2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Dataset Infant Anonymization with Pose and Emotion Retention

Mason Lary[1], Matthew Klawonn[2], Daniel Messinger[3], and Ifeoma Nwogu[1]
[1] Dept. of Computer Science and Engineering, University at Buffalo, NY
[2] Information Directorate, U.S. Air Force Research Lab
[3] Dept. of Psychology, University of Miami, Coral Gables, FL, USA

2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3503-9494-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/FG59268.2024.10581938

Abstract— **We demonstrate a procedure for the anonymization of infant subjects in videos such that salient behavioral information is retained. This method also creates a new identity that is consistent temporally across video frames. We present an overview of this anonymization process, which involves moving through the latent space of a generative model with an infant specific latent space traversal technique. We apply the technique on videos of infants, a historically difficult source of data, and make comparisons to other state-of-the-art anonymization systems. Metrics demonstrate an improved ability to retain emotional content of videos during the anonymization process, even during extreme emotions or poses, while maintaining a consistent identity throughout.**

I. INTRODUCTION

In the world of infant behavioral research, data transparency is sorely lacking. Due to privacy concerns, researchers may have to forego the sharing of videos and images of infants they have collected. This lack of fluidity means other teams must resort to collecting their own data for each experiment, a laborious, time-consuming, and potentially expensive task. To ease this burden, researchers may utilize tools specialized for anonymizing subjects in media, such as [10], [16], [19], and [22]. However, applying these existing methods to anonymize subjects in images can have severe side effects when applied to videos. Our goal is the anonymization of images and videos, while retaining the quality and salient aspects of the original media to be used in downstream tasks.

Fig. 1 demonstrates examples of side effects that existing anonymization methods can produce. Most notably, the emotional content of the anonymized video is significantly affected, rendering the generated data useless for tasks involving behavioral analysis. A common preprocessing task in behavioral studies of infants is the detection of facial action units (FACS) [7], [8], [23], a coding system for the muscles activated in the face for a given emotion. The loss of emotion in an anonymized image means the coding process produces significantly different results between the original and anonymized dataset, which is disastrous for methods that rely on FACS, such as in automated FACS recognition [8] or still-face experiments [28], [30]. In addition, Fig. 1 displays the lack of temporal consistency and inconsistencies with original and generated ages that plague other anonymization methods. As a result, the facial and skull morphology of the anonymized frames are often inconsistent with the original video and with one another. This makes it difficult to identify facial landmarks, a common preprocessing task in behavioral analysis [8], [26], [29].

To rectify these issues, we present a StyleGAN based method for anonymizing videos of infants without removing emotional content or pose information from the resulting video. In other words, the identity of an infant in a video anonymized using our method differs from the original by a significant amount, but the pose and emotions shown in the output video are close to those of the original. This method also boasts a higher degree of temporal consistency between anonymized frames than other methods. We accomplish this through a technique that traverses the latent space of the GAN in a disentangled way to find new latent vectors for generation that result in a new identity without losing the information necessary for behavioral analysis.

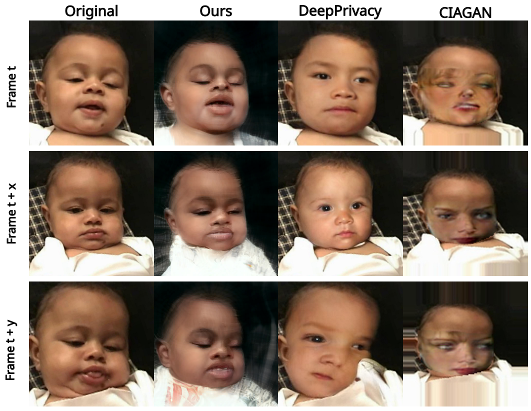

Fig. 1: A comparison of current GAN-based anonymization tools applied to a single infant video. These images demonstrate the loss in salient content from other anonymization methods. These images were taken from the SIBSMILE [21] dataset.

