

Facial Prior Based First Order Motion Model for Micro-expression Generation

Yi Zhang*
Youjun Zhao*
Yuhang Wen
Zixuan Tang
Xinhua Xu

School of Intelligent Systems Engineering,
Sun Yat-sen University

Mengyuan Liu[†]
nkliuyifang@gmail.com
School of Intelligent Systems Engineering,
Sun Yat-sen University
Guangdong Provincial Key Laboratory
of Fire Science and Technology

ABSTRACT

Spotting facial micro-expression from videos finds various potential applications in fields including clinical diagnosis and interrogation, meanwhile this task is still difficult due to the limited scale of training data. To solve this problem, this paper tries to formulate a new task called micro-expression generation and then presents a strong baseline which combines the first order motion model with facial prior knowledge. Given a target face, we intend to drive the face to generate micro-expression videos according to the motion patterns of source videos. Specifically, our new model involves three modules. First, we extract facial prior features from a region focusing module. Second, we estimate facial motion using key points and local affine transformations with a motion prediction module. Third, expression generation module is used to drive the target face to generate videos. We train our model on public CASME II, SAMM and SMIC datasets and then use the model to generate new micro-expression videos for evaluation. Our model achieves the first place in the Facial Micro-Expression Challenge 2021 (MEGC2021), where our superior performance is verified by three experts with Facial Action Coding System certification. Source code is provided in <https://github.com/Necolizer/Facial-Prior-Based-FOMM>.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision problems**: Image processing; • **Applied computing** → **Psychology**.

KEYWORDS

Micro-expression generation, Facial micro-expression, Generative adversarial network, Deep learning, Facial landmark

ACM Reference Format:

Yi Zhang, Youjun Zhao, Yuhang Wen, Zixuan Tang, Xinhua Xu, and Mengyuan Liu. 2021. Facial Prior Based First Order Motion Model for Micro-expression

*Both authors contributed equally to this research.

[†]Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8651-7/21/10...\$15.00

<https://doi.org/10.1145/3474085.3479211>

Generation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3474085.3479211>

1 INTRODUCTION

Micro-expressions are brief and involuntary facial expressions that usually last for 1/25 to 1/5 of a second. It is hard for human being to notice facial micro-expressions (FMEs) due to short duration, low intensity, and local-only occurrence [12, 19]. Recently, spotting micro-expressions with machine learning methods have attracted lots of attentions. Data-driving machine learning methods need large scale data to obtain an accurate yet robust model. However, building high-quality FME datasets would be expensive and time-consuming, which leads to the problems: 1) The size of constructed databases is usually too small to assure the robustness of machine learning models. 2) the imbalanced data distribution of currently available databases may lead to unsatisfying training results [8].

To solve this problem, one solution is facial micro-expressions (FME) generation. Conceptually, the goal of FME generation is to generate micro-expression on given template faces. Challenges in FME generation task are the follows: 1) Few references since it just started studying recently. 2) FMEs are subtle and hard-to-capture. Algorithms often have problem obtaining information of small changes and motions. 3) Traditional image processing methods utilizing handcrafted features get stuck in generation tasks, since the visual result of reconstruction is usually bad, and these methods have poor versatility. Deep learning methods need to be applied.

This paper presents a new method named Facial Prior Based First Order Motion Model for micro-expression generation, where we first extract the motion patterns using the regions feature computed according to the facial prior, and then utilize generative adversarial network to reconstruct the face and generate video. Visual results are involved to show the superior performance of our method. Also, evaluations from three experts verified that our method outperforms other methods in MEGC2021.

2 RELATED WORK

Image animation is a way to generate videos by animating objects in still images. Most popular ways of image animation is through deep learning, such as Generative Adversarial Networks (GANs) [4] and Variational Auto-Encoders (VAEs) [7]. However, these methods are only used in generating the macro, exaggerated expressions. Since FMEs are small, these methods may not lead to a convincing result.

Apparently, a satisfying result should take the subtle movement of the important facial areas into considerations.

