Generating Photorealistic Facial Expressions in Dyadic Interactions

Yuchi Huang
https://actnext.org/people/yuchi-huang
Saad Khan
https://actnext.org/people/saad-m-khan

ACTNext
ACT Inc.
IOWA, USA

Abstract

We propose an approach for generating photorealistic facial expressions for multiple virtual identities in dyadic interactions. To this end, we study human-human interactions to model one individual's facial expressions in the reaction of the other. We introduce a two level optimization of generative adversarial networks, wherein the first stage generates one's face shapes conditioned on facial action features derived from their dyadic interaction partner and the second stage synthesizes high quality face images from sketches. A 'layer features' L_1 regularization is employed to enhance the generation quality and an identity-constraint is utilized to ensure appearance distinction between different identities. We demonstrate that our model is effective at generating visually compelling facial expressions. Moreover, we quantitatively showed that generated agent facial expressions reflect valid emotional reactions to behavior of the human partner.

1 Introduction

Human communication involves both verbal and nonverbal ways of making sure our message is heard. A simple smile can indicate our approval of a message, while a scowl might signal displeasure or disagreement [9]. Moreover, the sight of a human face expressing fear elicits fearful responses in the observer, as indexed by increases in autonomic markers of arousal [22] and increased activity in the amygdala [19]. This process whereby an observer tends to unconsciously imitate the behaviour of the person being observed [3] has been shown to impact a variety of interpersonal activities such as collaboration, interviews and negotiations among others [2, 25].

Recent research in autonomous avatars [7, 11] aims to develop powerful human-agent interfaces that mimic such abilities. Not only do these avatar systems sense human behavior holistically using a multitude of sensory modalities, they also aim to embody ecologically valid human gestures, paralinguistics and facial expressions. However, producing realistic facial expressions in virtual characters that are appropriately contextualized and responsive to the interacting human remains a significant challenge. Early work on facial expression synthesis [5] often relied on rule based systems that mapped emotional states to predefined deformation in 2D or 3D face models. Later, statistical tools such as principal component analysis were introduced to model face shapes as a linear combination of prototypical expression basis [4]. A key challenge for such approaches is that the full range of appearance

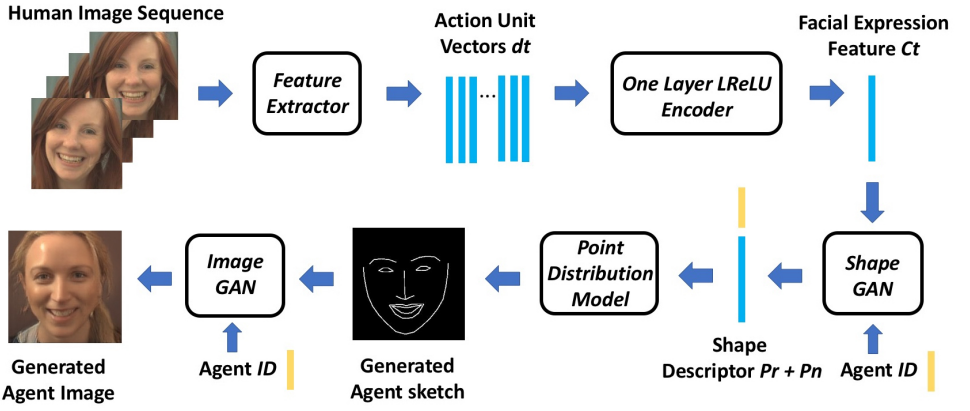

Figure 1: Our framework is composed of two stages of GANs, one to generate affective face sketches and the other to generate photorealistic facial expressions of agents. The inputs are facial action unit vectors d_t extracted from image sequences of human users.

variations required for convincing facial expression is far greater than the variation captured by a limited set of rules and base shapes. Advanced motion capture techniques have also been used to track facial movement of actors and transfer them to avatars [26] recreating highly realistic facial expressions. However, these solutions are not scalable to autonomous systems as they require a human actor in the loop to puppeteer avatar behavior. Recently, deep belief nets [23] and temporal restricted Boltzmann machines [30] were utilized as powerful yet flexible representation tools to model the variation and constraints of facial emotions and to produce convincing expression samples. While these approaches have shown promising results in transferring the same facial expression from one identity to another, they have not purported to model interaction dynamics of multiple person conversations.

