGURATION OF A BACKBONE ARE SHOWN IN BOLD, WHILE THE OVERALL BEST MODEL IS UNDERLINED

Backbone	Training Set	$F1_{Macro}$
MC3-18 [64]	R.D.	0.365
	R.D. + G.D. (R.D.:G.D.=1:1.5)	0.423
	R.D. + G.D. (R.D.:G.D.=1:15)	0.377
	R.D. + G.D. (R.D.:G.D.=1:150)	**0.463**
R(2+1)D-18 [64]	R.D.	0.344
	R.D. + G.D. (R.D.:G.D.=1:1.5)	0.362
	R.D. + G.D. (R.D.:G.D.=1:15)	0.400
	R.D. + G.D. (R.D.:G.D.=1:150)	**0.519**
R3D-18 [64]	R.D.	0.388
	R.D. + G.D. (R.D.:G.D.=1:1.5)	0.440
	R.D. + G.D. (R.D.:G.D.=1:15)	0.462
	R.D. + G.D. (R.D.:G.D.=1:150)	**0.546**

The training process is implemented based on the MEB [7] library, which provides tools for data loading and training micro-expression models. Specifically, the backbones use RGB as input, whose channel number is 3. All the samples are resized to 112×112. The recognition experiments are performed on a 4-2080Ti-GPU machine. The batch size is 32 for both real samples and generated data. We use the Adam optimizer with an initial rate of 10^{-4} and steps-based learning rate reduction policy for all the backbones. All the backbones were trained with a fixed number of iterations of 150, while the enrolled generated data are used only once in the overall training process. Therefore, the ratio of enrolled generated data to real samples is equal to the number of times that the real samples are repeated. Moreover, the weight of the generated data could be adjusted by λ flexibly. Specifically, when $\lambda = 0.01$ and the number of training instances on real samples is 150, the magnitude of the enrolled generated data is 1.5 times that of real samples.

As shown in Table VII, the $F1_{macro}$ score of all the action recognition backbones improves systematically and significantly. This result shows the effect of introducing generated data into the training process to further improve the recognition performance. The backbones used in the automatic evaluation are trained and augmented by means of generated

data, while no subject or motion information in the evaluated samples appears in the training set. Since the batch size might affect the optimization results, we conduct an additional group of experiments with a batch size of 64 with no generated data being introduced. In this setting, MC3-18, R(2+1)D-18 and R3D-18 obtain F1 scores of 0.439, 0.366, and 0.375, respectively, all of which are lower than the best results yielded by the methods trained on the synthetic dataset. These results verify the effectiveness that can only be achieved through the use of generated data.

V. Future Discussion

In this section, we discuss possible future research directions based on this paper from three perspectives: potential improvement of the generation algorithm, the application scenario of boosting micro-expression recognition, and the vision of using prior knowledge in expert-dependent domains.

A. Algorithm Improvement

In this paper, we propose a micro-expression generation framework that considers facial prior knowledge. By introducing a facial prior module, we encode the spatial features into an image animation framework, addressing the shortcoming that the motion information in the rare ME samples is so sparse that existing image animation method based on self-supervision fail to capture the spatial features of micro-expressions.

The proposed work suggests that the ME generation performance of the image animation framework could be systematically improved by introducing handcrafted features, while no other losses or constraints are needed. To achieve better performance in micro-expression generation, a promising improvement can be achieved by introducing better handcrafted features, especially spatial features that are specific to the micro-expressions, such as LBP-TOP [53]. Moreover, the proposed AWP is based on the Dlib library. A model that could predict more keypoints will provide more fine-grained features to the overall framework. We believe that the plug-and-play facial prior module provides researchers, not only computer vision researchers but those who have backgrounds in psychology, an efficient interface to encode better features into the framework. Similarly, better generative models can also be introduced. In summary, due to the concise and modular setup used in this paper, the generalizability of the core idea, which is, introducing facial priors to generative models, has been demonstrated. The scalable framework proposed in this article supports follow-up researchers to introduce prior knowledge and inductive bias of research concerns through modular design.

