Towards the Synthesis of Parent-Infant Facial Interactions

Renke Wang[1], Yeo Jin Amy Ahn[2],
Daniel Messinger[2], Ifeoma Nwogu[1,3]
[1] Department of Computer Science, Rochester Institute of Technology
[2] University of Miami
[3] University at Buffalo, SUNY

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667019

***Abstract*— This work is motivated by the need to automate the analysis of parent-infant interactions to better understand the existence of any potential behavioral patterns useful for the early diagnosis of autism spectrum disorder (ASD). It presents an approach for synthesizing the facial expression exchanges that occur during parent-infant interactions. This is accomplished by developing a novel approach that uses landmarks when synthesizing changing facial expressions. The proposed model consists of two components: (i) The first is a landmark converter that receives a set of facial landmarks and the target emotion as input and outputs a set of new landmarks transformed to match the emotion. (ii) The second component involves an image converter that takes in an input image, a target landmark and a target emotion and outputs a face transformed to match the input emotion. The inclusion of landmarks in the generation process proves useful in the generation of baby facial expressions; babies have somewhat different facial musculature and facial dynamics than adults. This paper presents a realistic-looking matrix of changing facial expressions sampled from a 2-D emotion continuum (valence and arousal) and displays successfully transferred facial expressions from real-life mother-infant dyads to novel ones.**

I. Introduction

This paper describes the first step in a larger study to develop a generative model for better understanding and synthesizing videos of parent-infant interactions. In recent years, awareness of the intricacies of parent-infant interactions in the first few months of life has become relevant in monitoring the mental health of both mother and infant [3]. Developmental disorders such as autism spectrum disorder (ASD) can be detected from brain and behavioral patterns presented in infants as early as 6 months of age [32]. However, diagnostic behavioral features for the disorder are not presented until children are 2 years of age or older[2]. More recently, child psychologists have examined the behaviors of "infant siblings of children with ASD", starting as early as 6 months, to learn more about potential behavioral patterns useful for early diagnosis[2]. The ability to successfully diagnose ASD early allows earlier intervention, which can reduce the negative effects of the disorder [11] and increase the positive long-term outcomes for the child, with results such as better verbal and overall cognition at school age[10].

This material is based upon work partly supported by the National Science Foundation under Grant No. 1846076.

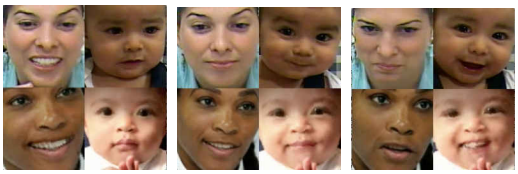

Fig. 1. Three different sets of frames, where the top row shows real images from the infant-sibling dataset; the bottom row shows images synthesized using the valence and arousal estimated from the top frames, and imposed on a sports celebrity with her 6-month-old daughter. More in Section IV-B

Recently, Ahn et al. [1] generated a large number of social communication variables, including patterns of vocal interactions between children and adults, to explore the development of objective measurements in ASD diagnostic and treatment monitoring. In this and many other studies, the number of parent-child pairs was typically approximately 70 or less. Gaining access to families willing and able to participate in studies such as these, cleaning and annotating the resulting data and finally performing analysis on the data can be very tedious and expensive; however the benefits of such studies cannot be overstated. Additionally, privacy concerns often limit the dissemination of such data to computational scientists for more automated studies.

For these and other reasons, our long-term goal is to develop a probabilistic generative technique for understanding and synthesizing new videos of parent-infant interactions through learning from real-life interactions. Beebe 2012 [3] describes how facial expression analysis is a key component of studying mother-infant interactions, as the mothers and their infants often tend to match each other's positive facial expressions until they both "build to a peak of positive facial excitement", along with the many other expression exchanges that happen between them.

Therefore, this work, given time-synchronized videos of 6-month-old infants interacting with a parent, the goal is to extract the facial expressions from an interacting pair and transfer them to a novel parent-infant dyad in a realistic manner. The interaction data used for this work was collected from an infant-sibling ASD study dataset, as described above.

II. Related works

A. Emotional expression representation

A common strategy in human emotion representation is to abstract emotions into discrete categories [15], such as the six commonly used ones: happy, sad, disgust, fear, surprise, and anger[1]. However, using discrete categories might result in the loss of some subtle variations in emotional expressions. For example, emotions can vary in intensity, and several emotions can occur at the same time. For this reason, the use of compound emotions such as happily surprised was proposed [14]. However, this further complicates the expression of emotions and it still does not describe the continuum of emotional expressions. The valance-arousal emotional state model[31] was introduced in the eighties, in an attempt to express emotions in 2D (or 3D) continuous space. Additionally, the facial action coding system (FACS) [16], was developed to decompose facial expressions into combinations of several fixed facial action units, that corresponded to specific musculature movements on the face. It provided one standard system for quantitatively describing facial expressions.

B. Generative adversarial network

Generative adversarial network(GAN)[18] is an effective generative model based on game theory. Currently, there are many different extensions of GANs, and the conditional GAN (CGAN)[24] is one of the more popular variants. CGAN uses conditional information to guide the image generation process; however, it is a supervised model, which requires ground truth target images in the training process.

