Automated Emotional Valence Estimation in Infants with Stochastic and Strided Temporal Sampling

Mang Ning
Department of ICS
Utrecht University
Utrecht, Netherlands
m.ning@uu.nl

Itir Onal Ertugrul
Department of ICS
Utrecht University
Utrecht, Netherlands
i.onalertugrul@uu.nl

Daniel S. Messinger
Department of Psychology
University of Miami
Miami, U.S.A.
dmessinger@miami.edu

Jeffrey F. Cohn
Department of Psychology
University of Pittsburgh
Pittsburgh, U.S.A.
jeffcohn@pitt.edu

Albert Ali Salah
Department of ICS
Utrecht University
Utrecht, Netherlands
a.a.salah@uu.nl

***Abstract*—We propose the first automated approach to estimate the emotional valence of infants from their facial behavior. We use the state-of-the-art transformer-based video masked autoencoder (VideoMAE) that is pre-trained on a large video dataset as a backbone, and finetune it on two large, well-annotated infant video datasets (SIBSMILE and MODELING). To augment the limited data, we propose a novel video temporal augmentation method called Stochastic and Strided Temporal Sampling (SSTS). We demonstrate the effectiveness of our approach for infant valence estimation by achieving 0.671 Concordance Correlation Coefficient (CCC) on SIBSMILE and MODELING. The experiments show that SSTS remarkably accelerates the training speed by 8 times while gaining the best valence estimation performance. Lastly, we suggest that face detection and cropping (coarse registration) is a promising alternative to landmark-based registration (i.e. fine registration) in data pre-processing when accurate infant facial landmark detectors are inaccessible.**

***Index Terms*—infant emotional valence estimation, facial expression recognition, video transformers**

I. Introduction

Facial expression recognition, as an integral part of affective computing research, has significantly advanced over the past 10 years [1]–[5] thanks to the progress in deep learning. Automated analysis of affect and emotion from facial expressions has various applications, such as detecting the severity of mental disorders [6], improving human-robot interactions [7], and enhancing user experience in virtual reality environments [8]. There are two major approaches to modelling facial expressions of emotions according to psychological research: categorical theory [9] and dimensional theory [10]. The former categorizes emotions into a set of distinct states, whereas the latter views emotions as points in a continuous space, defined by valence and arousal. The dimensional theory is currently gaining more popularity as it takes the intensity of the actions into account in addition to their presence, and allows for a more nuanced understanding of emotional states. More specifically, valence estimation in dimensional theory aims to evaluate the degree of positive or negative affect in an emotional state.

Recently, emotional valence estimation from facial expressions has been investigated for adults. For example, FaceBehaviorNet [11] and FATAUVA-Net [12] have shown good performance on Aff-Wild [13] and Aff-Wild2 [14] datasets to estimate the valence in the adult faces. However, research on valence estimation in infants is still missing. Considering that detecting facial affect plays a critical role in monitoring infant health and development, we propose an approach to automatically estimate valence in infant faces using two large video datasets: SIBSMILE and MODELING. Arousal is not considered in this paper because the above two datasets do not yet provide corresponding annotations.

Video-level analysis of facial expressions is more informative than frame-level analysis since subtle actions may not be perceived when motion information is missing and temporal context provides additional information to detect facial actions [15]. On the other hand, video understanding is much more challenging than image understanding as it requires modeling the temporal dynamics and demands larger amounts of training data. We finetune the state-of-the-art video transformer VideoMAE [16] that is pre-trained on large video databases using SIBSMILE and MODELING for the task of infant valence estimation. The impressive results demonstrate the effectiveness of VideoMAE, as well as verify the importance of pre-training.

As training deep video networks requires an extensive amount of samples, data augmentation has been widely used for alleviating data scarcity. Video data augmentation can further exploit temporal augmentation by randomly cropping a segment [16], [17] or sampling frames with a stride [18], [19] on account of the high redundancy in the time dimension. In the databases we introduce, each video sample has a one-second fixed duration, which differs from the diverse temporal length in other video datasets [20]–[22]. Thus, we propose Stochastic and Strided Temporal Sampling (SSTS) for valance estimation from videos. SSTS greatly enlarges the diversity of training samples and remarkably accelerates the training speed by 8x on SIBSMILE + MODELING while achieving the best performance of CCC=0.671.

