Continuous emotions: exploring label interpolation in conditional generative adversarial networks for face generation

Silvan Mertes, Florian Lingenfelser, Thomas Kiderle, Michael Dietz, Lama Diab, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Mertes, Silvan, Florian Lingenfelser, Thomas Kiderle, Michael Dietz, Lama Diab, and Elisabeth André. 2021. "Continuous emotions: exploring label interpolation in conditional generative adversarial networks for face generation." In *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications, July 7-9, 2021*, edited by Ana Fred, Carlo Sansone, and Kurosh Madani, 132–39. Setúbal: SciTePress. https://doi.org/10.5220/0010549401320139.

Continuous Emotions: Exploring Label Interpolation in Conditional Generative Adversarial Networks for Face Generation

Silvan Mertes, Florian Lingenfelser, Thomas Kiderle, Michael Dietz, Lama Diab and Elisabeth André
Lab for Human-Centered Artificial Intelligence, Augsburg University, Germany

Keywords: Generative Adversarial Networks, Face Generation, Conditional GAN, Emotion Generation, Label Interpolation.

Abstract: The ongoing rise of Generative Adversarial Networks is opening the possibility to create highly-realistic, natural looking images in various fields of application. One particular example is the generation of emotional human face images that can be applied to diverse use-cases such as automated avatar generation. However, most conditional approaches to create such emotional faces are addressing categorical emotional states, making smooth transitions between emotions difficult. In this work, we explore the possibilities of label interpolation in order to enhance a network that was trained on categorical emotions with the ability to generate face images that show emotions located in a continuous valence-arousal space.

1 INTRODUCTION

With the recent progress in the field of *Generative Adversarial Learning*, a broad variety of new algorithms have evolved that tackle the generation of artificial image data, leading to new possibilities for automated face generation and other generative tasks (Hong et al., 2019). State-of-the-art *Generative Adversarial Nets* (GANs) excel in terms of image quality of the generated outputs when compared to other generative model paradigms like *Variational Autoencoders*. However, common GAN architectures lack the capability to generate new data in a targeted way. Multiple modifications have been done to the original GAN framework to allow a controlled generation of new images. These modified architectures have proven their ability to create avatar images, i.e., images of human faces, that can be controlled regarding various human-interpretable features. In the context of emotional face generation, this allows to produce avatar images that are conditioned on a certain emotion.

The majority of datasets applicable to train those GANs are referring to a categorical emotion model, which means they are labeled on emotions like *Happy* or *Sad* in a discrete way. However, for many real-life use cases like emotional virtual agents, there is a demand for the generation of emotional faces in a more fine-grained way in order to strengthen the credibility and anthropomorphism of human-like avatars. Furthermore, avatars that are only capable of showing discrete emotions without gradation are impractical to use in scenarios where a smooth transition between different emotional states is required. Further use-cases include automated creation of textures for virtual crowd generation or data augmentation for emotion recognition tasks. Especially in the latter context, there is a huge demand for artificially created data, as continuous emotion recognition relies on non-categorical training data, whereas available datasets that are labeled with respect to dimensional features are rare. In all these scenarios, the use of a dimensional emotion model would be more sufficient to meet the posed requirements.

This work explores the possibility to train a *Conditional GAN* (cGAN) on a dataset of categorically labeled emotional faces and subsequently interpolate in the label space of that pretrained model in order to generate faces whose emotional expression can be controlled in a continuous, dimensional way. Thus, this paper aims to answer the question if label interpolation can be a tool for overcoming the disadvantage of categorical datasets for emotional face generation.