Deep generative models for image animations and video retargeting [1, 10, 18] have emerged in recent years. Some models have been proposed, such as Monkey-Net [13], FOMM [14] and MRAA [15] to get better performance in modeling object’s motion. These models encode motion information of the key points or areas in the videos, which are learned in self-supervised way. The performance of self-supervised depends on the diversity of samples. Concerning the motions of micro-expressions are too subtle for models to capture, self-supervised module should be replaced with prior knowledge about FMEs so that reconstructed videos could be better. Qiao et al [11] proposed Geometry-Contrastive GAN for Facial Expression Transfer which is related to our work. However, their work is limited to macro-expression transfer.

3 PROBLEM FORMULATION

We have a target image T , and a series of source images $S = \{S_k\}_{k=1}^n$, also called a driving video, where S_1 represents the onset frame and S_n represents the offset frame of a micro-expression. Suppose function $Motion(S_i, S_j)$ is the motion representation from frame S_i to S_j , and a function $Move(T_1, Motion(\cdot))$ transforms the target face according to the motion representation using distortion, rotation, or other affine transformation. Then this type of FME generation could be formulated as:

$$T_k = Move(T_i, Motion(S_j, S_k)) \quad (1)$$

There would be various subclasses if we take different T_i and S_j . Like inter-frame motion estimation, which is defined as:

$$T_k = Move(T_{k-1}, Motion(S_{k-1}, S_k)) \quad (2)$$

We believe this kind of problem modeling is preferable and our proposed method is based on this methodology. We hope for a model which could generate T_k having the same appearance of the face in S as well as containing the semantic information of motion extracted from the driving frame S_k . Let the generation model be \mathcal{G} , then the problem can be described as $T_k = \mathcal{G}(T_1, S_k)$. The generated video is $GT = \{T_k\}_{k=1}^n = \mathcal{G}(T, S)$. Two modules are needed in \mathcal{G} : motion extraction and face reconstruction, denoted as \mathcal{M} and \mathcal{R} (similar to $Motion(\cdot)$ and $Move(\cdot)$ defined above). Specifically, the reconstruction module should utilize information extracted from the motion extraction module. Then \mathcal{G} could be specified as $\mathcal{G}(\cdot) = \mathcal{R}(\mathcal{M}(\cdot))$.

Since the target face is practically different from driving faces, inspired by FOMM [14] and MRAA [15], we assume a virtual reference frame R and an operator \circ representing superposition of motions. Then our method can be formulated as:

$$GT = \mathcal{G}(T, S) = \mathcal{R}(T, \mathcal{M}(T, R) \circ \mathcal{M}(R, S)) \quad (3)$$

4 PROPOSED METHOD

Fig. 1 shows the general framework of our method. Given a target face, along with a video containing FME, Region-Focusing Module computes a facial prior map. Then the facial prior map is fused with the original frame as the input of Motion Prediction Module. This module uses the key points given explicitly along with local affine transformations to estimate complex motions by predicting

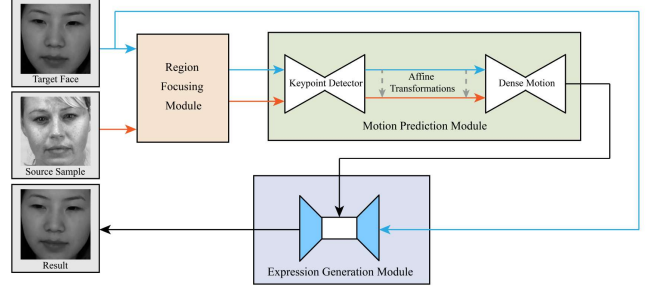


Figure 1: General framework of our method

backward optical flow. Besides, this module also predicts the occlusion map that indicates the part which could not estimate from the target image, improving the quality of the generated videos. In Expression Generation module, GAN makes use of the predicted optical flow and the occlusion map while encoding so that the decoded image contains the motion information, driving the target to generate FME videos.