Lastly, face registration is a commonly used pre-processing step in facial expression recognition in adults. It makes network training easier by aligning the input faces with a template face, typically based on the detected facial landmarks [15]. However, infant faces have different proportions, fewer wrinkles, and less texture compared to adults [23]. Therefore, facial landmark detectors trained on adults do not generalize well to infants, which may lead to incorrect alignment of

979-8-3503-2743-4/23/$31.00

infant faces. As there is no publicly available model for infant face landmark detection that is trained and evaluated on large infant databases, we recommend coarse registration as an alternative to fine registration because the former does not rely on landmark detection. Our experiments show that coarse registration achieves comparable performance to fine registration on SIBSMILE and MODELING but has the advantage of not requiring a landmark detector.

Overall, our contributions in this paper are threefold:

- We automatically estimate emotional valence in infant faces for the first time in the literature on two large, well-annotated infant databases using VideoMAE which is pre-trained on large video datasets and yields strong performances.
- We propose a video temporal augmentation method for the facial valence estimation task, demonstrating good efficiency and flexibility.
- We empirically show that coarse registration is a promising pre-processing method for automated valence estimation in the absence of accurate infant facial landmark detectors that can provide fine registration.

II. Related work

A. Facial Expression Recognition

The vast majority of research on facial expression recognition focuses on adults and can be grouped into three categories. The first group of works follows a categorical approach and aims to recognize the six universal expressions [2], [24], [25], including anger, fear, sadness, disgust, surprise, and happiness. The second group focuses on detecting the facial action units, which are the actions of the individual or a group of facial muscles, described in the Facial Action Coding System (FACS) [26]. These action units can describe complex facial expressions. Recently, several approaches have been proposed that focus on several aspects of AU detection including cross-domain detection [3], dynamics of actions [4], [5], and region-based detection [27], [15]. The third group of works considers dimensional affect analysis to understand human emotions where the arousal axis indicates the level of emotional activation, and the valence axis measures the level of pleasure [13]. Thanks to the advances in deep learning, many modern neural networks have been developed for automated valence and arousal estimation. Among them, FaceBehaviorNet [11] and FATAUVA-Net [12] present good performance on the datasets Aff-Wild [13] and Aff-Wild2 [14].

However, automated facial expression analysis in infants is extremely rare. Although a recent work [23] introduced an automated action unit detector in infants, the research on automated estimation of emotional valence from infant facial expressions is still missing. To fill this gap, we delve into the infant facial valence estimation with the state-of-the-art video transformers [16].

B. Video Data Augmentation

Data augmentation [28] has been largely used in various computer vision tasks to extend the training data distribution. Cropping, flipping, and color jittering [29], [30] are the commonly applied image augmentation methods in both small datasets [31] and large datasets [32]. Also, combing multiple data samples [33], [34] was exploited to achieve Vicinal Risk Minimization [35], and learning data augmentation strategies from data were introduced in [36] [37]. In addition to augmentation at the frame level, video augmentation introduces variation in the time dimension. Due to the high redundancy of two consecutive frames in a video, sparse temporal sampling [18], [19] and subsampling short video clips [16], [17] are often used for video classification. Specifically, given a raw video with a variable number of frames, a common way of augmentation is first randomly selecting a short, fixed-length clip with f consecutive frames, then sparsely drawing 16 or 32 frames from this clip, and performing augmentation (e.g. random cropping, color jittering) on each frame.

One reason for selecting a segment of the video is that the videos in the public datasets [20]–[22] are of different frame lengths (ranging from a few seconds to several minutes). Although this temporal sampling strategy facilitates constructing training samples with a stable temporal change rate, it also brings the issue of missing temporal information of the original video. In contrast, datasets for facial valence estimation do not have the variable duration problem since the annotation is typically assigned second by second. Therefore, our proposed augmentation method can model the complete temporal dependencies by involving the head and tail frames in a video clip.

C. Video Transformers

Recently, video transformers [38]–[40] have shown superior performance compared to 3D ConvNets [41]–[43] as the video feature learners due to their flexible self-attention mechanism [44] and the reduced inductive bias in visual encoding. Yet, transformers require large amounts of training data. When annotations are limited, supervised training may not be ideal. On the other hand, initially pre-training the video transformers in a self-supervised manner that aims to reconstruct the input from itself and then fine-tuning them for downstream tasks largely mitigates the demand for massive annotations. The masked autoencoding strategy, which is based on removing parts of input before feeding it into the model and reconstructing the full input, has been successful in NLP [45]. This strategy has been lately introduced into the computer vision on images [46] and videos [16], outperforming contrastive learning methods [47], [48]. VideoMAE [16] inherits the masking strategy from Masked Autoencoder [49] and performs self-supervised learning for video understanding, resulting in state-of-the-art performances on various video benchmarks. In this paper, we take advantage of the strength of VideoMAE in self-supervised video feature learning and fine-tune the network for infant valence estimation.