2 RELATED WORK

The possibility to create artificially generated images has experienced a great upswing with GANs, that firstly have been presented by (Goodfellow et al.,

Mertes, S., Lingenfelser, F., Kiderle, T., Dietz, M., Diab, L. and André, E.
Continuous Emotions: Exploring Label Interpolation in Conditional Generative Adversarial Networks for Face Generation.
DOI: 10.5220/0010549401320139
In *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications (DeLTA 2021)*, pages 132-139
ISBN: 978-989-758-526-5

2014). The basic idea of those GANs is, that two neural networks, the *Generator* and the *Discriminator* (or *Critic*), compete against each other in a minmax game. Therefore, the generator learns to create new samples that resemble a given training domain, whereas the discriminator learns to distinguish between real samples and fake samples generated by the generator. After training, the generator is able to transform random noise vectors into artificial image data.

In recent years, original GANs were extended in multiple ways. (Arjovsky and Bottou, 2017), (Gulrajani et al., 2017) and (Arjovsky et al., 2017) refined loss functions and training procedures in order to stabilize the training of GANs. (Radford et al., 2015) introduced *Deep Convolutional GAN* (DCGAN), that replaced the fully connected layers of both generator and discriminator networks with convolutional networks, resulting in the capability to create high quality image data. Further, their work presented the first attempts to generate highly realistic images of human faces. In their respective publication, it was also examined how the latent space of DCGAN implicitly models certain face features like the face pose. However, it can not be controlled which features are learnt by DCGAN as it is trained in an unsupervised setting. An even more sophisticated approach that was evaluated in the context of face-generation was presented by (Karras et al., 2017), who introduced their *Progressive Growing GANs* that are able to generate high-resolution face images.

A common problem with all those architectures is the fact, that the output results solely from random noise input, thus, it cannot be controlled in a human-interpretable way. Therefore, (Mirza and Osindero, 2014) introduced the concept of *Conditional GANs*. Here, the random noise vector is extended with additional label information. That information is used to condition the network to certain features during training. Therefore, a successfully trained cGAN can be used to generate new outputs in a targeted way. Multiple approaches have been published to use those cGANs to generate human faces that are conditioned on specific features. For example, (Wang et al., 2018) and (Gauthier, 2014) developed cGANs that are capable of generating images that can be controlled regarding different features like *glasses*, *gender*, *age*, *mouth openness*, among others. (Yi et al., 2018) used a cGAN to generate emotional face images to enhance datasets for emotion recognition systems.

However, these systems are either trained on categorical feature data to generate new face images conditioned on discrete classes, or they already use continuous label information during the training.

Another class of GANs focusses on style conversion problems. In the context of face generation, the tasks to be solved by those systems can be referred to as *Face Editing*. In contrast to the use-cases tackled by our approach, face editing does not aim to generate completely new data, but to modify existing image data (He et al., 2019; Royer et al., 2020; Choi et al., 2018; Liu et al., 2017; Lin et al., 2018). For example, (Ding et al., 2018) presented a system that is capable of transforming face images of certain emotions to other emotions in a continuous way. Although their approach does not explicitly rely on continuously labeled data, the variety regarding emotion intensity has to be represented in the training set. I.e., face images of varying emotion intensity have to be available, even if the intensity degree itself does not have to be known for every sample. They showed that their system is even capable of generating random new faces that show a certain emotion. However, they did not investigate if those newly generated face images can be controlled continuously. Also, they did not evaluate their system with respect to common dimensional models as the valence-arousal model, but rather explored if the intensity of the categorical emotions can be altered by their face editing system.

To the best of our knowledge, there exists no prior work that focuses on the generation of new emotional faces in a continuous way while using only discrete label data for training.

3 APPROACH

The following sections introduce our approach to use a cGAN that was trained on discrete emotions to generate faces showing continuous degrees of valence and arousal. We discuss differences in discrete and continuous emotion models and why we favour the latter approach.

3.1 Emotion Models

A categorical model subsumes emotions under discrete categories like happiness, sadness, surprise or anger. There is a common understanding of these emotional labels, as terms describing the emotion classes are taken from common language. However, this approach may be restricting, as many blended feelings and emotions cannot adequately be described by the chosen categories. Selection of some particular expressions can not be expected to cover a broad range of emotional states and could suffer from randomness.