4.1 Region-Focusing Module

In this section, we will illustrate the algorithm we designed to highlight regions of interest (ROI) for FME.

As mentioned above, there are many priors could be used to extract the feature represented the motion of the driving videos. We noticed that the Facial Action Coding Systems(FACs), which displays 68 facial landmarks in human faces, is a good prior to locate the regions in faces. With the pre-trained model in dlib [6], the 68 facial landmarks can be automatically predicted, as shown in Fig. 2 (a). But we noticed that not all regions are necessary since FME may only appear in some certain regions. Now the problem is how to focus on the ROI for the FME motion prediction, where the most obvious movements may occur.

We measure the importance of each pixel according to its distance away from the key points mainly selected from the 68 facial landmarks, which are colored in Fig. 2 (b). Some of the key points are moved horizontally and vertically from the facial landmarks by half the pupillary distance. Specifically, we denote the k^{th} key point in a facial image is $p_k(x_k, y_k) \in \mathbb{R}^{H \times W}$, and the certain i^{th} pixel in the facial image as $p_i(x_i, y_i) \in \mathbb{R}^{H \times W}$. Their Euclidean distance d_{ik} is calculated as

$$d_{ik} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \quad (4)$$

With the calculated distance, we could indicate the importance of this pixel using the gaussian kernel function, described as

$$r_{ik}(x_i, y_i) = e^{-\frac{d_{ik}^2}{2\sigma^2}} \quad (5)$$

in which e is the base of natural logarithm, σ is the variance of the gaussian distribution (set manually). Now the r_{ik} indicates the degree of interest of this pixel concerning the key point p_k . With each pixel’s degree of interest to the p_k , a map S_{km} (Fig. 3 (a)), which indicates weighted importance, is formulated as:

$$S_{km} = \sum_{r_{ik} \in \mathbb{R}^{H \times W}} r_{ik}(x_i, y_i) \quad (6)$$

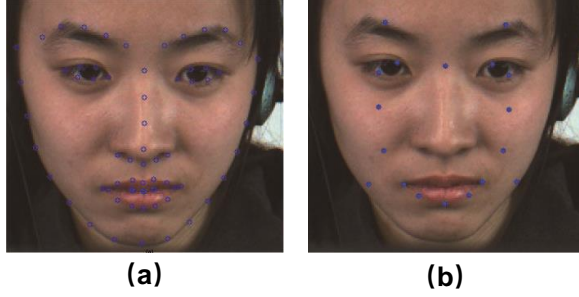


Figure 2: Facial prior generation: (a) Original detected key points using facial landmark detection method. (b) Modified key points used as facial prior for our framework.

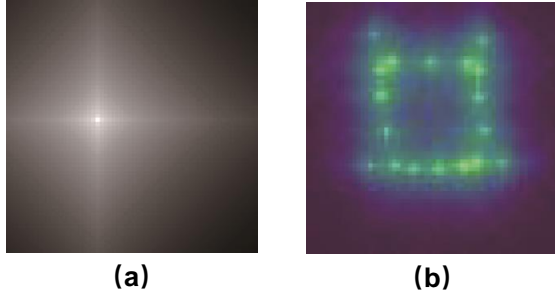


Figure 3: Facial prior map generation: (a) Contribution of one key point for the facial prior map. (b) Facial prior map generated with all modified key points.

Assume that n key points are selected, then we could get a synthesized facial prior map S_m in Fig. 3 (b), described as $S_m = \sum_k S_{km}$. The original frame image is fused with the facial prior map, forming the sample $S' \in \mathbb{R}^{(C+1) \times H \times W}$. S' more information related to the ROI. Together with the corresponding predicted key points, these manual-feature enhanced samples are next passed to the motion prediction module for further feature representation using neural network.