TABLE I
THE STATISTICS OF SIBSMILE AND MODELING

	SIBSMILE	MODELING
# subjects (infants)	25	16
# training video clips (training samples)	8252	4230
# training image frames	247k	127k
# testing video clips (testing samples)	2627	1375
# testing image frames	79k	41k
fps	30	30
frame resolution	720x486	1288x964

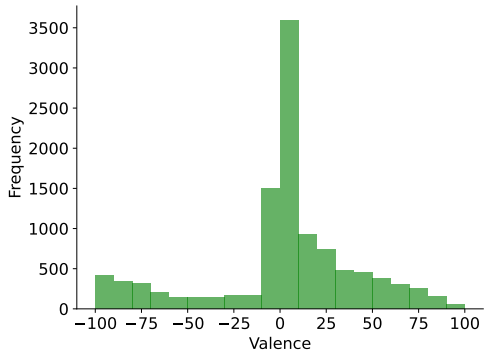
III. METHOD

A. Dataset and Pre-processing

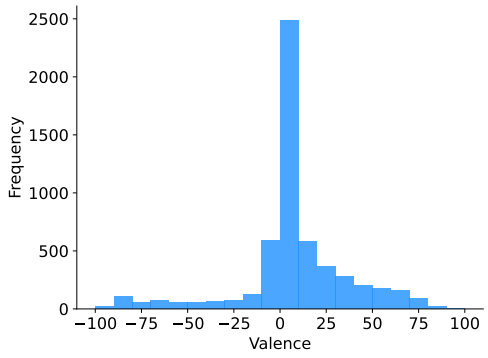
To the best of our knowledge, there are no large, manually annotated, publicly available infant valence datasets. We therefore first introduce two private infant facial expression datasets: SIBSMILE and MODELING. The statistics of these two datasets are shown in Table I. Similar to the MIAMI dataset [50] [51], the SIBSMILE and MODELING datasets record the interaction between mothers and six-month-old infants in a Face-to-Face/Still-Face/Resume (FF/SF/RE) protocol [52]. There are 25 and 16 babies with spontaneous behavior in SIBSMILE and MODELING, respectively. Training and test folds do not contain the same subjects. The main differences between SIBSMILE and MODELING are that the former is more in the wild, has a lower resolution, and has larger out-of-plane head motion. The annotations are generated by four domain experts and we use the mean as the ground truth valence score based on the high intraclass correlation coefficient (ICC) of raters we evaluated on both the MODEILNG and SIBSMILE datasets, where ICC(3, k)= 0.959 and ICC(3, k)= 0.936, respectively. Each second of video content is assigned a valence value (ranging from -100 to 100). As a result, we collected 10879 video clips (i.e. training samples and testing samples) for SIBSMILE and 5605 clips for MODELING, with each clip lasting one second. Fig. 1 presents the histograms of valence annotations. The distributions are similar to that of the Aff-wild2 dataset [14], in which most valence annotations are distributed around the neutral expression.

Since the raw video clips are mostly noisy, pre-processing is necessary before training the neural network. Following [11] [12], we apply face detection, resizing, and cropping to generate fixed-size frames. Additionally, [15] points out that face registration is beneficial to facial expression recognition, thus we conduct two different pre-processing methods. Coarse registration is based on 2D face cropping and scaling, whereas fine registration is based on 2D facial landmarking and alignment to a consensus shape.

a) Coarse registration: Considering the lack of publicly available infant face detectors, we use the state-of-the-art adult face detector [53] as an alternative. We find that this detector has a good generalization on infant data even though it has never been trained on infant faces. Throughout the pre-processing, we first apply face detection to locate the infant face in each frame. In the meantime, we remove the frames in which the face occurrence probability is below the threshold of 0.8. Then, we crop the face area according to the coordinates predicted by the face detector. Finally, each frame is resized to 384×384 pixels. Note that we also remove the video clips whose valid face frames are less than 24 because these clips have very few detected infant faces. As a result, each training clip in our SIBSMILE and MODELING datasets contains 24 to 30 face frames.