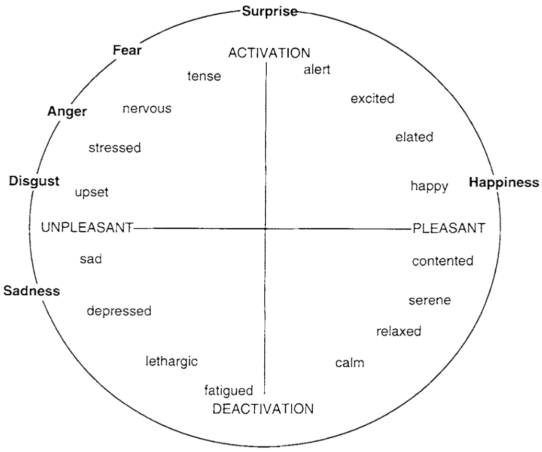

Figure 1: Russel's 2-dimensional valence arousal circumplex (Russell and Barrett, 1999).

An arguably more precise way of describing emotions is to attach the experienced stimuli to continuous scales within dimensional models. (Mehrabian, 1995) suggests to characterize emotions along three axes, which he defines as pleasure, arousal and dominance. (Lang et al., 1997) propose arousal and valence as measurements. These representations are less intuitive but allow continuous blending between affective states. They describe multiple aspects of an emotion, the combination of stimuli's alignments on these scales defines single emotions. In the case of the valence-arousal model, which is the more commonly used dimensional model, the valence scale describes the pleasantness of a given emotion. A positive valence value indicates an enjoyable emotion such as joy or pleasure. Negative values are associated with unpleasant emotions like sadness and fear. This designation is complemented by the arousal scale which measures the agitation level of an emotion (Figure 1).

Categorical as well as dimensional models are simplified, synthetic descriptions of human emotions and are not able to cover all of the included aspects. They are however useful and needed to model emotions as concepts to be presented to a machine. In our generation system for expressive faces, we prefer a dimensional approach over the categorical, in order to enable a seamless transition between displayed emotions. However, data collections featuring dimensional annotation for facial expressions are more sparse than the ones containing categorical labels (Section 4.1).

To alleviate this problem, we propose to use a cGAN that was conditioned on categorical emotions during training, and interpolate between those emotions in order to be able to create new images. Those newly generated face images show emotional states that are located in the continuous dimensional space of the valence/arousal model without having to correlate directly with discrete emotion categories.

3.2 Network Architecture

The network used in our experiments is largely based on a *Deep Convolutional GAN* (DCGAN) by (Radford et al., 2015). A detailed description of DCGAN can be found in their respective publication. Summarized, DCGAN follows the basic principles of traditional GANs while making use of convolutional and convolutional-transpose layers in order to improve the quality of generated outputs.

In order to allow for the targeted generation of images, the DCGAN was extended with the principles of a *cGAN*. In contrast to conventional GANs, cGANs add a conditioning component to the input vector. That vector is used to condition the generator network to certain features that shall be shown in the output images. Thus, during training, those feature information has to be fed as labels. All in all, the input to a cGAN consists of a random noise part z and a conditioning vector v. After a successful training procedure, the generator has learned to transform the random noise input to images that resemble the training domain, whereas the conditioning information is taken into account to direct those outputs to show the desired features. In the context of emotional face generation, the random noise part is responsible for the face itself, whereas the conditioning information leads to certain emotions of the face. Thus, two identical noises conditioned with different feature information should result in the same face showing different emotions.