II. Related Works

A. Generative Adversarial Networks

With the proliferation of generative technologies, GANs represent a staple of privatization methods [10], [11], [15], [20]. We base our method upon StyleGAN [14], a generative adversarial network well known for its ability to generate human faces, as well as its easily explored and well studied latent space for controlling generation. We specifically choose StyleGAN3 [13] to build our method, as it offers better temporal consistency when applied to videos than its predecessors.

Mason Lary is sponsored by a DoD SMART scholarship.

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

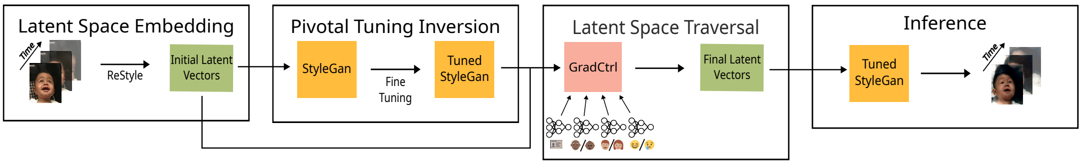

Fig. 2: An overview of our method. We embed video frames into the latent space of the StyleGAN. These latent vectors are used to fine-tune the StyleGAN model to produce results that closely match the dataset. The initial latent vectors and the fine-tuned StyleGAN model are then used to find the edited latent vectors that produce anonymized images. The fine-tuned StyleGAN maps these edited latent vectors to the final anonymized images.

Other GAN-based anonymization methods such as CIAGAN [20] and DeepPrivacy [11] typically train custom models with an objective function aimed at producing anonymized images. This results in networks that are able to successfully anonymize images at the cost of destroying any behavioral information from the original frames. By using StyleGAN as a starting point, we make use of several well-studied techniques, such as GAN inversion, pivotal tuning inversion, and an editable latent space, that aid in the anonymization process while allowing for the retention of behavioral attributes.

B. 3D Morphable Models

3D morphable models (3DMM) represent parameterized models of the human face, examples of which can be found in [6], [9], and [31]. Through parameters such as identity, pose, and expression, these methods are able to generate a wide variety of human faces as 3D models. To create an anonymous video using 3DMM techniques, one could vary pose and expression parameters across frames while keeping the identity parameter consistent.

However, infants are often out of the distribution of training data used to create these 3D models, leading to a morphology in the output video that would be found in a much older child or even an adult. Furthermore, retraining of these models requires data that comes from expensive 4D scanners, limiting the work that can be done to extend the data distribution. It is therefore imperative to have an anonymization method that can be trained using more abundant images and videos of infants.

III. Methodology

This section presents an overview of our method divided into three areas representing the three fundamental steps in obtaining our anonymized videos. The first step involves embedding video frames into the latent space of the GAN through an inversion process. Unfortunately, doing so naïvely yields latent vectors that generate subpar images lacking temporal consistency and similarity to the original frames. Therefore, in the second step, we leverage a process called Pivotal Tuning Inversion to fine-tune the StyleGAN backbone. This step allows us to better embed video frames into the latent space. The resulting latent vectors generate frames that are temporally consistent, and retain salient aspects of the original frames from which they come. In our third step, we add differentiable components to the StyleGAN backbone that allow us to edit the latent vectors found in step 2 according to our desired criteria.

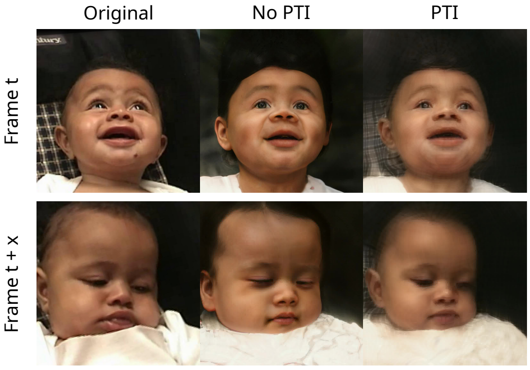

Fig. 3: A comparison between video frame inversion results for a StyleGAN model with and without PTI finetuning. The latent vector for both is the inverted image from the original column. It can be seen that the similarity in generated images to the original frame is higher after the PTI process.