4.2 Motion Prediction Module

Motion Prediction Module is a two-step model proposed by Siarohin et al [14]. This module can estimate the motion tendency from a driving frame image S in source video to the input target image T . The first step is utilizing a key point detector to extract key points in T and S as well as a local affine transformation to represent motion in the local area of each key point from S to T . Taylor expansion is used in calculating the affine transformation parameters of key points from S to T . Additionally, by introducing reference frame R , the prediction of transformation from S to T can be divided into two parts - from S to R and then R to T . During the second step, a dense motion network obtains a backward optical flow and a occlusion mask from S to T to decide which part of the image should be reconstructed after getting the transformation parameters from the previous step.

Inspired by this work, we designed Motion Prediction Module as follows. We firstly concatenate image and facial prior map S_m synthesized as an additional channel, instead of putting images

directly to the key point detector. Therefore, target matrix and driving matrix of $S' \in \mathbb{R}^{(C+1) \times H \times W}$ are sent into key point detector to extract features. The additive facial prior map represents the knowledge of occurrence area of micro-expression. With facial prior map, the output key points from key point detector can be located mainly focusing on the neighborhood of micro-expression areas and the movement of the entire head, which can be moved from S to T s using the key point trajectories in the subsequent local affine transformation.

4.3 Expression Generation Module

Generative adversarial network is taken in our Expression Generation Module. An auto-encoder structure is applied to generate image with a target image T as input. The backward optical flow and occlusive map calculated in 4.3 are sent into the encoder block to warp the feature map produced from two layers of down-sampling convolution. Then, a discriminator similar to pixel-to-pixel structure is used to decide if the reconstructed face image \hat{S} from generator is real or not.

Several losses are applied to train our framework since multiple networks are used in our system. We use an end-to-end perceptual loss of Johnson et al. [5] using pre-trained Vgg-19 and Mean Absolute Error (MAE) in other networks. Different weight coefficients are applied to our functions in which the perceptual loss is the highest. Given the reconstructed face image \hat{S} and driving frame S , the perceptual loss can be described as:

$$L_{perceptual}(\hat{S}, S) = \sum_{i=1}^I |F_i(\hat{S}) - F_i(S)| \quad (7)$$

where F_i represents i^{th} channel of the feature map in Vgg-19 and I is the number of feature channels in this layer.

5 EXPERIMENTS

We evaluated our method using CASME II [16], SAMM [2] and SMIC [9] datasets. Given 2 normalized template faces as target (selected from CASME I [17] and SMIC), we were required to generate FMEs using 9 certain driving videos as source. So we trained our model on the whole datasets except the 9 source videos and generated FMEs on target faces for expert evaluation. Preprocessing included converting to grayscale, detecting face, cropping the facial area, resizing to 256×256 , and concatenating the same grayscale to get an image with 3 channels. Our final results are also grayscale because in our opinions generated grayscale seem visually better than color images. Key points were selected as Fig. 2(b). Every frame in a video shared the same set of key points predicted by dlib according to the onset frame. In this way the computing costs were reduced. In addition, synthesized facial prior map S_m got normalized before entering Motion Prediction Module.

The generation challenge result of MEGC2021, presented in Table 1, shows that our method outperforms methods proposed by other participating teams and won the first place through subjective scoring by three experts who have Facial Action Coding System (FACS) certification [3], provided in <https://megc2021.github.io/GeneResultevaluation.html>. Fig. 4. presents the visual results of experiment conducted by ourselves, which compares our proposed method with other baseline methods. The numbers indicate the

Table 1: Overall evaluation of MEGC2021 Generation Challenge

Methods	Expert1	Expert2	Expert3	Overall	Normalized			
					Expert1	Expert2	Expert3	Overall
A(ID:3311)	85	51	37	173	85/140	51/107	37/76	1.57062
B(ID:3320)	104	66	66	236	104/140	66/107	66/76	2.228101
C(ID:3282)	140	107	56	303	140/140	107/107	56/76	2.736842
Ours(ID:3295)	139	101	76	316	139/140	101/107	76/76	2.936782