In our implementation, the conditioning information is given to the network as one-hot encoded label vector, where each element represents a certain emotion. As described in Section 4.1, the emotions *Neutral, Sad, Disgust, Fear, Angry* and *Happy* were used during training. Thus, the one-hot label vector v has the following form:

$$v = (v_1, v_2, ..., v_6) = \{0,1\}^6 \tag{1}$$

with

$$\sum_{i=1}^{6} v_i = 1 \tag{2}$$

3.3 Interpolation

During inference, we change the definition of the conditioning vector in order to allow a continuous interpolation between discrete emotional states. As introduced in Section 3.1, we derive the mapping between discrete emotions and continuous emotional states (i.e., valence-arousal space) from the idea of Russel's emotion system. Thus, the valence-arousal of a face image I can be represented by a tuple $VA(I) = (v, a)$, where v refers to the valence value and a to the arousal value. According to Russel's emotion system, an image x with $VA(x) = (0, 0)$, thus, representing the center of the emotion space, would show a neutral emotion. Certain emotions, like *Happy*, are represented by valence/arousal states that show quite extreme values. Therefore, to generate images with certain degrees of arousal or valence, we interpolate between those *extreme* emotions and the *neutral* emotion. With the term *extreme* emotions, we refer to all of the used categorical emotions except *Neutral*. By applying the interpolation technique, we do not use strictly one-hot encoded label vectors as conditioning information during inference, but interpolated label vectors that do not have to be of a binary structure. In order to do so, we have to redefine the conditioning vector v so that the single conditioning elements are not forced to a binary structure, but can take values in the interval $[0, 1]$:

$$v = (v_1, v_2, ..., v_6) = [0, 1]^6 \tag{3}$$

In order to keep the output of the conditioning layer consistent to the training, we found that retaining the restriction formulated in Equation 2 leads to better quality of the interpolated results than choosing the conditioning elements arbitrarily in the given intervals. In our experiments, we have adhered to perform interpolations between *Neutral* and a certain other emotion to maintain comparability between the emotions. It should be noted that the approach could easily be extended to interpolate between two extreme emotions. However, as we only use one extreme emotion and *Neutral* at once, the following constraint has to be added:

$$\exists i_{\in[2,6]} : v_1 + v_i = 1 \tag{4}$$

where v_1 represents the condition for *Neutral*. To generate an image that shall show a certain degree of valence v or arousal a, where $0 \leq a, v \leq 1$, we use the one-hot element of the an emotion that maximizes the specific value, for example *Happy* when dealing with valence, and decrease it to the desired degree, while simultaneously increasing the one-hot element that refers to *Neutral* to the same extent. Our hypothesis is, that due to the differentiable function that is approximated by the cGAN model during the training, those non-binary conditioning vectors lead to image outputs which are perceived as showing non-extreme emotions, thus, emotions that have can have valence/arousal values located anywhere in Russel's emotion system instead of just showing those values that are given during the training.

4 EXPERIMENTS & RESULTS

To evaluate, if label interpolation in the conditioning space is a valid approach to generate images with controllable valence-arousal values, we trained a cGAN model described in the previous chapter.

4.1 Dataset & Training

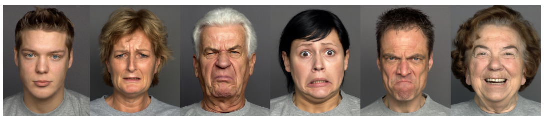

Figure 2: Exemplary data from FACES showing neutral, sad, disgust, fear, anger and happiness from left to right varying the age group.