A. Latent Space Embedding

StyleGAN, like many GANs, can be understood as a black box model g that maps a unique vector v to a unique image $g(v)$. Given a dataset of images $\{i_1, i_2, \ldots, i_N\}$ that come from video frames, an identity similarity function id, and an emotional similarity function e, where lower scores indicate lower similarity, our method attempts to find latent vectors which yield generated images that have dissimilar identities and similar emotion as compared to the original images. In other words, our objective function is:

$$\min_{v_1 \ldots v_N} \frac{1}{N} \sum_j id(i_j, g(v_j)) - e(i_j, g(v_j)) \tag{1}$$

The latent space of StyleGAN also has the property that directional paths in the latent space correspond to semantic changes in output images [4], [14], [25]. For instance, if $g(v)$ produces an image of a person with blonde hair, Δx could represent a direction in the latent space such that $g(v + \Delta x)$ is an image that has darker hair. We exploit this property of the latent space to design a traversal process, which we describe in more detail later in the paper. For now, let us focus on describing how to find v_j for a given image i_j. We require a way to invert an image i_j, such that the similarity

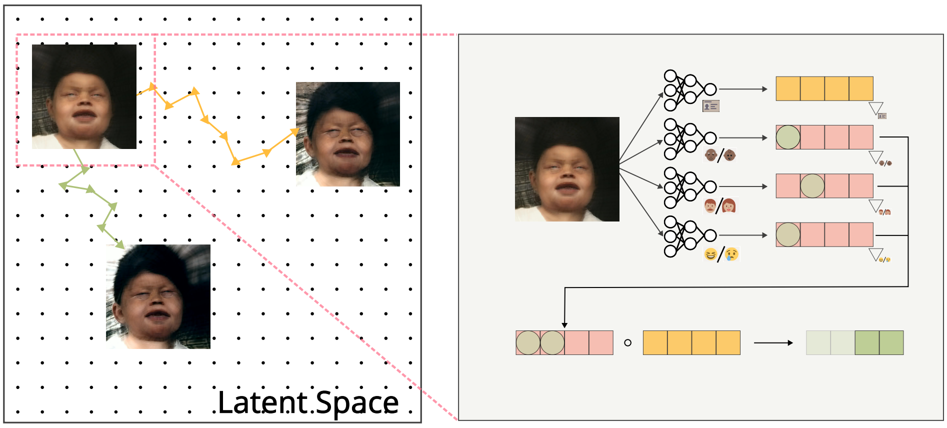

Fig. 4: An overview of the traversal method for the latent space. The top arrow is a path followed without disentangling, resulting in an image that looks much older than the original. The arrow at the bottom is a path that takes disentanglement into account. The disentanglement process is described on the right side of the diagram. Circles represent the top channels of a particular gradient, and the transparent areas of the final gradient represent channels that have been masked out.

between $g(inv(i_j))$ and i_j is maximized, where inv is the embedding process.

We turn to a technique called ReStyle, which can map an image into the StyleGAN latent space using an autoencoder [1], [2]. The process encodes translation and rotation information of the video subject into the latent vectors, meaning if video frames are close together in time, the location of their inverted vectors are also close together.

B. Pivotal Inversion Tuning

Unfortunately, using ReStyle alone produces images $\{g(v_1), g(v_2), \ldots, g(v_n)\}$ that lack similarity with the original images, as seen in the middle column of Fig. 3. We would like to match the generated images as closely as possible to the original in order to retain as much salient behavioral information as possible. To rectify this, we apply a technique known as Pivotal Tuning Inversion [24]. The basic idea of PTI is to fine-tune the StyleGAN3 model g' on only the latent vectors $\{v_1, v_2, \ldots, v_n\}$, such that the similarities between $\{i_1, i_2, \ldots, i_n\}$ and $\{g'(v_1), g'(v_2), \ldots, g'(v_n)\}$ are maximized.

This technique has an added benefit related to the editability of the latent vectors. Within the latent space of the GAN are certain regions where editing can be performed in a more disentangled way [27]. This means changing a latent vector in a semantic direction, such as to obtain darker hair color, does not change other semantics, such as whether the generated face wears glasses or not. By applying PTI to fine-tune StyleGAN3, the latent vectors from embedding the video end up in these more editable regions, benefiting the disentanglement of the anonymization method. The results of generation after this step are given in Fig. 3. It should be noted that this unfortunately causes a loss in generated image quality. However, the benefits of performing PTI outweigh the downsides when looking to retain as much of the information of the original videos as possible.