CycleGAN [41], DualGAN [40] and DiscoGAN [20] sidestep this issue by utilizing a cycle consistency loss. The cycle consistency loss reconstructs the generated image back to the input image and lets the reconstructed image and original image be as similar as possible. This helps the model reduce the space of possible mapping functions and thus enables the transfer of an image from one domain to another with unpaired data. Based on this, Choi et al. [7] proposed training models that support image transfer between multiple domains. This laid the foundation for emotion-based GANs.

C. Facial image manipulation

Facial expression generation techniques have been substantially explored. Many earlier works either manipulated the facial components of the input image [25], [37], [38] or used the target image as a comparison reference for the model to learn to generate a corresponding image[6], [30], [34].

In [22], the authors fitted a 3D morphable model (3DMM) [4] to an input neutral face image and deformed the reconstructed face into the target emotion. Then, the new face was blended with the original image to obtain the target face image. This strategy easily obtains high-quality synthesized images and can overcome cases of extreme head poses. However, unlike our proposed work, this model requires the input faces to be neutral, which greatly limits its flexibility.

In [13], the authors used an expression controller module to encode emotion into a more expressive vector, allowing the model to generate emotions of different intensities without the need for a target image of corresponding intensity. However, this model was still not able to generate a wide variety of emotions.

Vielzeuf et al. [36] projected each discrete emotion label into 3D space and then discovered the corresponding vector of dominance in the 3D space when given valance and arousal values, thus allowing the model to convert an image expression into any combination of valence-arousal-dominance.

Ganimation [28] was proposed to replace the emotion label with action units in an attempt to improve the smoothness of face animation. However, the problem was that the correspondences between the action units and emotions were often complicated and were measured based on discrete emotions. Again, it is not clear how well the use of a GAN based on action units would perform on baby faces.

D. Keypoint-guided image-to-image translation

Landmarks, as an extremely condensed facial feature representation, can preserve facial information such as the pose, gender, and facial structure. In the gender preserving generative adversarial network (GP-GAN) [12], the authors attempted to restore the original image by using only the landmarks as input. This confirmed that the semantic information contained in landmarks can further be used to generate new images. In [33], the authors used landmarks as a label for face swapping and used a landmark converter first to adjust the landmarks, which helped the expression better fit the target face. The experiments also demonstrated that the landmark changes could be successfully fed back to the generated faces. In [29], the authors used two different variational autoencoder (VAE) structures to encode the landmarks and the original face. They then concatenated the two vectors to generate the target face. However, this still required pairs of matched original and target faces for training. In [35], the authors proposed using landmarks as the label and used cycle consistency separately for landmarks and images. This was to ensure that the generated image retained both the semantic and landmark information of the original image and would not need a target image. However, this still required target ground truth landmarks as labels, thus reducing its flexibility as a completely unsupervised model.

III. Proposed Methodology

The core objective of our model is to transform any input face to express an arbitrary emotion under weak supervision. As shown in Figure 2, our model consists of two parts: 1. the landmark converter which converts the landmark x_l of the input image x_i into the target emotion c_e; and 2. the image converter which takes the converted landmarks y_l and target emotion c_e as labels to convert the input image x_i into the target image y_i.

[1]The emotion contempt is also sometimes included in these discrete categories.

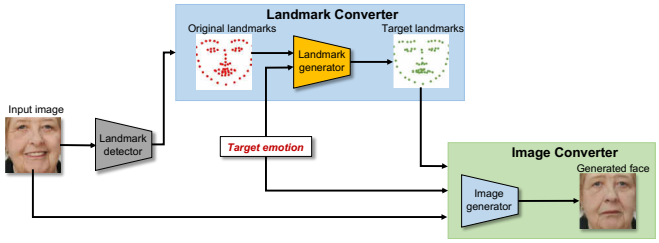

Fig. 2. High-level representation of the proposed avLandmarkGAN model, showing how an emotional target face image is generated from both an input face and target emotion labels via intermediate landmark generation.

A. Landmark converter

The landmark converter has a 2D GAN structure, and the input landmarks are a 2D vector: 2×68, following the landmark format in DLib [21]. The converter consists of a discriminator D_l and a generator G_l. The structure of landmark converter can be found in Figure 3

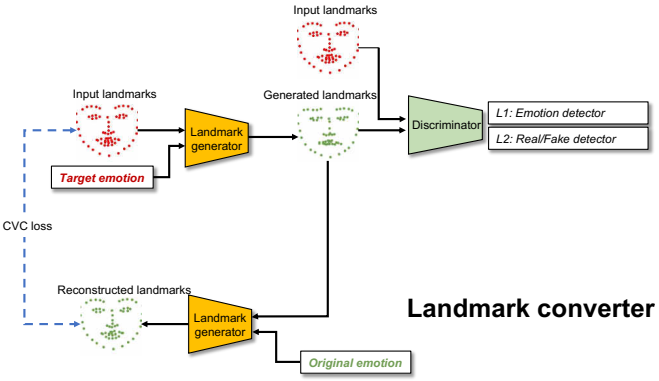

Fig. 3. 1. The landmark generator converts the input landmarks into generated landmarks based on the target emotion. 2. The discriminator will receive both the input landmarks and generated landmarks to distinguish between them and to detect the emotion with such landmarks. 3. The generator uses the generated landmarks and original emotion to generate the reconstructed landmarks and to calculate the cycle consistency loss with the input landmarks