This paper tackle the problem of generating photorealistic facial expressions for multiple-identities in human-agent interactions using conditional Generative Adversarial Networks (GAN) [10, 18]. Conditional GANs are generative models that learn a mapping from random noise vector z to output image y conditioned on auxiliary information x: $G : \{x, z\} \rightarrow y$. Previous work based on GANs and conditional GANs [18] has shown promise in a number of applications such as future frame/state prediction [16, 32], video generation [27], image manipulation [33], style transfer [15], text-to-image/image-to-image translation [14, 21] and 3D shape modeling [29]. Our work differs from these in that we do not employ conditions related to facial attributes of generated agent identities, but consider the influence of the interacting human's facial expressions in the virtual agent response. Similar to some recent work such as [12, 13, 24, 31], we developed a two-stage optimization of GANs that enable modeling of complex human behavior as well as compelling photorealisim in the generated agent (see Figure 1). The first stage generates non-rigid shape descriptors p_n of agents conditioned on human's expression. We designed a single layer LReLU (Leaky Rectified Linear) encoder that takes a sequence of action unit vectors [8] (of the interacting human) as input and outputs conditional features, which are in turn used by the *Shape GAN* to generate an intermediate representation, the non-rigid shape descriptors which are used to reconstruct affective face sketches. These sketches have the advantage of being rich enough to capture valid human-like emotions and generic enough to be mapped onto a number of different

agent face identities. The second stage termed Face GAN fleshes out affective face sketches into high-quality photorealistic images. In this stage, a new L_1 regularization term computed from layer features of the discriminator is employed to enhance the quality of generated images and novel identity constraint are proposed to ensure appearance distinction between different identities. We demonstrated that our approach is effective at generating images that depict visually compelling facial expressions. Moreover, we quantitatively showed that agent facial expressions in the generated images reflect valid emotional reactions to behavior of the human partner.

2 Datasets and Feature Extraction

In this study we learn ecologically valid models of human-agent interactions on two datasets: the interview dataset and SEMAINE dataset [17]. Our interview dataset consists of 31 dyadic Skype interviews for undergraduate university admissions. Each prospective student was interviewed by the same interviewer (Caucasian female) who followed a predetermined set of questions designed to gather evidence of the candidate's English speaking ability. For the purpose of this study we treat the interviewer as the only *agent* who's behavior we aim to model as a response to the stimulus i.e. human partner (interviewee). An advantage of this dataset is that it provides a significant amount of data under varying stimuli (31 different interviewees) to adequately model the interviewer's i.e. the agent's behavior with different lighting conditions, backgrounds, outfits and hair styles. SEMAINE dataset [17] contains over 140 dialogues between 4 'operators' (who are persons simulating a machine) and 16 'users' (who are humans). An operator plays one of four personalities and tries to keep the corresponding user engaged: 'Poppy' (happy), 'Spike' (angry), 'Prudence' (sensible) and 'Obadiah' (gloomy), as shown in the first column of Figure 8. one individual can play an operator in one session and a user in another conversation. To test the ability of our model to generate facial expressions of multiple identities, on this dataset we utilize the 'users' as agents and 'operators' as human partners.

We used OpenFace [1] to process sampled image frames of humans and agents. Three features are extracted: a 17-d vector d representing likelihoods of 17 facial action units such as inner brow raiser and lid tightener; a 40-d shape descriptor $p = [p_r, p_n]$; a 136-d vector s of 68 landmarks. p can be obtained by fitting s to a Point Distribution Model (PDM) provided by OpenFace, in which p_r (p_n) represents the rigid (non-rigid) shape transformation. With a chosen p_r and the PDM, a p_n can produce a set of face landmarks which forms an affective face sketch (we link these landmarks by piece-wise linear lines of one pixel width). To train and test GAN models, We create training/testing sets on above two datasets respectively by uniformly sampling data pairs of human video sequences (as input to our system) and corresponding agent images (as ground truth of output).