(a) SIBSMILE



(b) MODELING

Fig. 1. Histograms of valence annotations of SIBSMILE and MODELING datasets

b) Fine registration: Since 3D face registration requires a standard morphological model of infant faces which is not available yet, we apply 2D face registration to infant faces in SIBSMILE and MODELING. The difference in pre-processing between coarse and fine registration is that fine registration adds two more steps after face detection: face landmark detection and similarity transformation. We employ 2D-FAN [53] network to detect 68 face landmarks [54]. Then, five landmarks (left eye center, right eye center, tip of nose, left corner of mouth, right corner of mouth) are used for a 2D similarity transformation [55]. The registration is accomplished by linearly transforming the original face to a template face [56] whose eye centers are horizontal and the face is in the center of the image. Note that, unlike 3D face registration that achieves a frontal face, 2D face registration cannot eliminate out-of-plane motion and does not guarantee a frontal face.

B. Data Augmentation - Stochastic and Strided Temporal Sampling

High temporal redundancy is a general prior in video data [57], therefore, sparsely sampling partial frames from the raw video [18] is often adopted to reduce the computational overload and yield the fixed size of input frames for neural networks. Large video datasets (SSV2 [20], Kinetics-400 [21], UCF101 [22]) have a variable duration and each clip lasts from several seconds to minutes, so it is common to sample a segment in the clip for training to avoid computational explosion. But this could be problematic since such segments can not ensure capturing the complete temporal information. On the contrary, each training clip in SIBSMILE and MODELING has a fixed temporal duration (1 second).

We introduce our data augmentation method called Stochastic and Strided Temporal Sampling (SSTS) for the video valence estimation task. Fig. 2 and Alg. 1 show the procedure of SSTS which consists of two steps: stochastic sampling and strided sampling. Firstly, we randomly draw 24 frames from the raw clip whose frame length N varies from 24 to 30 (Alg. 1 line 2). In this way, we not only introduce stochasticity for the temporal sampling but also tackle the issue of the variable length of the input clip (as VideoMAE requires a fixed length of input frames). Subsequently, we propose the strided sampling by introducing the stride parameter s. Specifically, we select the starting frame index i at random and then sample the remaining frames with the stride s, forming the training sample with frame index $[i, i+s, i+2s...]$ (Alg. 1 line 4). Again, stochasticity is involved in the strided sampling step. For example, in Fig. 2, the orange frames correspond to a training sample when $i = 0$ and $s = 6$ and the green frames form a training sample with $i = 1$ and $s = 6$.

Algorithm 1 Stochastic and Strided Temporal Sampling

1: **Initialize** frame_ind $= [0, 1, ..., N]$
2: inds $=$ np.random.choice(frame_ind, 24, replace=False)
3: $i \sim [0, 1, ..., s-1]$
4: selected_inds $=$ range(i, 24, s)
5: final_inds $=$ inds[selected_inds]
6: **return** final_inds

We emphasize that the main advantages of SSTS are that the training cost is tremendously reduced for video transformers and s introduces flexibility to suit different dataset volumes and model complexities. More details are given in Section IV-B and Section IV-A. Except for SSTS, we also apply RandAugment [58] to each frame, where color jittering, rotation, shearing, and brightness are randomly and jointly applied.

C. Network

We use VideoMAE [16] as the network since it is an efficient feature learner and outperforms other modern video transformers [38]–[40] and traditional 3D ConvNets [41]–[43] in the tasks of video classification. The overall architecture of our infant valence estimation system is shown in Fig. 3. Given the SSTS-processed training clip, we first divide each frame into 16×16 patches and encode each patch into tokens. Then, self-attention is implemented among each patch token, modelling the spatiotemporal dependencies. Following VideoMAE [16], we use ViT-small [59] as the transformer encoder, thus the spatial-temporal self-attention mechanism in our model incorporates that in ViViT [38]. Note that the computation increases quadratically with the number of patches. Thereby, the length of input frames matters in both the training and inference stages and this hyperparameter should be selected carefully. Finally, the encoder outputs the valence by mapping the latent features to the valence space with a fully-connected layer.

D. Implementation Details

In this chapter, we describe the main hyperparameters we used for experiments. Since training the video transformer from scratch is difficult and requires an excessive amount of training data, we consider finetuning the pre-trained VideoMAE to estimate the valence. Concretely, we use VideoMAE pre-trained on the SSV2 dataset [20] and finetune the model until the convergence (max. 100 epochs). The size of the input clip is $[24/s, 3, 224, 224]$ where $24/s$ denotes the number of frames and 224 determines the height and width of the clip. All experiments are carried out with a batch size of 64 and a learning rate of 0.0005 using a single Nvidia A100 GPU on Pytorch 1.11 platform. Also, we apply exponential moving average (0.999) and weight decay (0.05) [60] in the training stage. Finally, the VideoMAE, denoted by $f_\theta(\cdot)$, is trained with the mean squared error (MSE) between the predictions and targets:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2 \tag{1}$$