The input landmarks are detected by a pretrained face alignment network (FAN) model [5] and the goal is to convert these landmarks x_l to y_l through conditioning with c_e. To ensure the identity constancy, we use cycle consistency in our model, i.e. we reconstruct y_l back to the x_l on which it was originally based. In addition, to address steganography, the issue of messages being hidden in the latent variables of the model, as discussed in [41], we propose a strategy to calibrate the generated landmarks y_l. The loss function of the landmark converter consists of the following parts:

1) Adversarial loss: Similar to the vanilla adversarial loss [18], the adversarial loss L_{adv} can be formulated as follows:

$$L_{adv} = E_{x_l}\left[\log D_l\left(x_l\right)\right] + E_{x_l,c_e}\left[\log\left(1 - D_l\left(G_l\left(x_l, c_e\right)\right)\right)\right] \tag{1}$$

2) Emotion loss: This is similar to the classification loss in StarGAN [7], but we use an L2 regression loss instead, because the emotion labels we use consist of two continuous values. The emotion loss function consists of two terms. The first term L_e^{real} is to force D_l correctly predict the emotion c_e' of a real image. The second term L_e^{fake} forces G_l to generate more suitable faces. Thus, the emotion loss is:

$$\begin{aligned} L_e^{real} &= E_{x_l,c_e'}\left[\left\|D_l\left(x_l\right) - c_e'\right\|_2\right] \\ L_e^{fake} &= E_{x_l,c_e}\left[\left\|D_l\left(G_l\left(x_l, c_e\right)\right) - c_e\right\|_2\right] \end{aligned} \tag{2}$$

3) Reconstruction loss: Following the cycle consistency paradigm, we send y_l back to the generator and use the original emotion label c_e' to reconstruct a similar landmark x_l', which is supposed to be similar to x_l. Hence, the reconstruction loss L_{rec} can be formulated as:

$$L_{rec} = E_{x_l,c_e,c_e'}\left[\left\|x_l - G_l\left(G_l\left(x_l, c_e\right), c_e'\right)\right\|_1\right] \tag{3}$$

4) Face feature center loss: In our initial experiments, we find that even though we have L_{rec}, y_l usually still has errors. For example, y_l gives pose directions opposite to and eye positions different from those of x_l. This is because the capacity/size of the landmark vector is small. Compared to 128*128*3 images, landmarks have only 2*68 points, so they are more likely to encounter steganography issues. This issue occurs here when the generator uses tricks to hide certain information and then attempts to recover it in the reconstruction step.

To address this, we propose the center of facial feature loss L_{ffc}. This is the novel loss that we introduce in this work. In L_{ffc} the landmarks are split into 7 groups: face shape, left eye, right eye, left eyebrow, right eyebrow, nose and mouth. We then calculate the group centers, which are represented by h_j. The group centers are expected to be constant for all the different emotions. Therefore, the loss function is as follows:

$$L_{ffc} = E_{x_l,c_e}\left[\sum_{j=1}^{7}\left\|h_j\left(x_l\right) - h_i\left(G_l\left(x_l, c_e\right)\right)\right\|_1\right] \tag{4}$$

5) Full objective: Finally, the complete loss function for D_l and G_l is:

$$\begin{aligned} L_{D_l} &= -L_{adv} + \lambda_e^l L_e^{real} \\ L_{G_l} &= L_{adv} + \lambda_e^l L_e^{fake} + \lambda_{rec} L_{rec} + \lambda_{ffc} L_{ffc} \end{aligned} \tag{5}$$

B. Image converter

Similar to the landmark converter, the image converter is a cycle consistency based GAN. The input is the image, and the image converter generates a new image conditioned on both the landmarks and a target emotion.

The image converter receives two sets of labels: 1. The landmark label c_l, which is the landmark y_l generated from the landmark converter. It helps the model to understand the outline of the target expression. 2. The emotion label c_e, which is the same as that used in landmark converter. It complements the details not provided in the landmark, such as nasolabial folds, teeth, etc.