3 Approach Details

3.1 Agent Face Sketch Generation

Shape GAN generates contextually valid face sketches of agents conditioned on interacting human's behavior. Figure 2 (Left) summarizes the architecture of Shape GAN. For time t, a 10-d random noise z (sampled from the uniform prior $\mathcal{U}(-1,1)$) , an 17-d feature c_t and a

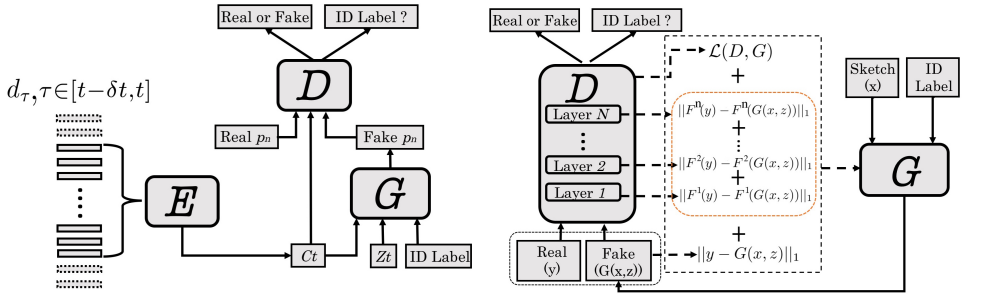

Figure 2: Left: the architecture of Shape GAN, illustrating the relationship between the discriminator (D), the generator (G) and the encoder (E), which is utilized to extract temporal context of a human partner's facial expression. Right: the architecture of Face GAN and how our layer feature loss works.

one-hot vector l representing identity label are used to form input of Generator (G). A one-layer Leaky Rectified Linear (LReLU) encoder is employed to compute each element of c_t from a sequence of temporal AU descriptors d_τ ($\tau \in [t-\delta t, t]$) of interacting humans:

$$c_t^i = LReLU\Big(\sum_{\tau \in [t-\delta t, t]} w_\tau^i d_\tau^i + b^i\Big), \tag{1}$$

where $LReLU(.)$ is a Leaky Rectified Linear function. In the training phase, d_t, corresponding real shape descriptors p_n and agent id labels l are sampled to form the training set. c_t is also used as conditional features in the Discriminator (D). G/D consists of five/three fully connected layers, each followed by batch normalization and rectified linear processing; D outputs a real/fake image classification and an id classification (to constain the training of G). To guarantee that G produces valid shapes, a L_1 loss $\mathcal{L}_{L1}(G)$ is combined with $\mathcal{L}(D,G)$, the original loss of Conditional GAN as follows:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y\sim p_{data}(x,y),z\sim p_z(z)}[||y - G(x,z)||_1], \tag{2}$$

$$G^* = \arg\min_G \max_D \mathcal{L}(D,G) + \lambda \mathcal{L}_{L1}(G), \tag{3}$$

where x here denotes an input facial expression feature c_t, y represents a corresponding real shape descriptor p_n^r and $G(x,z)$ is a fake shape descriptor generated. In our experiments we fixed $\lambda = 20$ and found it worked well. With a sampled p_r and the Point Distribution Model, p_n can produce face landmarks which form a sketch.

3.2 Photorealistic Agent Face Generation

Face GAN fleshes out agent affective face sketches into high-quality photorealistic images. We developed a framework similar to Isola et al. [14] with following two novel techniques.

Layer feature loss. In the original image-to-image translation method [14], the GAN objective is combined with a L_1 loss (in the form of Equation 3) to enhance image quality of outputs. In this context, x in Equation 3 denotes an input sketch image, y represents a corresponding real video frame and $G(x,z)$ is a fake face image generated from Generator (G).