E. Evaluation Metric

Following [14], the main metric we use for evaluating the valence estimation performance is Concordance Correlation Coefficient (CCC) [61] which measures the agreement between the targets and network predictions. The value of CCC

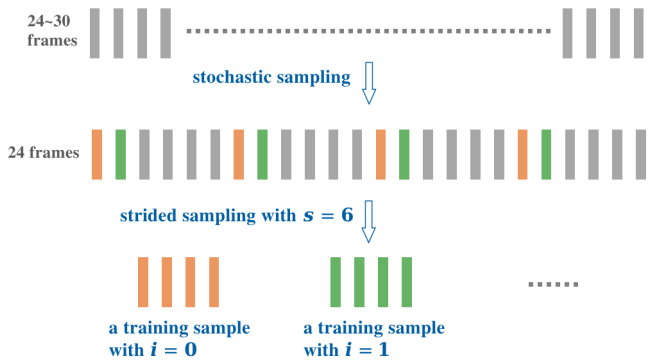


Fig. 2. A diagram of our Stochastic and Strided Temporal Sampling. Figure best viewed in color.

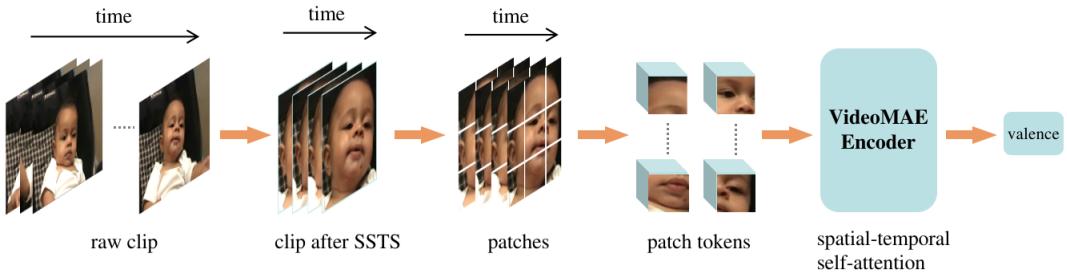


Fig. 3. The architecture of our infant valence estimation system

ranges in $[-1, 1]$ and the upper bound indicates the best performance. CCC is computed by:

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{2}$$

where s_x and s_y correspond to the variances of the target valence x and predicted valence y, respectively. $\bar{x}$ and $\bar{y}$ are the means and s_{xy} is the covariance between x and y. Note that, CCC is substantively identical to an absolute agreement ICC. As an auxiliary criterion, the root-mean-square error (RMSE) is also used to measure the average Euclidean distance between x and y. The upper bound of RMSE is 200 in this paper as the infant valence varies in $[-100, 100]$. We notice that the variance of each evaluated CCC/RMSE is small and can be ignored. Thus we only report a single value of these two metrics. The majority of the experiments are conducted on the composite dataset SIBSMILE + MODELING, since training the video transformer is data-hungry.

IV. Results

A. Efficient Training with SSTS

We analyze the effect of our proposed SSTS under different stride s on the composite dataset SIBSMILE + MODELING which contains 12482 training clips and 4002 testing clips. We train all models until convergence (max. 100 epochs) and compute both CCC and RMSE on the testing set. Note that $s = 1$ means using the entire 24 frames for training, which is regarded as the baseline. We refer to $s \geqslant 2$ as the groups trained with our SSTS method.

The results in Table II show that all models (except for $s = 12$ with cropping) trained with SSTS outperform the baseline in both CCC and RMSE metrics, regardless of the pre-processing method being coarse or fine registration. Compared with the baseline $s = 1$, using SSTS with $s = 6$ (i.e. the number of input frames is 4) not only obtains the best CCC and RMSE, but also accelerates the training stage by 8x, showing significant training efficiency. When s reaches 12, SSTS can even speed up the training by 11x while maintaining a competitive CCC and RMSE.

We argue the reason why $s = 6$ achieves the best performance on infant valence prediction is that the input of 6 frames best matches the temporal redundancy in the dataset SIBSMILE + MODELING. Intuitively, the more redundancy of the video clip, the larger s we can use. Given the fixed model complexity, a big s would bring about the information loss in the time dimension, and a small s will result in difficulties in model training or underfitting. This observation leads to another good property of our SSTS method: it can be adapted to different datasets and networks by simply searching for the optimal s within a small range. However, for facial valence estimation tasks on videos in the 24-30 fps range, we expect $s = 6$ to work well.