The structure of the image converter can be found in Figure 4, which is similar to that of the landmark converter. Its loss function can be divided into five parts:

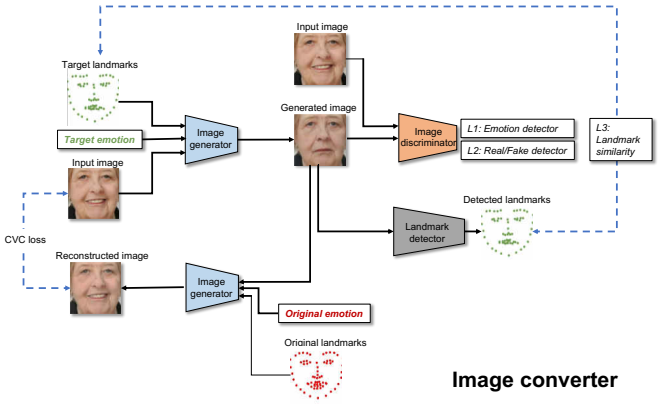

Fig. 4. 1. The image generator uses the converted landmarks as its target landmarks and associates with the target emotion to convert the input image. 2. The discriminator receives both the input image and the generated image to distinguish them and to detect the emotion in the generated image. 3. The generator reconstructs a generated image with the original emotion that should be similar to the input image.

1) Adversarial loss: Similar to the landmark converter, discriminator D_i and generator G_i play the minmax game by optimizing the adversarial loss L_{adv}^{img}:

$$\begin{aligned} L_{adv}^{img} =& E_{x_i}[\log D_i(x)] \\ &+ E_{x_i,c_e,c_l}\left[\log\left(1 - D_i\left(G_i\left(x_i, c_e, c_l\right)\right)\right)\right] \end{aligned} \tag{6}$$

2) Emotion loss: This is same as the loss described in the landmark converter, and can be written as:

$$\begin{aligned} L_e^{real} &= E_{x_i,c_e'}\left[\left\|D_i\left(x_i\right) - c_e'\right\|_2\right] \\ L_e^{fake} &= E_{x_i,c_e,c_l}\left[\left\|D_i\left(G_i\left(x_i, c_e, c_l\right)\right) - c_e\right\|_2\right] \end{aligned} \tag{7}$$

3) Landmark loss: In addition to the emotion, landmarks are also used as conditional labels in the generation process. We use the L1 loss to deal with the sparsity of landmarks' variation.

$$\begin{aligned} L_l^{real} &= E_{x_i,c_l'}\left[\left\|D_i\left(x_i\right) - c_l'\right\|_1\right] \\ L_l^{fake} &= E_{x_i,c_e,c_l}\left[\left\|D_i\left(G_i\left(x_i, c_e, c_l\right)\right) - c_l\right\|_1\right] \end{aligned} \tag{8}$$

4) Reconstruction loss: Similar to the reconstruction loss in landmark converter, the L_{rec} ensures the identity consistency of the generated image by minimizing the following function:

$$L_{rec} = E_{x_i,c_e,c_l,c_e',c_l'}\left[\left\|x_i - G_i\left(G_i\left(x_i, c_e, c_l\right), c_e', c_l'\right)\right\|_1\right] \tag{9}$$

5) Full objective: Finally, the complete loss function for the generator and discriminator in the image converter is:

$$\begin{aligned} L_{D_i} &= -L_{adv} + \lambda_e^i L_e^{real} + \lambda_l L_l^{real} \\ L_{G_i} &= L_{adv} + \lambda_e^i L_e^{fake} + \lambda_l L_l^{fake} + \lambda_{rec} L_{rec} \end{aligned} \tag{10}$$

IV. Implementation and Results

A. Generating Smooth Transitions between Expressions

Both the landmark converter and image converter are trained on the AffectNet dataset [26], which provides both valence and arousal labels for each image. We use the pretrained FAN model to detect landmarks in each image in the dataset and discard all images where either faces are not detected or any of the landmarks are missing. This results in 234,815 images in the training set and 3,159 images in the validation set. To fit our GANs, the faces are resized to 128*128*3.

Because the input of the landmark converter is a sequence of 2*68 coordinates instead of a 3*128*128 array of image pixels, we treat the x- and y-coordinates as two channels to obtain a vector with a length of 136 (2×68) and use this as input. In the image converter, we draw the landmarks on a 128*128*1 white image and then directly concatenate it to the original image as the landmark label.

For the discriminator, we require it to provide the specific values for 136 coordinates, thus ensuring that the generator truly applies the landmarks to the image, rather than simply hiding the coordinates within some vectors.

Training schedule:: The landmark converter and image converter share many hyper-parameters: the batch size is 16; the learning rate is 0.0001 with a linear decay factor of 0.99997 over every epoch; the emotion loss weight $\lambda_e^l = \lambda_e^i = 5$; the landmark loss weight $\lambda_{lm} = 10$; the reconstruction weight $\lambda_{recon} = 10$; the face feature center weight $\lambda_{ffc} = 10$; the arousal and valence vectors are concatenated and flattened into a 1-D vector to calculate the regression loss; and the parameters are optimized with the Adam optimizer and trained for 400,000 iterations.

The weight represents the relative importance of each loss during training. The weights we provide above attempt to adjust each loss to a similar scale and thus ensure that each loss can fully perform its role. To further verify the reliability of our model, we perform ablation experiments for all the other weights in the loss function except for λ_{recon}, for which a value has already been suggested in [7], [9]. To show the effect of each loss term, we individually adjusted each weight to show its impact on the model. The Frechet inception distance (FID) [19] and root mean square error (RMSE) are used to determine the performance of our model. The FID score has been popularly used in many different works [8], [23], [39] to indicate the realisticness of the generated image. It was calculated between 10,000 generated images and our training set. To calculate the RMSE, we first train a RESNET-18 to predict valence and arousal using AffectNet, and use this to predict the emotions of all generated images. We then calculate the RMSE between the predicted emotions and real emotions. RMSE therefore reflects the extent of similarity between the generated emotional values are to the real ones.