We must also acknowledge that intricate loss functions or constraints are beneficial to the task. Subtle motion translation is a very challenging topic. A dedicated loss function for the FME generation task is a promising means to further improve the generation quality.

B. Boosting Recognition

An important application scenario of FME generation is using the generated data to improve the ME recognition performance. In this paper, we conducted an experiment to validate the effectiveness of using generated data to improve recognition. Meanwhile, we propose a method to balance the proportion of generated data by adjusting the weight of loss from generated data.

We believe that data generation is beneficial to ME recognition task not simply as a data augmentation technique. For example, existing micro-expression datasets suffer from extreme class imbalance. The micro-expression generation technique can easily address this issue. For a task with a serious lack of effective data such as micro-expression recognition, data generation could be beneficial in many ways. More strategies for how we utilize the generated data are worth exploring.

C. Prior Knowledge

This paper is based on a belief that by manually encoding prior knowledge, we can systematically improve data-driven methods in domains that lack effective data, have low-tensity features, and are highly dependent on expensive expert knowledge. Specifically, in this paper, prior knowledge is the spatial information of each video frame, which can be used to improve the performance of image animation. We look forward to newly proposed image or video generation methods based on this vision contributing to psychology.

VI. Conclusion

This paper focuses on a new facial micro-expression (FME) generation task that aims to generate novel FME videos. Different from mainstream image animation methods that pay more attention to encoding macro motion information, the FME generation task aims to encode subtle facial motion information. To this end, we utilize facial action units (AUs) and present a facial prior module to enhance the ability of the general image animation module to capture subtle facial motion features. In addition to the technical improvement, this paper provides a detailed protocol of how to evaluate the generated FME samples automatically, as well as a strategy to utilize the generated data as an augmentation to FME analysis, which would provide subjective metrics and specific application scenarios for this task. Extensive experiments on three benchmark datasets, namely, CASME II, SAMM, and SMIC, verify the effectiveness of our facial prior module, which consistently improves general image animation frameworks. In addition to using expert evaluation, we present public evaluation and automatic evaluation protocols and results, which also illustrate the superior performance of our proposed facial prior module. Extensive recognition experiments demonstrate the effectiveness of using generated data to enhance the performance of general action recognition backbones on the FMER task.

References

[1] W.-J. Yan et al., "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86041.

[2] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *J. Nonverbal Behav.*, vol. 37, no. 4, pp. 217–230, Dec. 2013.

[3] S. Porter and L. ten Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions," *Psychol. Sci.*, vol. 19, no. 5, pp. 508–514, May 2008.

[4] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1449–1456.

[5] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2028–2046, Oct. 2022.

[6] X. Ben et al., "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5826–5846, Sep. 2022.

[7] T. Varanka, Y. Li, W. Peng, and G. Zhao, "Data leakage and evaluation issues in micro-expression analysis," *IEEE Trans. Affective Comput.*, no. 1, pp. 1–12, 2023.

[8] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018.

[9] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.

[10] C. H. Yap, C. Kendrick, and M. H. Yap, "SAMM long videos: A spontaneous facial micro- and macro-expressions dataset," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 771–776.

[11] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME): A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 424–436, Oct. 2018.

[12] J. Yu, C. Zhang, Y. Song, and W. Cai, "ICE-GAN: Identity-aware and capsule-enhanced GAN with graph-based reasoning for micro-expression recognition and synthesis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[13] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 7137–7147.

[14] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13648–13657.

[15] Y. Zhang, Y. Zhao, Y. Wen, Z. Tang, X. Xu, and M. Liu, "Facial prior based first order motion model for micro-expression generation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4755–4759.

[16] M. H. Yap, J. See, X. Hong, and S.-J. Wang, "Facial micro-expressions grand challenge 2018 summary," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 675–678.

[17] W. Merghani, A. Davison, and M. Yap, "Facial micro-expressions grand challenge 2018: Evaluating spatio-temporal features for classification of objective classes," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 662–666.

[18] Y. Li, X. Huang, and G. Zhao, "Joint local and global information learning with single apex frame detection for micro-expression recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 249–263, 2021.

[19] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 6977–6991, 2020.

[20] S.-J. Wang et al., "Micro-expression recognition using color spaces," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6034–6047, Dec. 2015.