B. Ensemble Inference with SSTS

In this section, we discuss ensemble inference for video valence estimation. Similar to the concept of Ensemble Learning [62], we reduce the prediction variance by averaging the multiple predictions when using the strided temporal sampling method (SSTS). More specifically, we receive s predictions $[y_1, y_2, ..., y_s]$ given a testing clip. Then, the final prediction $\hat{y}$ is computed by $\hat{y} = \sum_{i=1}^{s} y_i/s$. All the results in Table II are computed by this ensemble inference strategy.

We now compare the difference between ensemble inference and single inference, in which the latter means $\hat{y}$ is a randomly selected sample in the prediction sequence $[y_1, y_2, ..., y_s]$. We measure the performance of these two inference methods on the dataset SIBSMILE, MODELING and SIBSMILE + MODELING with $s = 6$ in all experiments. The results are presented in Table III, from which we can see that ensemble inference surpasses single inference in all three datasets and in both coarse and fine registration pre-processing settings. Especially in the dataset MODELING processed by coarse registration, ensemble inference leads to the most CCC gain, where the CCC increases from 0.683 to 0.732. The RMSE is also improved by ensemble inference in all experimental settings.

It is worth pointing out that learning the infant valence on MODELING dataset is easier than the learning on SIBSMILE dataset since the former has a larger resolution and a higher frequency of capturing the frontal faces of babies. This explains the performance gap on the dataset SIBSMILE and MODELING in Table III.

C. Stochasticity of SSTS

In our proposed SSTS method, both *stochastic sampling* and *strided sampling* steps involve stochasticity (see Fig. 2). To analyze the effect of randomness in SSTS, we design the SSTS

TABLE II
OBTAINED CONCORDANCE CORRELATION COEFFICIENT (CCC), ROOT MEAN SQUARE ERROR (RMSE) AND TRAINING TIME ON SIBSMILE + MODELING USING SSTS WITH DIFFERENT STRIDES (S)

Stride (s)	s=1 (baseline)		s=2		s=3		s=4		s=6		s=8		s=12	
	Coarse	Fine	Coarse	Fine	Coarse	Fine	Coarse	Fine	Coarse	Fine	Coarse	Fine	Coarse	Fine
CCC	0.630	0.631	0.668	0.647	0.644	0.633	0.648	0.643	**0.671**	**0.667**	0.635	0.667	0.621	0.659
RMSE	25.35	25.33	24.09	24.43	25.02	24.44	25.04	24.31	**23.68**	**23.86**	24.98	24.23	25.45	23.95
Training time (min)	946		423		292		223		160		123		85	

TABLE III
CONCORDANCE CORRELATION COEFFICIENT (CCC) AND ROOT MEAN SQUARE ERROR (RMSE) WITH SINGLE AND ENSEMBLE INFERENCE

		Coarse Registration		Fine Registration	
		Single	Ensemble	Single	Ensemble
SIBSMILE	CCC	0.605	**0.61**	0.633	**0.643**
	RMSE	30.21	**30.08**	28.55	**28.41**
MODELING	CCC	0.683	**0.732**	0.712	**0.724**
	RMSE	17.08	**16.14**	16.52	**16.48**
SIBSMILE + MODELING	CCC	0.633	**0.671**	0.663	**0.667**
	RMSE	25.94	**23.68**	24.27	**23.86**

without stochasticity by always selecting the first 24 frames in the stochastic sampling step and fixing the starting frame as 1 in the strided sampling step. Then we compare its performance with standard SSTS on the benchmark SIBSMILE + MODELING with $s = 6$. The results of Table IV highlight the importance of stochasticity of SSTS as the CCC drops significantly from 0.671 to 0.619 in the coarse registration group and drops from 0.676 to 0.653 in the fine registration group. In addition, randomness can greatly enhance the diversity of training samples, so the coarse registration group, which relies more heavily on diversity, drops more than the fine registration group after losing stochasticity.

D. Coarse vs. Fine Face Registration

Coarse and fine face registration are the two commonly used pre-processing methods for facial expression recognition. We now analyze their effect on infant valence estimation. Table II summarizes the performances of these under various strides of SSTS. When the stride $s \leqslant 6$, coarse registration generally achieves better results than fine registration. The biggest gap occurs in the group of $s = 2$ where the CCC of coarse registration is 0.668 and fine registration receives 0.647 of CCC. On the contrary, fine registration outperforms coarse registration in the settings of $s = 8$ and $s = 12$ (i.e. the number of input frames is 3 and 2, respectively). Our explanation for this phenomenon is that face cropping introduces more spatial variations, which help learn a robust network but require a relatively large number of input frames. On the other hand, landmark-based alignment constrains the position of face features learned by the network, hence the network can be trained using a small number of input frames, but at the risk of being sensitive to the adversarial samples.