As shown in Table I, the model is very sensitive to λ_e in the image converter. Too small of a weight leads to a sharp decrease in accuracy, while too high leads to a sharp decrease in realisticness. The other parameters are less sensitive. In our model, the landmarks and images used are 2-dimensional and do not take into account the 3-dimensional spatial structure of the face. Therefore some potentially useful information may be lost. To explore further, we replaced

TABLE I

ABLATION STUDY SHOWING MODEL PERFORMANCE WITH DIFFERENT LOSS FUNCTIONS AND CORRESPONDING WEIGHTS.

weight	*metrics*	λ_e^l	λ_{ffc}	λ_e^i	λ_l
0	FID	25.66	25.80	20.92	25.50
	RMSE	0.188	0.199	0.56	0.200
0.1	FID	24.40	26.43	20.84	25.97
	RMSE	0.204	0.205	0.56	0.192
1	FID	24.24	24.54	21.89	25.58
	RMSE	0.193	0.202	0.31	0.208
5	FID	**22.13**	**22.13**	**22.13**	25.33
	RMSE	**0.200**	**0.200**	**0.200**	0.208
10	FID	26.25	24.16	26.59	**22.13**
	RMSE	0.203	0.192	0.177	**0.200**
100	FID	24.02	24.84	64.92	23.44
	RMSE	0.202	0.198	0.178	0.198

all the landmarks in the model with 3D ones. However, from our experimental results, the best performance of our model using 3D points was $FID = 26.7$ and $RMSE = 0.211$, demonstrating that 3D points do not improve model performance.

To qualitatively demonstrate the continuous transitions between expressions, we generate *avImage matrices* as in Figure 5. Additional image matrices are presented in the supplementary material.

To validate the output of our proposed end-to-end avLandmarkGAN, since AffetNet provides both discrete and continuous labels for each face, we computed the valence and arousal averages of each of the 8 discrete emotions for the faces, and the results are presented in Table IV-A. We then synthesized expressive images using these arousal and valence values, and the results are shown in Figure 6.

TABLE II

AVERAGES OF CONTINUOUS EMOTIONS FOR EACH DISCRETE EMOTION LABEL, USED TO GENERATE THE IMAGES IN FIGURE 6.

	Valence	**Arousal**
Neutral	0.0043	0.0040
Happy	0.6652	0.1342
Sad	-0.6864	-0.2585
Surprise	0.2306	0.6812
Fear	-0.1156	0.7741
Disgust	-0.7312	0.4342
Anger	-0.3646	0.6604
Contempt	-0.5548	0.5880

Observations: From the images shown in Figure 5, we demonstrate that the proposed avLandmarkGAN can successfully generate new emotionally expressive images on a 2-D valence-arousal continuum. By computing the average valence and arousal values for the corresponding discrete labels, we successfully generated faces that approximately match the discrete emotion labels. Although successful for emotions such as happy, sad, surprise and fear, it is challenging for the model to differentiate between disgust (typically characterized by AU_9[2], the nose wrinkler + others), anger (often characterized by AU_4, the brow lowerer + others) and contempt (characterized by AU_12 or 14, the lip corner puller or the dimpler, expressed strictly on only one side of the face). There are not enough differentiating examples of these 3 labels in the training dataset; hence, additional data is required to better cover the spectrum.

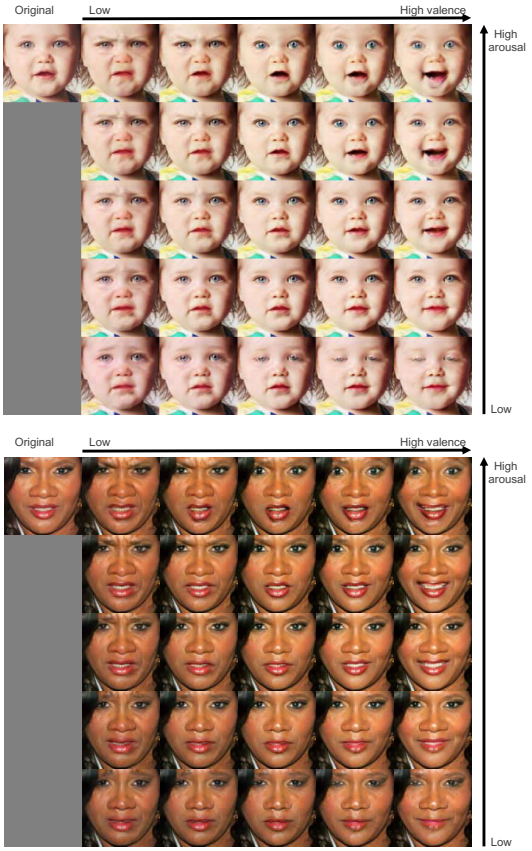

Fig. 5. Two avImage matrices generated by avLandmarkGAN with continuous arousal and valence labels imposed on an original face from AffectNet (top left)