As described in Section 1, datasets that are labeled with respect to dimensional emotional models are rare. Further, those few datasets that contain those continuous label information (like *AffectNet* by (Mollahosseini et al., 2019) or AFEW-VA by (Kossaifi et al., 2017)) are recorded in the wild. Usually, variability in data is an advantage for deep learning tasks, however, for many generative tasks that focus on modeling specific feature dimensions like emotions, variety in data leads to inconsistent and unnatural results. Thus, for our usecase, a uniformly recorded dataset with low variability outside the facial emotion is required. To address the lack of continuously labeled and consistently recorded datasets meeting these requirements for emotional face generation, we explore the use of label interpolation to overcome the disadvantages of categorical labeled datasets. Such datasets containing discrete emotion labels are widely available. A dataset fitting our purpose particularly well is the FACES dataset (Ebner et al., 2010). The FACES dataset contains 2052 images showing emotional facial expressions from 171 men and women of different ages (58 young, 56 middle-aged and 57 old). Each participant shows two versions of the emotions *Neutral*, *Sadness*, *Disgust*, *Fear*, *Anger* and *Happiness*. Despite the relatively small size of the dataset, the consistency of the images in terms of lighting, background, and viewing angle makes it stand out as a training dataset and an adequate choice for training GANs. For instance, all images in the dataset have the same blank background and the subjects wear the same plain grey top as depicted in Figure 2. All images were resized to the target size of 256x256 pixels prior to training. No further preprocessing was necessary.

The model was trained for 10,000 epochs on all 2052 images of the FACES dataset using *Adam* opti-

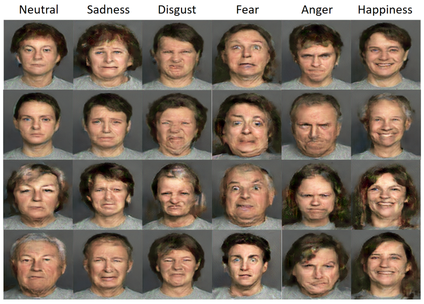

Figure 3: Example outputs of the trained cGAN model.

mizer with a learning rate of 0.0001. Example outputs of the trained model, conditioned on one-hot vectors of all 6 used emotions, are shown in Fig. 3.

Our evaluation process is two-folded. First, we wanted to evaluate whether extreme emotions can be generated by the cGAN without interpolation and whether this output is recognized as the correct emotion by real humans. Secondly, we decided to perform a more fine-grained analysis on the continuously generated outputs by the use of our approach. In order to do so, we decided to perform an evaluation using a pretrained valence-arousal regression model. As the focus of this paper is the exploration of the possibility to interpolate between discrete label information, the absolute numerical values of such a regression model are not a good metric for our purposes. Instead, we want to explore if the interpolation mechanism is able to model the full bandwidth of valence and arousal values that can occur between a neutral emotion and certain extreme emotions. Thus, in the first step, we conducted a tiny user study to evaluate if the trained model itself is capable of generating those extreme emotions that it was trained on. The second evaluation step was designed as a computational evaluation of the interpolation mechanism to explore if the trained cGAN is capable of modeling a smooth transition between the emotional states in terms of valence and arousal.

4.2 User Evaluation of cGAN Model

In our user study, we evaluated the capabilities of the cGAN to create images of discrete emotions that were generated using the corresponding one-hot vector encoding. In total, 20 participants of ages ranging from 22 to 31 years (M = 25.8, SD = 2,46, 40% male, 60% female) took part in the survey.

The survey consisted of 36 images, half of which were selected out of the FACES dataset, while the other half was generated by our trained cGAN. In total, 6 images were shown for each emotion, 3 from the FACES dataset and 3 generated faces. The images that were taken from the FACES dataset were resized to 256*x*256 pixels in order to keep consistency with the artificially created images. For every image and emotion, the participants were asked how much they agreed with the image showing a certain emotion by the use of a 5-point Likert scale (1 = strongly agree, 5 = strongly disagree).

The results are shown in Fig. 4. As can be seen, the artificially generated images were perceived in a convincingly similar way as the original images from the FACES dataset. For every emotion, it becomes clearly visible that the emotion is predominantly recognized correctly by the participants. The emotion *Sadness* stands out, as here, generated images were recognized even better than the original ones, as *sad* images from the FACES dataset more often were confused with *Disgust*. Thus, the trained cGAN model turns out to be a suitable base for further interpolation.