This L_1 loss only computes the difference between real and generated images on pixel level and does not model the abstract content discrepancy. By using the loss function in Equation 3, we found the quality of generated images is not satisfactory in practice. Consider that the discriminator (D) is a convolutional neural network, in which lower level layers extract local features such as edges or corners of an input image, while higher-level layers extract more abstract information such as parts and contours. Inspired by this fact, we proposed a L_1^{layer} loss on the 'layer features' of D between real and generated images:

$$\mathcal{L}_{L1}^{layer}(G) = \mathbb{E}_{x,y\sim p_{data}(x,y),z\sim p_z(z)} \sum_l ||F^l(y) - F^l(G(x,z))||_1, \tag{4}$$

$$\mathcal{L}^* = \mathcal{L}(G) + \lambda(\mathcal{L}_{L1}(G) + \mathcal{L}_{L1}^{layer}(G)), \tag{5}$$

where F^l denotes flattened feature representation in layer l; $\mathcal{L}(G)$ is the original generator loss of conditional GAN; $\mathcal{L}^*$ is the final objective that combines the pixel level loss $\mathcal{L}_{L1}(G)$ and the layer feature loss $\mathcal{L}_{L1}^{content}(G)$). Figure 2 (Right) illustrates how this layer feature loss works. In this way, both the precise appearance and real content difference is constrained in our model to force the transferred images matching the original ones. We also fixed λ in Equation 5 to be 20 and found it worked well.

With this approach, D's task remains unchanged, i.e., distinguish real facial expression images from generated ones, but G's job is to not only fail D, but also produce images matching the content of real samples y (the input to D) in an L_1 sense. The noise signal of z is not explicitly fed into this stage; instead randomness is only provided in the form of dropout, applied on first 3 layers in the encoding network of the generator at both training and inference time. As shown in the experiment, this content loss $\mathcal{L}_{L1}^{content}$ produces images with much higher quality, compared with the pixel level loss L_{L1}.

Identity constraint with two-pass optimization. On SEMAINE dataset we model the face expression generation of multiple agents. Initially we adopted an 'Info-GAN' like model [6] to provide identity constraint, as shown in Figure 2 (Right). D not only distinguishes real images from generated ones, but also outputs an identity classification result which we expect to match the input label. We use its loss to constain the training of G; however, even with a large weight on this loss, G can only generate images with serious defects and blurs, as shown in Figure 3. Each of those images seems to contain facial features from multiple identities. To overcome this problem, we further impose a two-pass optimization on each mini batch. In the first pass, a sketch (from the ID i) and the label i are used as input and the loss of G is computed according to Equation 5. To compute the layer loss and the pixel-level loss, in this pass the generated image will be compared with the real image

Figure 3: Defective SEMAINE agent Images produced by Face GAN with 'Info-GAN'-like identity constraint. Each image was generated from a different ID and the same face sketch, but contains blurry facial features which are suspected to come from multiple identities.

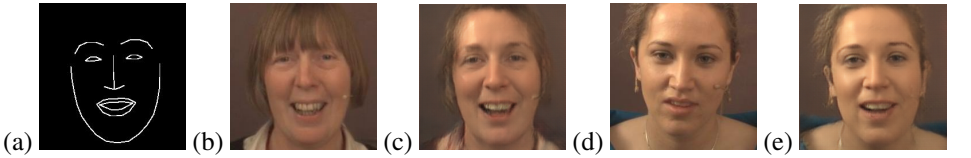

Figure 4: a) an input sketch of identity i; b) the real image of the sketch; c) the generated image of the sketch; d) a real image of identity j; e) the generated image of j. Given the identity label j and the sketch of i, our system is able to generate an image e) of j which contains very similar facial expression to that of i's images, b) and c).

Figure 5: First row: face images generated from different video interview session labels which illustrate different outfits, hair styles, lighting conditions or backgrounds. Second row: images generated from a fixed session label. Notice in each column two images are generated from the same sketch, therefore have the exactly same facial expressions.

corresponding to the input sketch. In the second pass, the same sketch from ID i and a different label $j, j \neq i$ are used as input; the generated image will be compared with a randomly sampled real image from ID j (we hope a sketch from ID i can still generate an image with a similar appearance to ID j given that the input ID label is j and $j \neq i$)). The loss of G in the second pass is computed as $\mathcal{L}^{**} = \mathcal{L}(G) + \lambda \mathcal{L}_{L1}^{layer}(G)$, that is, we only consider the layer feature loss without imposing a 'pixel-level' matching between the generated image and the randomly sampled image from ID j. As Figure 4 shows, with this two-pass optimization, even a sketch of label i can generate an image of j given the label j as input.