B. Effects of Landmarks on the Generation of Baby Faces

To quantitatively test the effects of adding the landmark converter to our architecture, specifically those on the baby faces in our dataset (described in more detail in Section IV-C), we do the following:

1) We implemented the avGAN[36], which is an architecture similar to ours, but uses a different discriminator architecture and does not implement landmarks.
2) In the interest of fairness, we developed a classifier similar to the discriminator used in the avGAN based on the ResNet-18 architecture.
3) We then selected 49,000 frames from our *infant-sibling* dataset. The dataset is annotated only with valence labels. The selected frames contain frontal faces, and the

[2]AU is the abbreviation of action unit. Each AU represents the contractions of specific facial muscles and is used in the facial action coding system (FACS) [17].

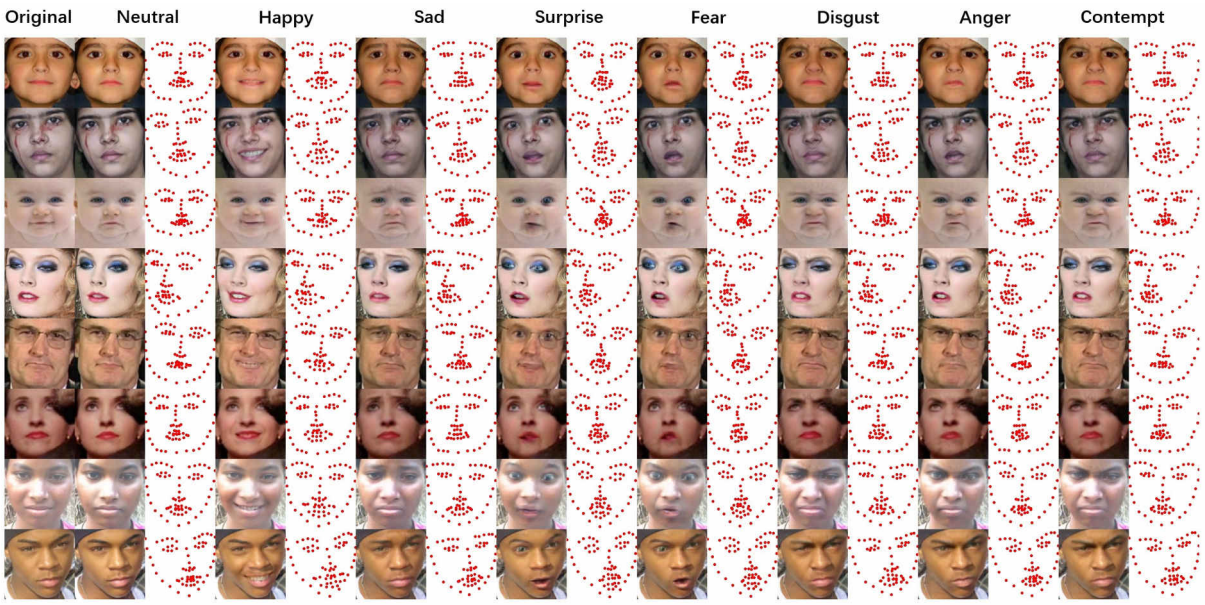

Fig. 6. Image showing the results of generating discrete emotions using AffectNet valence and arousal labels. The first columns show the original image and the next sets of 2 columns show face images and landmarks synthesized with the converted emotion labels. The conversion values are shown in Table $IV - A$

.

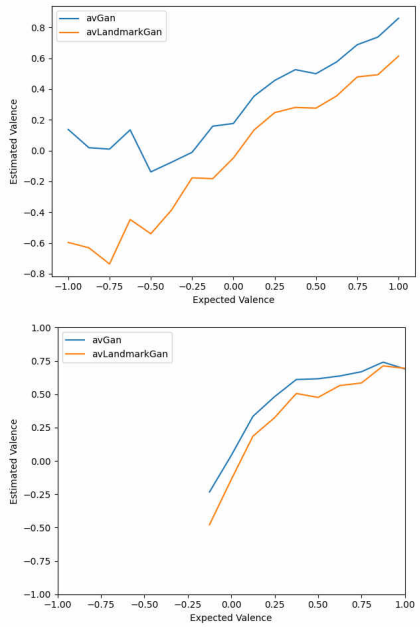

Fig. 7. Plots comparing avGAN and avLandmarkGAN showing the estimated (predicted) versus the expected (labeled) valence values for baby (top) and mom (bottom) faces from the *infant-sibling* dataset. NOTE: Moms collectively did not display negative affect in the experiment.

standard deviation on valence labels between annotators was not greater than 0.2.