Another reason why fine registration performs worse than coarse registration (when $s \leqslant 6$) is that the accuracy of fine registration relies on the performance of the face landmark detection model. In this paper, we use the face alignment network trained on adult faces to get infant face landmarks due to the lack of public infant face alignment datasets or models validated on large infant datasets. Based on the observation that our employed face landmark model makes poor predictions occasionally, we believe that a custom infant face landmark detector would contribute to the performance of face registration and valence estimation to some extent. However, considering that annotating the infant face landmarks and training the network are expensive, coarse registration is a good alternative to fine registration in the scenario of infant valence estimation.

TABLE IV
CONCORDANCE CORRELATION COEFFICIENT (CCC) AND ROOT MEAN SQUARE ERROR (RMSE) USING WITH SSTS AND WITHOUT STOCHASTICITY

	Coarse Registration		Fine Registration	
	SSTS	No Stochasticity	SSTS	No Stochasticity
CCC	**0.671**	0.619	**0.676**	0.653
RMSE	**23.68**	25.63	**23.86**	24.29

E. Pre-training Is Necessary

To investigate the importance of pre-training for VideoMAE in the downstream tasks, we also train VideoMAE from scratch with $s = 6$ on the SIBSMILE + MODELING dataset and compare its performance with the model pre-trained on the SSV2 dataset [20]. The results on Table V demonstrate that the video transformer without pre-training on large datasets fails to learn the infant valence by showing -0.007 and -0.004 CCC (the zero CCC means no correlation between predictions and targets). It is worth pointing out that the large dataset SSV2 contains generic human action clips (e.g., playing basketball) and most videos do not involve human faces. Even under this great task difference, pre-training on a large-scale dataset is still vital in infant valence estimation. We believe that pre-training VideoMAE on large face datasets could further improve the performance of infant valence prediction and it is an interesting direction for future work.

F. Generalization

To test the generalization of the VideoMAE trained on infant valence estimation task, we implement the cross-domain

TABLE V
CONCORDANCE CORRELATION COEFFICIENT (CCC) AND ROOT MEAN SQUARE ERROR (RMSE) WITH AND WITHOUT PRE-TRAINING

	Coarse Registration		Fine Registration	
	No pre-training	Pre-training	No pre-training	Pre-training
CCC	-0.007	0.671	-0.004	0.667
RMSE	33.78	23.68	33.61	23.86

TABLE VI
OBTAINED CONCORDANCE CORRELATION COEFFICIENT (CCC) FOR CROSS-DOMAIN GENERALIZABILITY

	Coarse Registration		Fine Registration	
train on $\Rightarrow$ test on $\Downarrow$	SIBSMILE	MODELING	SIBSMILE	MODELING
SIBSMILE	–	0.193	–	0.089
MODELING	0.577	–	0.746	–

training and inference based on SIBSMILE and MODELING datasets. Table VI presents the CCC results under stride $s = 6$. It is clear that the model trained on SIBSMILE performs well on the MODELING dataset with a 0.577 CCC score using coarse registration and a 0.746 CCC score via fine registration. By contrast, the testing on SIBSMILE using the model trained on MODELING is unsatisfactory. There can be two reasons for this phenomenon: First, SIBSMILE has more data samples than MODELING. Secondly, the head pose in SIBSMILE is more diverse than that in MODELING. Models trained with small databases containing mostly frontal faces cannot generalize well to more in-the-wild databases.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we investigate the infant valence estimation problem and demonstrate the effectiveness of using a pre-trained VideoMAE in this problem. Correspondingly, we propose the temporal augmentation method SSTS to speed up the network training and improve the valence estimation performance. Finally, we suggest using coarse registration via face cropping in the pre-processing stage when an accurate face landmark detection model is not available. Regarding future work, the impact of the precise infant face registration model and the model pre-training using face data on infant expression recognition are two questions worth studying.

ETHICAL IMPACT STATEMENT

The study was IRB approved. We obtained permission from the guardians of the babies to conduct scientific research using the videos of infants. We also acquired consent from the mother of the infant whose face is visible in this manuscript to use their images for publication. However, in order to protect the user privacy of the subjects, we can not open source the datasets of SIBSMILE and MODELING, including the checkpoints of the models trained with these two datasets.

Data used to train our models contains videos of 6-month-old infants. As this is the first study to automatically estimate valence in infants, it serves as a proof-of-concept. If it is implemented on a larger scale, caution should be exercised to test it on more diverse samples of different age groups.