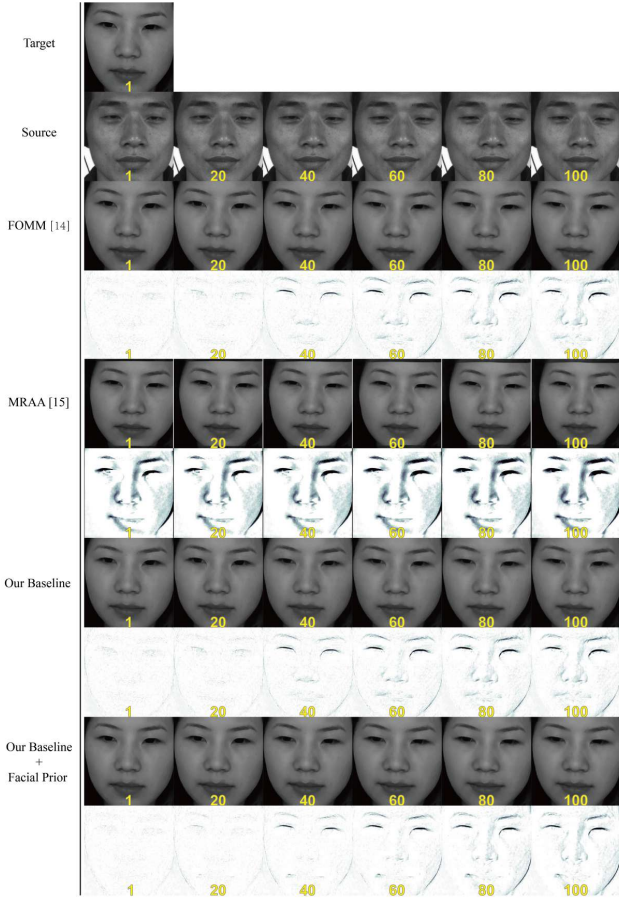


Figure 4: Results of different methods. To facilitate observation, we also show images on the 4th, 6th, 8th, 10th Lines, which are differential images between the 3rd, 5th, 7th, 9th Lines and the target image.

current frame numbers. We suggest visiting our GitHub page above for GIF to get more intuitive visual results. Compared with FOMM and MRAA, our method could generate more spontaneous, natural, and smooth FMEs with less noise. Also, the effectiveness of facial prior is proved compared with baseline. Moreover, we trained our model on 4 different training sets, which are SAMM, SMIC-HS, CASME II, and a mix of these three. We found that different training datasets would lead to nuances, as shown in Fig. 5.