4) For each frame, frame #1 is presented to both GANs as input, along with the target valence label (annotated in the dataset). The ResNet-18-based classifier estimated the valence of the 2 resulting synthetic images (one from avGAN and the other from avLandmarkGAN).
5) The correlation factors between the estimated (or predicted) versus the annotated (or expected) valence values on the synthesized baby faces are **avGAN=0.49 and avLandmarkGAN=0.70**. On the synthesized mother faces, the correlation values are avGAN=0.64 and avLandmarkGAN=0.65.
6) The plots of estimated versus expected valence labels on the baby and mom faces from the dataset are shown in Figure 7.

Observations: Including landmark generation as an intermediary step during facial expression synthesis leads to improvements in the expressions generated on baby faces. Using the expert annotations provided, we note that the avGAN without landmarks performs as well as our proposed avLandmarkGAN on adult mother faces whereas it performs almost 30% worse on baby faces.

This could be related to the findings of studies by Oster who observed significant differences between the facial structures and dynamics of adults and babies [27], and this led to the development of *Baby FACS*, a system specifically dedicated to coding the facial movements on babies' faces.

C. Results of Transferring Parent-Infant Emotions

a) Dataset: The *Infant-Sibling* dataset is collected in a still-face experiment involving an infant and a parent facing each other. The experiment consists of 3 phases: (i) the parent interacts very positively and animatedly with the infant; (ii) the mother turns away and maintains a "still face" and remains unresponsive; and finally, (iii) the repair or reunion phase, where the mother faces the infant and resumes play as before. We focus on the emotional reactions in all 3 phases.

The dataset provided to us contains 60 dyadic pairs of

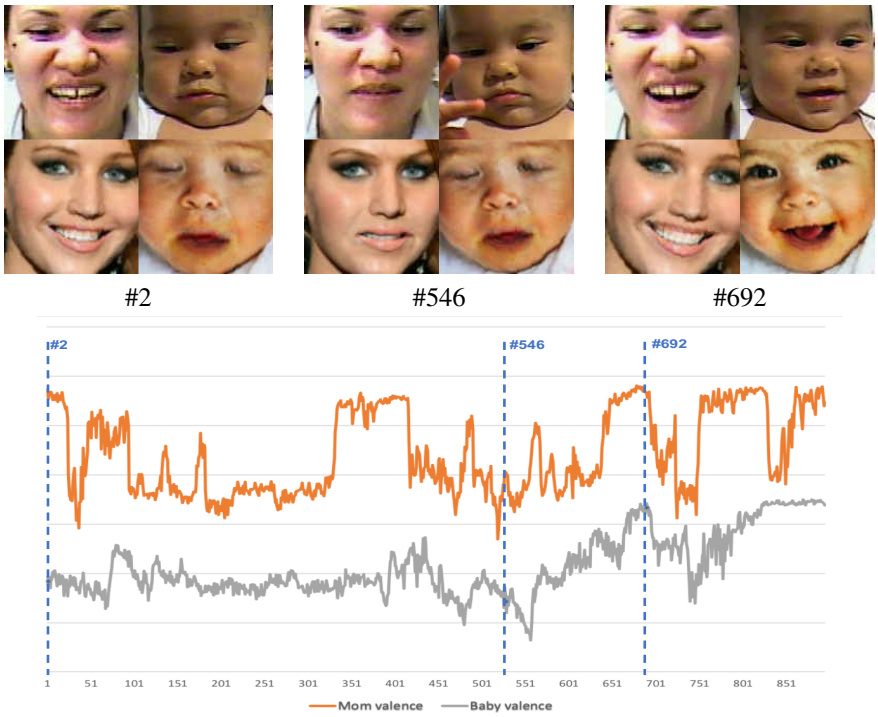

Fig. 8. Three different sets of frames where the top row shows real images from the infant-sibling dataset and the bottom row shows the emotions transferred onto an adult face and a baby face obtained from AffectNet.

parents and infants at age six months, with video-recordings of the three phases for each dyad. One camera faces the infant, and another faces the parent. The data from both cameras are already time-aligned. The video recordings are converted to frames at 30 fps and each frame (mother and infant) is annotated for valence on a scale of {-100,100}, which we normalize to {-1,1}.

b) Emotion Transfer: To transfer emotions, we select a video (any arbitrary video) from the provided dataset and predict the valence and arousal values on each frame for the parent and infant using the avLandmarkGAN discriminator. Taking any arbitrary starting image, such as that in Figure 1, where we use an image of the tennis player Serena Williams and her 6-month-old daughter, we use the avLandmarkGAN to synthesize new images of their faces with expressions that match a real-life parent-infant dyad as shown in Figure 1. Another example of transferred emotions is shown in Figure 8. The top row shows the original dyad, and the bottom row shows the emotion-synthesized pair. Frame #2 shows high valence for mother but low arousal from baby; frame #546 shows medium valence from both, but low arousal from the baby, and #692 show very high valence and arousal from both mother and baby. The bottom graph displays the valence values for mother and baby (as predicted by the ResNet classifier), and the extracted frames are highlighted in the graph.