C. Latent Space Traversal

To control generation, we provide to the model an attribute a to edit, as well as a set of attributes $\{b, c, \ldots, z\}$ that should be retained in the output. Given these attributes, we require differentiable functions $\{f_a, f_b, \ldots, f_z\}$ that can score a latent vector v_j, with a high score indicating that $g'(v_j)$ is likely to have the given attribute, as described in [3]. As an example, $f_{female}(v_j)$ for a latent vector v_j could provide a probability that $g'(v_j)$ will be classified as female. We can calculate $\{\nabla_{v_j} f_a(v_j), \nabla_{v_j} f_b(v_j), \ldots, \nabla_{v_j} f_z(v_j)\}$ with respect to v_j. Changing the latent vector in the direction of one of these gradients should correspond to a greater expression of the underlying attribute. For instance, $g'(v_j + \alpha * \nabla_{v_j} f_{female}(v_j))$ and $g'(v_j - \alpha * \nabla_{v_j} f_{female}(v_j))$ should generate images that have a higher chance of being classified as female and male, respectively. α serves as the step size.

Attributes can be entangled with each other in the latent space, such that moving in one semantic direction causes changes in another. This is demonstrated in Fig. 4, where the age and identity of the generated image are shown to be entangled. The entanglement arises from high magnitude components in $\{\nabla_{v_j} f_b(v_j), \ldots, \nabla_{v_j} f_z(v_j)\}$ coinciding with components of $\nabla_{v_j} f_a$. To combat this, we calculate the positions of the top-K magnitudes in each gradient $\{\nabla_{v_j} f_b(v_j), \ldots, \nabla_{v_j} f_z(v_j)\}$ and set the same positions in $\nabla_{v_j} f_a$ to 0, giving $(\nabla_{v_j} f_a)'$. Changing a latent vector in the direction of $(\nabla_{v_j} f_a)'$ should therefore cause a smaller change in the expression of the attributes that should stay the same, allowing for disentangled editing.

For anonymizing infant behavioral research videos, we set the attribute to change to be the identity of the generated image, while the ones to retain are emotion, age, and gender. We chose to add scorers for age and gender after early experimental results showed that changing identity often changed these attributes as well, and in so doing negatively affected the facial and skull morphology of the images.

For our scoring functions, we train single layer neural

TABLE I: A comparison of methods on the SIBSMILE dataset.

	Identity		Arousal	Valence	Pose
	Cos ↓	TC ↑	RMSE ↓	RMSE ↓	RMat ↓
PTI + Traversal	.55	**.79**	.29	.19	25.79
PTI	.69	**.79**	**.22**	**.15**	**20.59**
Traversal	.44	.76	.37	.22	22.14
DeepPrivacy	.28	.41	.48	.33	80.33
CIAGAN	**.21**	.64	.58	.38	57.14

networks $\{f_{id}, f_{emo}, f_{gen}, f_{age}\}$ on random latent vectors v_j, generating labels for each v_j by applying pretrained classifiers from [18] to $g'(v_j)$. The identity attribute is an exception to this rule; we generate latent vectors v_j for video frames from the embedding process previously mentioned, as well as random latent vectors r_j. ArcFace [5] is used on generated images $\{g'(v_j); g'(r_j)\}$ to create a unit vector i representing the identity of a face. We choose a label for each vector based on a threshold for the cosine similarity between i and the average identity vector from all frames from the current video.

IV. Experiments

To test our method, we utilize the SIBSMILE dataset [21], which includes videos of infants undergoing a still face experiment [28]. Throughout this experiment, infants display a wide range of pose and emotion, presenting a challenging task for anonymization methods. There are 174 videos from 58 infants in 3 different phases of the still face experiment. For traversing the latent space, we use 300 steps per image and an α of -75. Our method is compared against the DeepPrivacy and CIAGAN models, two state-of-the-art anonymization methods [11], [20]. We also perform an ablation study by separating the two components of our method, only applying PTI or latent space traversal to latent space images.