[21] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, and G. Zhao, "Domain regeneration for cross-database micro-expression recognition," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2484–2498, May 2018.

[22] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "LEARNet: Dynamic imaging network for micro expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 1618–1627, 2020.

[23] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 8590–8605, 2020.

[24] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 32–47, Jan. 2019.

[25] T. Zhang et al., "Cross-database micro-expression recognition: A benchmark," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 544–559, Feb. 2022.

[26] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[27] N. V. Quang, J. Chun, and T. Tokuyama, "CapsuleNet for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.

[28] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "MEGC 2019—The second facial micro-expressions grand challenge," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[29] X. Li et al., "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 563–577, Oct. 2018.

[30] J. Li, C. Soladié, R. Séguier, S.-J. Wang, and M. H. Yap, "Spotting micro-expressions on long videos sequences," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[31] J. Li, C. Soladié, and R. Séguier, "Local temporal pattern and data augmentation for spotting micro-expressions," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 811–822, Jan. 2023.

[32] S.-J. Wang, Y. He, J. Li, and X. Fu, "MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos," *IEEE Trans. Image Process.*, vol. 30, pp. 3956–3969, 2021.

[33] J. LI, S.-J. Wang, M. H. Yap, J. See, X. Hong, and X. Li, "MEGC2020—The third facial micro-expression grand challenge," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 777–780.

[34] H. Pan, L. Xie, and Z. Wang, "Local bilinear convolutional neural network for spotting macro- and micro-expression intervals in long video sequences," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 749–753.

[35] L.-W. Zhang et al., "Spatio-temporal fusion for macro- and micro-expression spotting in long video sequences," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 734–741.

[36] S.-J. Wang, S. Wu, X. Qian, J. Li, and X. Fu, "A main directional maximal difference analysis for spotting facial movements from long-term videos," *Neurocomputing*, vol. 230, pp. 382–389, Mar. 2017.

[37] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.

[38] J. Li et al., "CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2782–2800, May 2023.

[39] P. Ekman and W. V. Friesen, "Facial action coding system," *Environ. Psychol. Nonverbal Behav.*, 1978.

[40] J. F. Cohn and P. Ekman, "Measuring facial action," in *The New Handbook of Methods in Nonverbal Behavior Research*, vol. 525, no. 9–64. 2008, p. 1.

[41] J. Zhu, B. Wang, W. Sun, and J. Dai, "Facial expression recognition video analysis system based on facial action units : A feasible engineering implementation scheme," in *Proc. 13th Int. Symp. Comput. Intell. Design (ISCID)*, Dec. 2020, pp. 238–243.

[42] M. Nadeeshani, A. Jayaweera, and P. Samarasinghe, "Facial emotion prediction through action units and deep learning," in *Proc. 2nd Int. Conf. Advancements Comput. (ICAC)*, vol. 1, Dec. 2020, pp. 293–298.

[43] L. Zhou, Q. Mao, and M. Dong, "Objective class-based micro-expression recognition through simultaneous action unit detection and feature aggregation," 2020, *arXiv:2012.13148*.

[44] Y. Li, X. Huang, and G. Zhao, "Micro-expression action unit detection with spatial and channel attention," *Neurocomputing*, vol. 436, pp. 221–231, May 2021.

[45] A. Davison, W. Merghani, C. Lansley, C.-C. Ng, and M. H. Yap, "Objective micro-facial movement detection using FACS-based regions and baseline evaluation," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 642–649.

[46] W. Merghani and M. H. Yap, "Adaptive mask for region-based facial micro-expression recognition," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 765–770.

[47] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Learning bases of activity for facial expression recognition," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1965–1978, Apr. 2017.

[48] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[49] S. Sharma, K. Shanmugasundaram, and S. K. Ramasamy, "FAREC—CNN based efficient face recognition technique using Dlib," in *Proc. Int. Conf. Adv. Commun. Control Comput. Technol. (ICACCCT)*, May 2016, pp. 192–195.