Although there is only one agent in the interview dataset, the lighting conditions, backgrounds, outfits or hair styles may vary from one interview session to another. On this dataset, we use video session IDs as 'identity' constraint and enforce above two-pass optimization. We are able to generate various facial expressions with similar outfits, hair styles, lighting conditions and background, as illustrated in Figure 5.

4 Experiments

4.1 Results on the Interview Dataset

In Figure 6 we show image pairs of human partners and the generated agent. The top row in each pair shows the last frame of a human partner's video clip which is used to generate the

Figure 6: The figure shows image pairs of human partners and the generated agent on the interview dataset. Each image in the top row shows the last frame of a human partner's video clip that is used to generate the virtual agent response below it. Notice how the generated agent frames embody an appropriate emotional response to the human partner.

virtual agent response in the row below it. Notice how the generated agent images embody an appropriate emotional response to the human partner. While evaluating generative models that can sample but not estimate likelihood directly is a challenging problem[10, 18], we designed two innovate methods to quantitatively test both the validity of generated agent's emotional response in the context of interacting human partner's behavior and the quality of photorealistic agent images.

Experimental Setting. From the interview dataset we randomly sampled 70,000 video clips, 100 frames in length each from the human partner videos and extracted their action unit vectors. For each human partner video clip a single frame from the associated agent video (last frame of clip) is also sampled for training. All face images are aligned and processed by OpenFace [1] to generate ground truth face shape descriptors and sketches. For testing, we randomly sampled 7000 human partner video clips V^h (100 frames each and no overlap with the training set) and used these as input to generate 7000 facial expression images of the virtual agent I^g. Separately, we also constructed a set of 7000 true agent images I^r sampled directly from the interviewer videos, with each image corresponding exactly to the last frame of associated human partner's clip in V^h.

Testing validity of generated agent behavior. We used an Euclidean distance between Action Unit vectors of two corresponding images to compute the facial expression difference. Our hypothesis is that on average, the paired expression distance between a real interviewer image $i^r_j \in I^r$ and corresponding generated virtual agent clip $i^g_j \in I^g$ should be significantly smaller than the distance with a control group $i^r_k, (k \neq j)$ randomly selected from I^r. We formed two distance groups Dis^{paired} and Dis^{rand}, where Dis^{paired} contains 7000 paired distance values $Dis(i^r_j, i^g_j)$ and Dis^{rand} contains 7000 random distance values $Dis(i^r_j, i^g_k), k \neq j$ (for each j we randomly sampled k for 100 times to compute an average distance). Our hypothesis could be tested by a lower-tailed, two sample t-test in which the null/alternative hypotheses is defined as H_0: $\mu_1 = \mu_2$ and H_α: $\mu_1 < \mu_2$ respectively, in which μ_1 (μ_2) represents the mean of Dis^{paired} (Dis^{rand}) (as shown in Table 1). We adopted Matlab function *ttest2* to conduct this test in which Satterthwaite's approximation [28] was used for the case that equal variances of two distributions are not assumed. At a significance level of 0.05, the computed p-value is 6.4×10^{-7}. The alternative hypothesis H_α is accepted at the 0.05 significance level, which concludes a statistically significant reacting effect of our model.

Measuring image quality. To measure the quality of images generated by Face GAN and compare our layer feature loss with the L_1 loss in [14], we extracted face sketches from each sample in I^r (defined above) and input them to two pre-trained Face GAN models,

	Dis^{paired}	Dis^{rand}
Sample Mean($\pm$ STD)	0.157 ± 0.231	0.248 ± 0.216

Table 1: Sample means and standard deviations of two distance groups.

Table 2: The confusion matrix described in Section 4.2. The horizontal dimension indicates predicted classes of generated agents and the vertical dimension denotes the actual classes of their real counterparts.