Videos of different parent-infant interactions with their transferred dyads are provided in the supplemental material.

c) Video post-processing: To create a realistic-looking generated video using the synthesized images, we introduced two transformations: rotation and resizing. We first calculate the target rotation angle θ based on current iterations i using the following function:

$$\theta = \frac{\sin(a * i) + \sin(b * i)}{2} * m$$

where $a, b \in [-2, -1] \cup [1, 2])$ and $m \in [0, 10]$; a, b and m are randomly initialized and are reinitialized when the rotation angle becomes zero again. The rotation is also accompanied by a resizing factor to prevent the resulting image from having a black portion (empty pixels after rotation); the resize factor s_r is calculated by the following equation:

$$s_r = \frac{\tan(\theta)}{(1 + \tan(\theta)) * \sin(\theta)}$$

Next, we obtain the resize factor s, also based on the current iteration i, which can be calculated by the function:

$$s = 1 - \frac{abs(\sin(c * i) + \sin(d * i))}{2} * n$$

where $c, d \in [-2, -1] \cup [1, 2])$ and $n \in [0, 0.2]$. We take the smaller value $min(s, s_r)$ to perform the resizing.

The net effect is a smooth and more realistic motion in the synthesized videos. This is applied purely for aesthetic reasons.

d) Observations: We observed that the synthetically generated expressions indeed appear to interact in a similar manner as the original dyads. Paying particular attention to the babies' faces, both the arousal and valence transfer readily to the synthetic images, resulting in natural dynamics such as eyes closure (for low arousal) and lip widening (for high valence).

V. Conclusion and Future Work

In conclusion, we successfully developed a 2-part model that generated realistic-looking faces when provided with an input image and a target emotion. The target emotion can be drawn from a continuous 2-D space.

By integrating landmarks in the model, we successfully generated baby faces whose predicted emotion labels correlate strongly with human annotations, even more so than state-of-the-art models.

As a part of a larger ongoing study in analyzing the behavioral patterns of at-risk infants, we are interested in evaluating a generative model of parent-infant interactions, to understand the types of interactions occurring in different phases of the collection experiment.

References

[1] Y. A. Ahn, J. Moffitt, Y. Tao, S. Custode, M.-L. Shyu, L. Perry, and D. S. Messinger. Objective measurement of social communication behaviors in children with suspected asd during the ados-2. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 360–364, 2020.

[2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Arlington, VA: American Psychiatric Press, 2013.

[3] B. Beebe. Mother–infant research informs mother–infant treatment. *Clinical Social Work Journal*, 38(1):17–36, 2010.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[5] 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*, pages 1021–1030, 2017.

[6] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

[7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[8] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[9] C. Chu, A. Zhmoginov, and M. Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017.

[10] M. L. E. Clark, Z. Vinen, J. Barbaro, and C. Dissanayake. School age outcomes of children diagnosed early and later with autism spectrum disorder. *Journal of autism and developmental disorders*, 48(1):92–102, 2018.

[11] G. Dawson, S. Rogers, J. Munson, M. Smith, J. Winter, J. Greenson, A. Donaldson, and J. Varley. Randomized, controlled trial of an intervention for toddlers with autism: the early start denver model. *Pediatrics*, 125(1):e17–e23, 2010.

[12] X. Di, V. A. Sindagi, and V. M. Patel. Gp-gan: Gender preserving gan for synthesizing faces from landmarks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1079–1084. IEEE, 2018.

[13] H. Ding, K. Sricharan, and R. Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[14] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[15] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

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

[17] E. Friesen and P. Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[20] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017.

[21] D. E. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[22] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *arXiv preprint arXiv:1811.05027*, 2018.

[23] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019.

[24] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[25] U. Mohammed, S. J. Prince, and J. Kautz. Visio-lization: generating novel facial images. *ACM Transactions on Graphics (ToG)*, 28(3):1–8, 2009.

[26] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[27] H. Oster. The repertoire of infant facial expressions: an ontogenetic perspective. 2005.

[28] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018.

[29] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang. Emotional facial expression transfer from a single image via generative adversarial nets. *Computer Animation and Virtual Worlds*, 29(3-4):e1819, 2018.

[30] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. In *International conference on machine learning*, pages 1431–1439. PMLR, 2014.

[31] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[32] M. D. Shen and J. Piven. Brain and behavior development in autism from birth through infancy. *Dialogues in clinical neuroscience*, 19(4):325, 2017.

[33] P. Sun, Y. Li, H. Qi, and S. Lyu. Landmarkgan: Synthesizing faces from landmarks. *arXiv preprint arXiv:2011.00269*, 2020.

[34] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson. Generating facial expressions with deep belief nets. *Affective Computing, Emotion Modelling, Synthesis and Recognition*, pages 421–440, 2008.

[35] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, and Y. Yan. Cycle in cycle generative adversarial networks for keypoint-guided image generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2052–2060, 2019.

[36] V. Vielzeuf, C. Kervadec, S. Pateux, and F. Jurie. The many variations of emotion. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.

[37] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas. Expression flow for 3d-aware face component transfer. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011.

[38] R. Yeh, Z. Liu, D. B. Goldman, and A. Agarwala. Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961*, 2016.

[39] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2019.

[40] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.