[50] S. Mohanty, S. V. Hegde, S. Prasad, and J. Manikandan, "Design of real-time drowsiness detection system using Dlib," in *Proc. IEEE Int. WIE Conf. Electr. Comput. Eng. (WIECON-ECE)*, Nov. 2019, pp. 1–4.

[51] H. Yuhong, "Research on micro-expression spotting method based on optical flow features," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4803–4807.

[52] S. Zhao et al., "A two-stage 3D CNN based learning method for spontaneous micro-expression recognition," *Neurocomputing*, vol. 448, pp. 276–289, Aug. 2021.

[53] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[54] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Process., Image Commun.*, vol. 74, pp. 129–139, May 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596518310038

[55] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[56] T. Varanka, W. Peng, and G. Zhao, "Micro-expression recognition with noisy labels," *Electron. Imag.*, vol. 33, no. 11, pp. 157-1–157-8, Jan. 2021.

[57] J. Zhang, F. Liu, and A. Zhou, "Off-TANet: A lightweight neural micro-expression recognizer with optical flow features and integrated attention mechanism," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, Oct. 2021, pp. 266–279.

[58] M. Wei, W. Zheng, Y. Zong, X. Jiang, C. Lu, and J. Liu, "A novel micro-expression recognition approach using attention-based magnification-adaptive networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2420–2424.

[59] H. Li, M. Sui, Z. Zhu, and F. Zhao, "MMNet: Muscle motion-guided network for micro-expression recognition," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1–7.

[60] S. P. Teja Reddy, S. Teja Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[61] M. Verma, S. K. Vipparthi, and G. Singh, "Non-linearities improve OrigiNet based on active imaging for micro expression recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[62] M. Verma, M. S. K. Reddy, Y. R. Meedimale, M. Mandal, and S. K. Vipparthi, "AutoMER: Spatiotemporal neural architecture search for microexpression recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6116–6128, Nov. 2022.

[63] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[64] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High-fidelity face manipulation with extreme poses and expressions," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2218–2231, 2021.

[65] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "Marionette: Few-shot face reenactment preserving identity of unseen targets," in *Proc. AAAI*, vol. 34, 2020, pp. 10893–10900.

[66] J. Zhang et al., "FReeNet: Multi-identity face reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5325–5334.

[67] X. Wu et al., "F^3A-GAN: Facial flow for face animation with generative adversarial networks," *IEEE Trans. Image Process.*, vol. 30, pp. 8658–8670, 2021.

[68] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, "Unsupervised discovery of object landmarks as structural representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2694–2703.

[69] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, "Unsupervised learning of object landmarks through conditional image generation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 4020–4031.

[70] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-GAN: Unsupervised video retargeting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 119–135.

[71] O. Wiles, A. Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 670–686.

[72] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2372–2381.

[73] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.

[74] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7083–7092.

[75] X. Fan, A. R. Shahid, and H. Yan, "Facial micro-expression generation based on deep motion retargeting and transfer learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4735–4739.

[76] J. Li, C. Soladie, and R. Seguier, "A survey on databases for facial micro-expression analysis," in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 241–248.

[77] G. D. Sad, F. Reyes, and J. Alvarez, "FaceTrack: Asymmetric facial and gesture analysis tool for speech language pathologist applications," in *Proc. 1st Workshop Facial Micro-Expression: Adv. Techn. Facial Expressions Gener. Spotting*, Oct. 2021, pp. 1–10.

[78] Z. Dong, G. Wang, S. Lu, W.-J. Yan, and S.-J. Wang, "A brief guide: Code for spontaneous expressions and micro-expressions in videos," in *Proc. 1st Workshop Facial Micro-Expression: Adv. Techn. Facial Expressions Gener. Spotting*, Oct. 2021, pp. 31–37.

[79] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME)2: A database for spontaneous spotting and recognition," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 1–14, Jan. 2017.

[80] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2016, pp. 694–711.

[81] Z. Geng, C. Cao, and S. Tulyakov, "3D guided fine-grained face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9813–9822.

[82] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[83] Y. Xu, S. Zhao, H. Tang, X. Mao, T. Xu, and E. Chen, "FAMGAN: Fine-grained AUs modulation based generative adversarial network for micro-expression generation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4813–4817.