4.3 Computational Evaluation of the Interpolation Mechanism

As was shown in the user study, the cGAN is able to model discrete emotions when conditioned on one-hot vectors. To verify if our approach based on interpolating in that label space is enhancing the network with the ability to generate images with varying, continuous degrees of valence and arousal, we performed a computational evaluation. In order to do so, we fed 5,000 random noise vectors into the cGAN, 1,000 for each emotion. The conditioning vector was initially conditioned on a neutral emotion. Additionally, for each noise vector, interpolation steps towards the respective extreme emotion were conducted. Therefore, 10 interpolation steps in intervals of 0.1 were done per noise vector, so that the last interpolation step equals a one-hot vector that is conditioned on the corresponding extreme emotion. Each of the resulting images was then evaluated with a pre-trained valence/arousal-assessment model. Example outputs of various interpolation steps between *Neutral* and the other five emotions are depicted in Fig. 5.

The structure of that model was based on the MobileNetV2 architecture (Sandler et al., 2018), which was adapted for multi-task learning of the following two tasks: recognition of continuous valence/arousal values and detection of eight discrete emotion classes (*Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger* and *Contempt*). We selected this architecture because it achieves similar results to Inception- or ResNet-based models but can be trained more quickly. The network was trained for both tasks simultaneously on the AffectNet dataset (Mollahosseini et al., 2019) us-

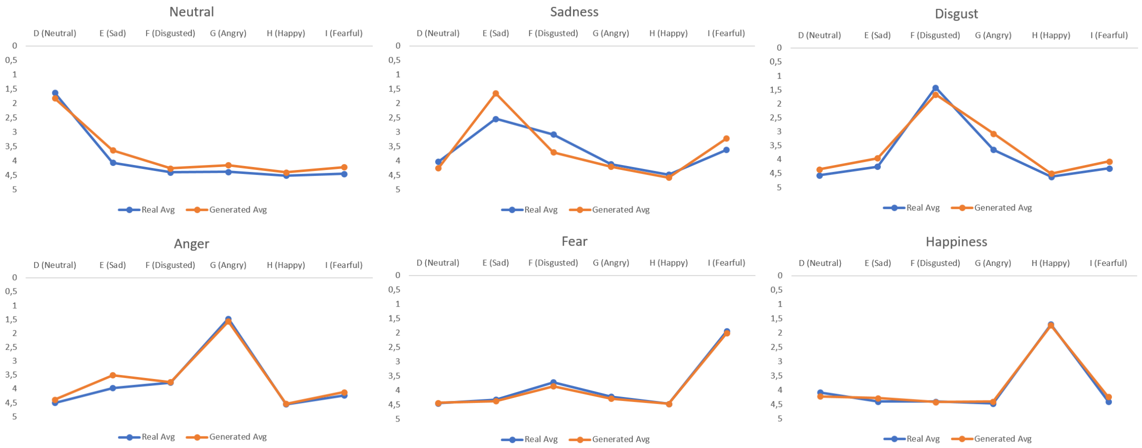

Figure 4: Results of the user study. Blue graphs show the perceived emotion of real images from the FACES dataset, while orange graphs show the perceived emotion of outputs of the cGAN conditioned on one-hot vectors. The y-axis represents the degree of the participant's agreement with the corresponding emotions that are represented by the x-axis.