Our goal is to analyze quantitatively, for each anonymization method, the degree to which edited video frames are anonymized, retain the same arousal, valence, and pose of the original frames, and demonstrate temporal consistency. We compare with state-of-the-art and with ablations in Table I. We use valence and arousal measures from [26] to assess each approach's ability to retain emotional features between corresponding original and anonymized frames. Valence measures how positive or negative and emotion is while arousal represents its intensity, with both being standard features for emotional analysis. RMSE is taken as a metric of similarity between results. We compare the identity of edited and original frames by feeding both through the identification network from [5] and taking a cosine similarity (Cos). Identities under the ArcFace model are generally considered the same if the angle between identity vectors are less than 45°, or a cosine similarity above .71 [5]. To demonstrate temporal consistency (TC), we compute the weighted cosine similarity between identity vectors for generated frames spaced 3 seconds apart. Finally, we measure pose of the original and anonymized image using FLAME [17]. We compute a metric (RMat) between the rotation matrices of the two poses, as given in [12].

Compared to the state of the art, our method performs better on every metric except the anonymization metric. However, based on the Arcface threshold mentioned previously, this level of anonymization is acceptable. This demonstrates the ability of our process to anonymize a subject while retaining behavioral attributes. Comparing to the ablation results, applying only PTI results in marginally better behavioral attribute retention while essentially failing to anonymize the subject. Applying only the traversal process demonstrates a better anonymization process than our method at the cost of losing emotional retention. Our method combines the behavioral retention of the PTI process with the anonymization of the traversal process.

V. Future Work

Our results demonstrate promise in the realm of utility preserving anonymization methods. However, there is still room for improvement, especially in the degree of anonymization offered. While it was not present in this work, future contributions should involve differential privacy in order to prove that identities are sufficiently hidden. Differential privacy guarantees would increase the likelihood of this work being practically useful to behavioral researchers. As it is, the results of this paper show an avenue towards anonymizing sensitive identity information while retaining salient features for downstream tasks.

Another avenue of future work involves the quality of output images, especially by replacing only the face, rather than the whole frame. This would more closely mimic CIAGAN, but would provide overall better image quality. Diffusion models might provide a path towards this goal due to their ability to produce high fidelity output videos and their ability to perform image infilling. It is also possible that diffusion models contain latent spaces with a lower degree of entanglement, allowing for an easier editing experience during anonymization.

VI. Conclusion

Infants are underrepresented in computer vision data sets due to tedious data collection and hurdles preventing data sharing. Models trained on images of humans have a tendency to break down on the out-of-distribution images, which we demonstrated with videos of infants. In spite of these limitations, we have developed a technique for anonymizing videos of infants such that emotional and pose content is preserved during the process. We have displayed the efficacy of this technique on a difficult dataset containing videos of infants undergoing a behavioral study and showing extreme emotions, occlusions, and shifts in pose. For researchers in human behavioral research, this process is a step forward towards the goal of sharing sensitive datasets that could aid the research community as a whole.

References

[1] Y. Alaluf, O. Patashnik, and D. Cohen-Or. ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6691–6700, Montreal, QC, Canada, Oct. 2021. IEEE.

[2] Y. Alaluf, O. Patashnik, Z. Wu, A. Zamir, E. Shechtman, D. Lischinski, and D. Cohen-Or. Third Time's the Charm? Image and Video Editing with StyleGAN3, Jan. 2022. arXiv:2201.13433.

[3] Z. Chen, R. Jiang, B. Duke, H. Zhao, and P. Aarabi. Exploring Gradient-Based Multi-directional Controls in GANs. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, volume 13683, pages 104–119. Springer Nature Switzerland, Cham, 2022. Series Title: Lecture Notes in Computer Science.

[4] M. J. Chong, H.-Y. Lee, and D. Forsyth. StyleGAN of All Trades: Image Manipulation with Only Pretrained StyleGAN, Nov. 2021. arXiv:2111.01619.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, June 2019. ISSN: 2575-7075.