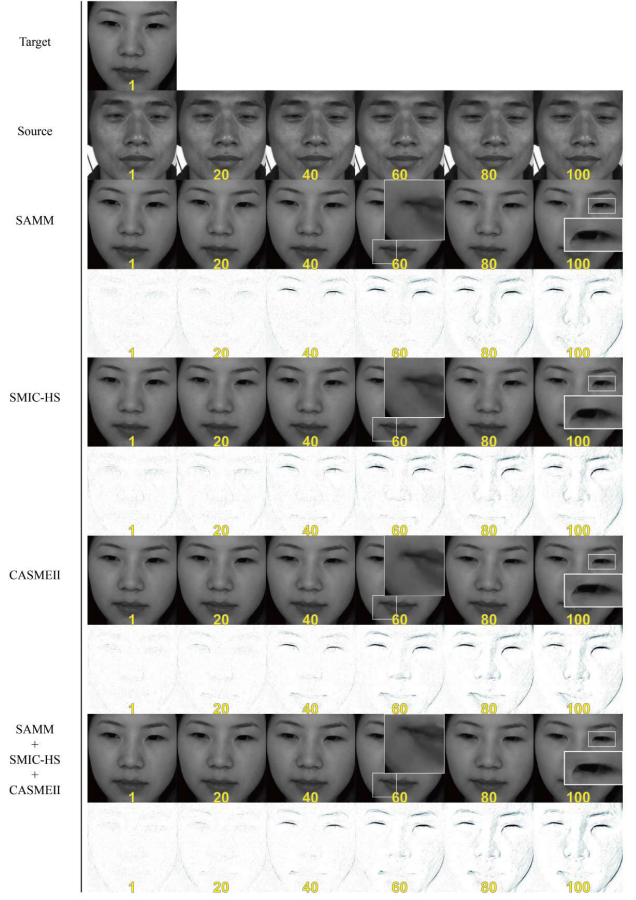


Figure 5: Our results with different training datasets

6 CONCLUSION

This paper presents a facial-prior-based first order motion model for facial micro-expression generation. Specifically, We take prior of the facial area into considerations and design a region-focusing module to get facial prior features, a motion prediction module to estimate facial motions and an expression generation module to generate FME videos. By training on CASME II, SAMM and SMIC datasets, our method achieves superior performances verified by three experts with Facial Action Coding System certification.

REFERENCES

- [1] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. RecycleGAN: Unsupervised Video Retargeting BT - Computer Vision – ECCV 2018, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.).

- Springer International Publishing, Cham, 122–138.
- [2] A K Davison, C Lansley, N Costen, K Tan, and M H Yap. 2018. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Transactions on Affective Computing* 9, 1 (2018), 116–129. <https://doi.org/10.1109/TAFFC.2016.2573832>
 - [3] P Ekman and W Friesen. 1978. Facial action coding system: a technique for the measurement of facial movement.
 - [4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) (NIPS'14). MIT Press, Cambridge, MA, USA, 2672–2680.
 - [5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution BT - Computer Vision – ECCV 2016, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 694–711.
 - [6] V Kazemi and J Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1867–1874. <https://doi.org/10.1109/CVPR.2014.241>
 - [7] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]
 - [8] Anh Cat Le Ngo, Raphael Chung-Wei Phan, and John See. 2015. Spontaneous Subtle Expression Recognition: Imbalanced Databases and Solutions. In *Computer Vision – ACCV 2014*, Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang (Eds.). Springer International Publishing, Cham, 33–48.
 - [9] X Li, T Pfister, X Huang, G Zhao, and M Pietikäinen. 2013. A Spontaneous Micro-expression Database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–6. <https://doi.org/10.1109/FG.2013.6553717>
 - [10] Yahui Liu, Marco De Nadai, Gloria Zen, Nicu Sebe, and Bruno Lepri. 2019. Gesture-to-Gesture Translation in the Wild via Category-Independent Conditional Maps. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 1916–1924. <https://doi.org/10.1145/3343031.3351020>
 - [11] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. 2018. Geometry-Contrastive GAN for Facial Expression Transfer. arXiv:1802.01822 [cs.CV]
 - [12] Xun-bing Shen, Qi Wu, and Xiao-lan Fu. 2012. Effects of the duration of expressions on the recognition of microexpressions. *Journal of Zhejiang University SCIENCE B* 13, 3 (2012), 221–230. <https://doi.org/10.1631/jzus.B1100063>
 - [13] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating Arbitrary Objects via Deep Motion Transfer. arXiv:1812.08861 [cs.GR]
 - [14] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*, H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, and R Garnett (Eds.), Vol. 32. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf>
 - [15] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion Representations for Articulated Animation. arXiv:2104.11280 [cs.CV]
 - [16] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. 2014. CASME II: an improved spontaneous micro-expression database and the baseline evaluation. *PLoS one* 9, 1 (2014), e86041. <https://doi.org/10.1371/journal.pone.0086041>
 - [17] Wen-Jing Yan, Q Wu, Yong-Jin Liu, Su-Jing Wang, and X Fu. 2013. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–7. <https://doi.org/10.1109/FG.2013.6553799>
 - [18] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. 2019. DwNet: Dense warp-based network for pose-guided human video generation. arXiv:1910.09139 [cs.CV]
 - [19] Ling Zhou, Qirong Mao, and Ming Dong. 2021. Objective Class-based Micro-Expression Recognition through Simultaneous Action Unit Detection and Feature Aggregation. arXiv:2012.13148 [cs.CV]