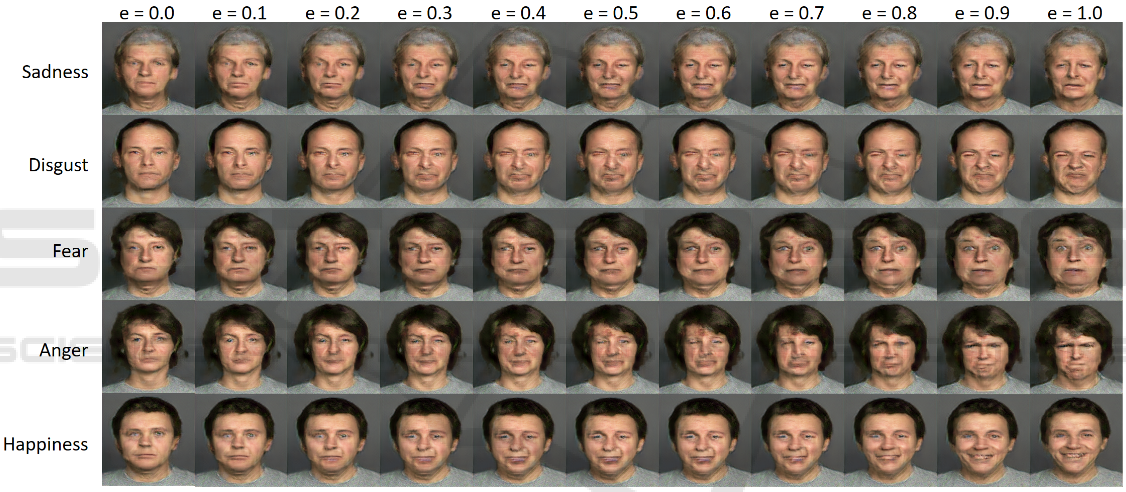

Figure 5: Example outputs of the interpolation mechanism. Each row shows a set of interpolation steps, where in each step, the emotion portion e was increased by 0.1, whereas the neutral portion was decreased by the same amount.

ing an *Adam* optimizer with a learning rate of 0.001 for 100 epochs. After training, the secondary model head for discrete emotion detection was removed and only the valence/arousal head was used to evaluate the cGAN model.

We assessed the valence/arousal values for every output image of the interpolated cGAN and averaged them over the 1,000 samples per emotion. The results are shown in Fig. 6.

5 DISCUSSION

As can be derived from the plots depicted in Fig. 6, the interpolation mechanism is able to condition the cGAN to produce face images of various valence/arousal values. Further, those values that are taken during the interpolation procedure are mainly located in the intervals defined by the respective start and end points of the corresponding interpolation, i.e., the samples that are conditioned with a binary one-hot vector.

However, it can be seen that in the cases of *Sadness* and *Disgust*, during interpolation, the valence first takes slightly higher values than the respective valence values at the starting sample, before finally descending to the level of the interpolation end point. For *Anger*, both valence and arousal values are going up and down until finally arriving at a similar level as where they started. It should be noted, that the plots are showing only the valence and arousal values that were tracked by the valence/arousal assess-

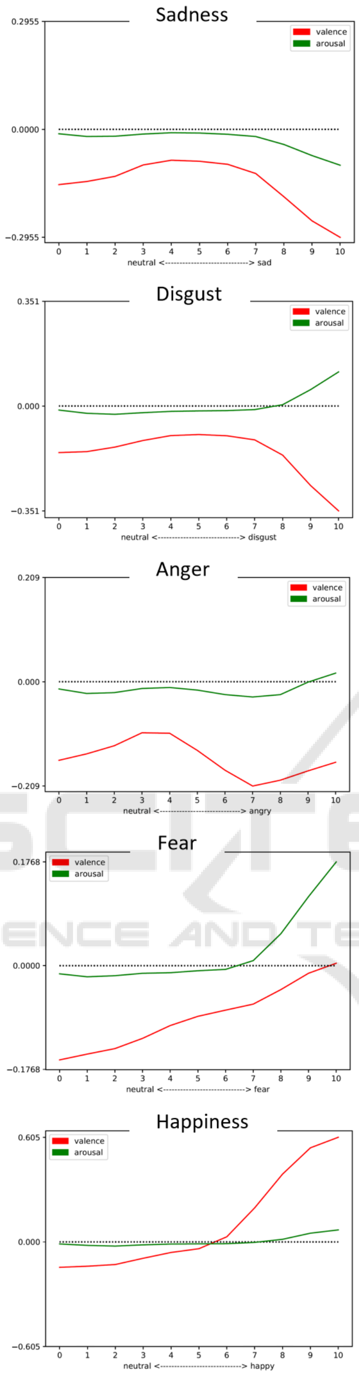

Figure 6: Computational Evaluation of our interpolation approach. Red graphs show valence, while green graphs show arousal. The x-axis represents the interpolation steps. Each interpolation step was performed by increasing the corresponding emotion vector element by 0.1, while decreasing the neutral vector element by 0.1.