[6] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. pages 285–295. IEEE Computer Society, June 2019.

[7] P. Ekman and W. V. Friesen. Facial Action Coding System, Jan. 2019. Institution: American Psychological Association.

[8] I. O. Ertugrul, L. A. Jeni, W. Ding, and J. F. Cohn. AFAR: A Deep Learning Based Tool for Automated Facial Affect Recognition. *IEEE International Conference on Automatic Face & Gesture Recognition*, 2019, May 2019.

[9] Y. Feng. DECA: Detailed Expression Capture and Animation (SIGGRAPH2021), Dec. 2022. original-date: 2020-04-08T00:51:49Z.

[10] H. Hukkelas and F. Lindseth. DeepPrivacy2: Towards Realistic Full-Body Anonymization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1329–1338, Waikoloa, HI, USA, Jan. 2023. IEEE.

[11] H. Hukkelås, R. Mester, and F. Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization, Sept. 2019. arXiv:1909.04538.

[12] D. Q. Huynh. Metrics for 3D Rotations: Comparison and Analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, Oct. 2009.

[13] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-Free Generative Adversarial Networks. In *35th Conference on Neural Information Processing Systems*, 2021.

[14] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4217–4228, Dec. 2021.

[15] S. M. S. M. Khorzooghi and S. Nilizadeh. StyleGAN as a Utility-Preserving Face De-identification Method, Dec. 2022. arXiv:2212.02611.

[16] M. Klemp, K. Rösch, R. Wagner, J. Quehl, and M. Lauer. LDFA: Latent Diffusion Face Anonymization for Self-driving Applications. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3199–3205, Vancouver, BC, Canada, June 2023. IEEE.

[17] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):1–17, Nov. 2017.

[18] J. Lin, R. Zhang, F. Ganz, S. Han, and J.-Y. Zhu. Anycost GANs for Interactive Image Synthesis and Editing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14981–14991, Nashville, TN, USA, June 2021. IEEE.

[19] T. Ma, D. Li, W. Wang, and J. Dong. CFA-Net: Controllable Face Anonymization Network with Identity Representation Manipulation, Oct. 2021. arXiv:2105.11137.

[20] M. Maximov, I. Elezi, and L. Leal-Taixé. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5446–5455, June 2020. arXiv:2005.09544.

[21] M. Ning, I. O. Ertugrul, D. Messinger, J. Cohn, and A. A. Salah. Automated Emotional Valence Estimation in Infants with Stochastic and Strided Temporal Sampling.

[22] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or. Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics*, 39(6):225:1–225:14, Nov. 2020.

[23] H. Oster and M. Dondi. Facial Action Coding System for Infants and Young Children (Baby FACS)_preconference Workshop (ICIS 2022). July 2022.

[24] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or. Pivotal Tuning for Latent-based Editing of Real Images. *ACM Transactions on Graphics*, 42(1):6:1–6:13, Aug. 2022.

[25] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the Latent Space of GANs for Semantic Face Editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9240–9249, Seattle, WA, USA, June 2020. IEEE.

[26] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, Jan. 2021. Number: 1 Publisher: Nature Publishing Group.

[27] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics*, 40(4):133:1–133:14, July 2021.

[28] E. Tronick, H. Als, L. Adamson, S. Wise, and T. B. Brazelton. The Infant's Response to Entrapment between Contradictory Messages in Face-to-Face Interaction. *Journal of the American Academy of Child Psychiatry*, 17(1):1–13, Dec. 1978.

[29] R. Wang, Y. J. Amy Ahn, D. Messinger, and I. Nwogu. Towards the Synthesis of Parent-Infant Facial Interactions. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, Dec. 2021.

[30] M. K. Weinberg, M. Beeghly, K. L. Olson, and E. Tronick. A Still-face Paradigm for Young Children: 2½ Year-olds' Reactions to Maternal Unavailability during the Still-face. *The journal of developmental processes*, 3(1):4–22, 2008.

[31] W. Zielonka, T. Bolkart, and J. Thies. Towards Metrical Reconstruction of Human Faces. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 250–269, Cham, 2022. Springer Nature Switzerland.