%	Poppy	Spike	Prudence	Obadiah
Poppy	**71.4**	4.6	19.7	4.3
Spike	2.1	**63.9**	12.3	21.7
Prudence	15.6	9.7	**60.3**	14.4
Obadiah	3.4	16.6	6.8	**73.2**

where the first one was trained with our layer feature loss while the second one with a pixel-level L_1 loss [14]. In this way we obtained two generated image sets I_1^g and I_2^g. In this experiment, we masked all generated images by an oval shape to compare the appearance difference in the face area only. To compute the similarity of an image $I_1^g(k)$ (or $I_2^g(k)$) to its true counterpart in $I^r(k)$, we used the mean squared error (MSE) to measure the difference between 7000 pairs of $i_1^g(k)$ (or $i_2^g(k)$) and $i^r(k)$. The average MSE ($\pm STD$) of two sets on 7000 pairs are $\mathbf{47.03 \pm 53.71}$ and $\mathbf{87.88 \pm 94.86}$ respectively. Our method outperforms the method of [14] by cutting down the mean squared error (MSE) almost by half (from 87.88 to 47.03). These results quantitatively verified that our model with layer feature loss produces better images. Some samples in Figure 7 illustrate the image quality comparison between two sets of generated images.

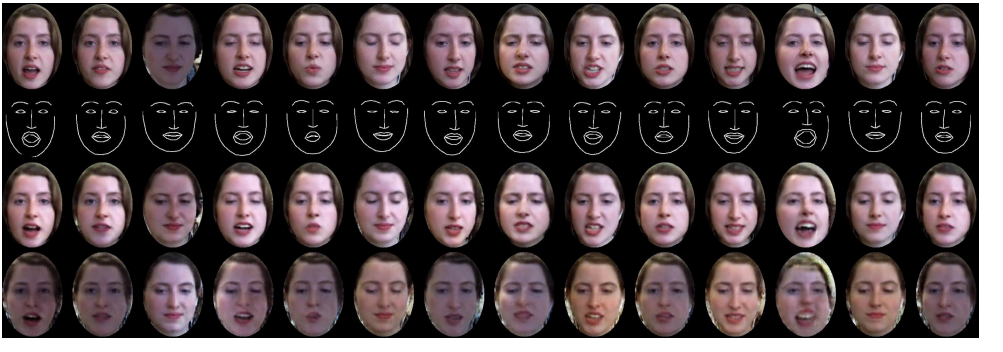

Figure 7: Image quality comparison between real images (I^r, Row 1), corresponding images generated with layer feature loss (I_1^g, Row 3) and images generated with a pixel-level L_1 loss only (I_2^g, Row 4). Row 3 and Row 4 are generated from the same sketches in Row 2.

4.2 Results on SEMAINE Dataset

On SEMAINE Dataset, our model employs a training (testing) set containing 10,000 (6000) sampled sequence-image data pairs. As introduced in Section 2, one individual can play a human partner in one session and an agent in another conversation. Images of agents can be categorized to 4 classes based on their human partner's personality. However, a large portion of them only express neutral emotions no matter what personality their human partners

Figure 8: The first column shows the last frame of 4 input sequences of 4 human partners. From top to bottom, each person plays one of 4 personalities respectively: Poppy, Spike, Prudence and Obadiah. On the right, each row illustartes the generated facial expression images of 8 agents corresponding to the input on the left.

played. To analyze the most representative cases, Euclidean distances are computed between AU vector of each real image of agents and all AU vectors of agent images from a different personality class. We summed up these distances (for each image) and selected top-ranked 4000 (2400) agent images and corresponding sequences of human partners from training

Figure 9: Smaple agent images generated from different typies of human partners: (a) Poppy, (b) Spike, (c) Prudence and (d) Obadiah.

(testing) set. These two sets of agent images are the most distinctive samples of 4 classes and contains much less neutral expressions. By using the selected 2400 image sequences of humans from the testing set, we generate 2400 fake agent face images (some generated samples are illustrated in Figure 8 and Figure 9). With the 4000 real agent images from the training set, we trained a convolutional network to perform a 4-way classification according to their human partner's personalities. We utilize this network on 2400 generated images and the confusion matrix is shown in Table 2. Results in Figure 8, 9 and Table 2 demonstrate that our model is effective at generating visually compelling facial expressions of agents that reflect valid emotional reactions to behavior of partners: the predicted categories of generated agents generally align well with the true classes of their real counterparts.