ment model described in the previous section. That model seems to attribute similar valence/arousal values to generated angry faces as to generated neutral faces when conditioned with a binary one-hot vector, although the user study has shown that both of those emotions were perceived in a correct way. As can be seen in Fig. 1, *Anger* should show substantially different valence and notably different arousal than a neutral emotion according to Russel's emotion model. Thus, we assume that the valence/arousal assessment model has its flaws when it comes to dealing with certain emotional states. The fact, that the valence values of neutral faces already are located notably below zero, as well as the observation that images that are showing *Fear* are averaging in a higher valence than *neutral* images, although also being perceived correctly in our user study, strengthens this assumption. However, the values are predominantly evolving into the right direction during the whole interpolation procedure and modeling the whole range in the desired valence and arousal intervals between the discrete emotions. Thus, the interpolation mechanism can indeed be used to generate face images of continuous emotional states, generating a variety of samples in the dimensional valence-arousal model. The fact, that the values are not evolving in a linear way, i.e., the plots appear rather as curves than as straight lines, does not take away much from the results, as for a more even interpolation, the single interpolation step intervals can easily be modified, e.g., instead of using the same step interval for every single interpolation step, higher intervals can be used in ranges where the target features are changing slower.

6 CONCLUSION & OUTLOOK

In this work, we have explored the capabilities of a continuous interpolation through a discrete conditioning space of a cGAN. We strived for a possibility to generate images of emotional faces, where the emotions are not restricted to categorical structures, but can transition freely in the valence-arousal space. Our experiments showed that our interpolation mechanism is able to achieve that goal, although the performance of the approach is heavily dependent on the emotion that is used for the interpolation. All in all, the approach shows great potential to be applied as a tool for continuous emotional face generation. In future work, we plan to extend our work by applying it to more complex cGAN architectures in order to optimize the quality of generated images. Further, we plan to examine the applicability of label interpolation by the use of other datasets that show a higher diversity.

ACKNOWLEDGEMENTS

This work has been funded by the European Union Horizon 2020 research and innovation programme, grant agreement 856879.

REFERENCES

Arjovsky, M. and Bottou, L. (2017). Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). 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.

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

Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 42(1):351–362.

Gauthier, J. (2014). Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2.

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

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*.

He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. (2019). Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478.

Hong, Y., Hwang, U., Yoo, J., and Yoon, S. (2019). How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys (CSUR)*, 52(1):1–43.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36.

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. In Lang, P. J., Simons, R. F., and Balaban, M. T., editors, *Attention and orienting: Sensory and motivational processes*, pages 97–135. Psychology Press.

Lin, J., Xia, Y., Qin, T., Chen, Z., and Liu, T.-Y. (2018). Conditional image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5524–5532.

Liu, M.-Y., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*.

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 121(3):339–361.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Royer, A., Bousmalis, K., Gouws, S., Bertsch, F., Mosseri, I., Cole, F., and Murphy, K. (2020). Xgan: Unsupervised image-to-image translation for many-to-many mappings. In *Domain Adaptation for Visual Understanding*, pages 33–49. Springer.

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Wang, Y., Dantcheva, A., and Bremond, F. (2018). From attributes to faces: a conditional generative network for face generation. In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE.

Yi, W., Sun, Y., and He, S. (2018). Data augmentation using conditional gans for facial emotion recognition. In *2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama)*, pages 710–714. IEEE.