ACKNOWLEDGMENT

This research was supported in part by NIH award R01MH096951 and National Science Foundation award 1052736.

REFERENCES

[1] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Recognition of genuine smiles," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 279–294, 2015.

[2] D. Kollias, A. Tagaris, and A. Stafylopatis, "On line emotion detection using retrainable deep neural networks," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–8.

[3] I. O. Ertugrul, J. F. Cohn, L. A. Jeni, Z. Zhang, L. Yin, and Q. Ji, "Cross-domain au detection: Domains, learning approaches, and measures," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[4] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–8.

[5] L. Yang, I. O. Ertugrul, J. F. Cohn, Z. Hammal, D. Jiang, and H. Sahli, "Facs3d-net: 3d convolution based spatiotemporal representation for action unit detection," in *2019 8th International conference on affective computing and intelligent interaction (ACII)*. IEEE, 2019, pp. 538–544.

[6] J. F. Cohn, L. A. Jeni, I. Onal Ertugrul, D. Malone, M. S. Okun, D. Borton, and W. K. Goodman, "Automated affect detection in deep brain stimulation for obsessive-compulsive disorder: A pilot study," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 40–44.

[7] N. Rawal and R. M. Stock-Homburg, "Facial emotion expressions in human–robot interaction: A survey," *International Journal of Social Robotics*, vol. 14, no. 7, pp. 1583–1604, 2022.

[8] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang, and H. Yu, "Realistic facial expression reconstruction for vr hmd users," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 730–743, 2019.

[9] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[10] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[11] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Face behavior a la carte: Expressions, affect and action units in a single network," *arXiv preprint arXiv:1910.11111*, 2019.

[12] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien, "Fatauva-net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 17–25.

[13] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: valence and arousal'in-the-wild'challenge," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 34–41.

[14] D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018.

[15] I. Onal Ertugrul, L. Yang, L. A. Jeni, and J. F. Cohn, "D-pattnet: Dynamic patch-attentive deep network for action unit detection," *Frontiers in computer science*, vol. 1, p. 11, 2019.

[16] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *arXiv preprint arXiv:2203.12602*, 2022.

[17] L. Wang, Z. Tong, B. Ji, and G. Wu, "Tdn: Temporal difference networks for efficient action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1895–1904.

[18] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[19] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[20] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[21] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[22] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[23] I. Onal Ertugrul, Y. A. Ahn, M. Bilalpur, D. S. Messinger, M. L. Speltz, and J. F. Cohn, "Infant afar: Automated facial action recognition in infants," *Behavior research methods*, pp. 1–12, 2022.

[24] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[25] T. Dalgleish and M. Power, *Handbook of cognition and emotion*. John Wiley & Sons, 2000.

[26] P. Ekman and W. V. Friesen, "Facial action coding system: a technique for the measurement of facial movement," 1978.

[27] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "Eac-net: Deep nets with enhancing and cropping for facial action unit detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2583–2596, 2018.

[28] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition—tangent distance and tangent propagation," in *Neural networks: tricks of the trade*. Springer, 2002, pp. 239–274.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[34] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[35] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, "Vicinal risk minimization," *Advances in neural information processing systems*, vol. 13, 2000.

[36] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[37] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 2020, pp. 566–583.

[38] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[39] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6824–6835.

[40] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[42] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[43] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[46] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653–9663.

[47] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[48] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[49] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[50] M. Chen, S.-M. Chow, Z. Hammal, D. S. Messinger, and J. F. Cohn, "A person-and time-varying vector autoregressive model to capture interactive infant-mother head movement dynamics," *Multivariate behavioral research*, vol. 56, no. 5, pp. 739–767, 2021.

[51] Z. Hammal, J. F. Cohn, and D. S. Messinger, "Head movement dynamics during play and perturbed mother-infant interaction," *IEEE transactions on affective computing*, vol. 6, no. 4, pp. 361–370, 2015.

[52] L. B. Adamson and J. E. Frick, "The still face: A history of a shared experimental paradigm," *Infancy*, vol. 4, no. 4, pp. 451–473, 2003.

[53] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1021–1030.

[54] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 397–403.

[55] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2879–2886.

[56] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[57] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 3, pp. 436–450, 2012.

[58] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[60] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *Advances in neural information processing systems*, vol. 4, 1991.

[61] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[62] L. Rokach, "Ensemble-based classifiers," *Artificial intelligence review*, vol. 33, pp. 1–39, 2010.