5 Discussion

A key feature of our approach is that the generative model for agent behavior is learned not as a function of predefined rules or the generated identity's own/self attributes but rather the behavior of their interaction partner. Consistent with much of the literature in behavioral neuroscience on the significance of non-verbal modes of communication through facial expressions [9, 20], our approach demonstrates that even when limiting observations to facial expressions of the interacting human partner we can generate agent behavior depicting valid emotional responses. Our model can be enhanced by fusing a multitude of modalities to more comprehensively model behavior of the human partner, including speech, paralinguistics, interaction context and past behavior of the virtual agent itself.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[2] S. Barsade. The ripple effect: Emotional contagion and its influence on group behavior, 2002.

[3] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual detection of behavioural mimicry. In *IEEE Humaine Association Conference on Affective Computing and Intelligent Interaction*, Chicago, 2013.

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. SIGGRAPH '99, 1999.

[5] The Duy Bui, Dirk Heylen, Mannes Poel, and Anton Nijholt. Generation of facial expressions from emotion using a fuzzy rule based system. In Markus Stumptner, Dan Corbett, and Mike Brooks, editors, *AI 2001: Advances in Artificial Intelligence*, pages 369–391. Springer, Berlin, 2001. URL http://doc.utwente.nl/60383/.

[6] Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances In Neural Information Processing Systems 29*. 2016.

[7] D. Devault, A. Rizzo, and L.P. Morency. Simsensei: A virtual human interviewer for healthcare decision support. In *Thirteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2014.

[8] P. Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press*, 1978.

[9] Chris Frith. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535), 3453-3458*, 2009.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*. 2014.

[11] M.E. Hoque, M. Courgeon, M. Mutlu, J.C. Martin, and R.W. Picard. Mach: My automated conversation coach. In *Ubicomp*, 2013.

[12] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *arXiv preprint arXiv:1612.04357*, 2016.

[13] Yuchi Huang and Saad Khan. Dyadgan: Generating facial expressions in dyadic interactions. In *CVPR Workshops*, July 2017.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv e-prints*, November 2016.

[15] C. Li and M. Wand. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. *ECCV*, April 2016.

[16] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *ArXiv e-prints*, November 2015.

[17] Gary McKeown et al. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.*, 2012.

[18] M. Mirza and S. Osindero. Conditional generative adversarial nets. *ArXiv e-prints*, November 2014.

[19] J. S. Morris, A. Ohman, and R. J. Dolan. A subcortical pathway to the right amygdala mediating âĂŸunseenâĂŹ fear. In *National Academy of Science*, 1996.

[20] Alex Pentland and Tracy Heibeck. Honest signals: how they shape our world. In *MIT press*. 2010.

[21] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. *ArXiv e-prints*, May 2016.

[22] A. Ohman & J. J. Soares. Emotional conditioning to masked stimuli: expectancies for aversive outcomes following nonrecognized fear-relevant stimuli. *Experimental Psychology*, 1998.

[23] J. M. Susskind, G. E. Hinton, Movellan J. R., and A. K. Anderson. Generating facial expressions with deep belief nets. In *In Affective Computing, Emotion Modelling, Synthesis and Recognition*. I-Tech Education and Publishing, 2008.

[24] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4), July 2017.

[25] A. A. Tawfik, L. Sanchez, and D. Saparova. The effects of case libraries in supporting collaborative problem-solving in an online learning environment. *Technology, Knowledge and Learning*, 19(3):337–358, 2014.

[26] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. NieÃ§ner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[27] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NIPS 2016, December 5-10, 2016, Barcelona, Spain*, pages 613–621, 2016.

[28] B. L. Welch. The generalization of student's problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.

[29] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, pages 82–90, 2016.

[30] Matthew D. Zeiler, Graham W. Taylor, Leonid Sigal, Iain A. Matthews, and Rob Fergus. Facial expression transfer with input-output temporal restricted boltzmann machines. In *NIPS*, 2011.

[31] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *arXiv preprint arXiv:1612.03242*, 2016.

[32] Y. Zhou and T. L. Berg. Learning Temporal Transformations From Time-Lapse Videos. *ECCV*, August 2016.

[33] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.