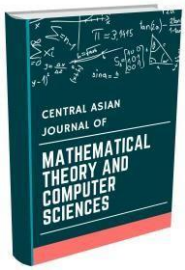

CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES

https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS

Volume: 06 Issue: 01 | January 2025 ISSN: 2660-5309

Article

Facial Expression Manipulation via Latent Space Generative Adversarial Network

Wafaa Razzaq[*1]

1. College of Nursing, University of Thi-Qar, Nasiriyah, Iraq

* Correspondence: wafaa_razzaq@utq.edu.iq

Abstract: Style Generative Adversarial Network (StyleGAN) stands out as the state-of-the-art architecture for generating highly realistic synthetic faces. Its implementation projects an image into its latent space, which can be manipulated by means of directional curves modifying features of the original image. However, its high dimensionality makes the manual search for a directionality that produces a given feature or gesture impractical. This work proposes a pseudo-auto encoder type neural architecture that manipulates the latent projection by alternating the appearance of the face. This is done by encoding the facial gesture with Action Units vectors. A dynamic of expressions was achieved that allows the transition from one gesture to another without having to go through the neutral, improving the naturalness of the gestural dynamics.

Keywords: StyleGAN2, Latent Space, Facial Expressions

1. Introduction

In this work, we will address the temporal dynamics of the portrait as an object capable of evolving in a self-generative way through the study of the change in appearance of human faces. For this, we will use synthetic images created by StyleGAN2[1], a generative adversarial network (GAN) that generates hyper-realistic faces . StyleGAN is a groundbreaking type of neural network developed by researchers for generating highly realistic images. It is particularly renowned for its ability to synthesize images that exhibit intricate details and a high degree of realism. We will seek to reflect the changes in the person's mood in their daily life in the portrait. StyleGAN allows us to edit images through manipulations in their associated latent space. These visual changes allow us to change their appearance or expression, which can be translated as a new emotional state of the model.

Related Work

The main works in the area that seek to master the dynamics of the manifold to obtain expected outputs of the model can be broadly classified by using unsupervised and supervised training.

[1-11] Unsupervised training is mainly useful in the case of manipulations where labeled datasets do not exist or are difficult to generate [3]. Supervised training uses datasets with labels given by experts to train either another network or the GAN itself in order to achieve particular manipulations, such as facial expressions. This work proposes

Citation: Razaaq, W. Facial Expression Manipulation via Latent Space Generative Adversarial Network. Central Asian Journal of Mathematical Theory and Computer Sciences 2025, 6(1), 103-108.

Received: 11th Jan 2025
Revised: 24th Jan 2025
Accepted: 31st Jan 2025
Published: 20th Feb 2025

the use of a pseudo-autoencoder type neural architecture with supervised training that, thanks to the use of Action Units (AUs) vectors [12], obtains a natural transition of a facial expression to another. Undesired effects and biases obtained from the pseudo-autoencoder training are also presented.

2. Materials and Methods

2.1 Expression Encoder (Latent Space)

Generative Adversarial Neural Networks (GAN) [13] consist of a system where a discriminator D is trained to maximize the probability of recognizing between a real image and a synthetic one, and a generator G that tries to fool it, generating realistic synthetic images from a random seed Z. Once the GAN has been trained, G performs a non-linear mapping of this random seed Z to generate a hyper-realistic image-type output of the form: $g : Z \rightarrow X$.

In the case of StyleGAN [14] its generator, instead of learning a direct mapping between Z and X, first maps Z to another intermediate Latent Space W. This space is composed of 18 channels that encode facial information (gender, background, eye color, etc.). Then, W is mapped to X where [14] shows that the entanglement in the geometry of W is smaller than the entanglement of the geometry in Z, that is, the entanglement in the geometry of W is smaller than the entanglement in the geometry of Z, that is, the entanglement in the geometry of Z is smaller than the entanglement in the geometry of Z. That is to say, it simplifies the space itself, allowing modifications of a latent vector w to remain in the model manifold. This latent space, although not completely disentangled, allows us that given $w \in W$ and a direction ϕ we can operate in its neighborhood using successive linear interpolations to explore the space in the manifold and observe modifications continuously in the data $g(w + (\phi - w) \cdot \delta)$. Where δ is a proportionality factor.

2.2 Expression Encoder(Pseudo-auto encoder architecture)

A model inspired by the architecture of autoencoders was developed. Its objective is that, from a latent vector win, a latent output vector $wout$ is projected, which corresponds to the same identity, pose, age,etc., but its facial expression is determined by a vector of the dynamics of actions units uin. An example of this transition is exemplified in Figure 2. We will see this architecture in detail in the following paragraphs.

Autoencoders are a multilayer neural architecture whose output seeks to exactly replicate the input data. Formally, let $A()$ be the autoencoder model, such that $A(x) = \hat{x}$ such that $\hat{x} \simeq x$. The parts of an autoencodercan be divided into an encoder, which projects the input into a latent space of lower dimension, and a decoder, which reconstructs the initial information without loss of information.The latent vector win is the input to the network, which consists of a two-layer encoder with 512 units each. The vector of the dynamics of the action units uin is concatenated to the latent output. This vector, with 43 elements, is copied 5 times in order to increase the connections. Then, the decoder consists of 2 layers of 512 neurons each, with the final output being $wout$, such that $|win| = |wout|$.

The vector of the dynamics of the action units uin guides the transition of the gesture of the person's face. This vector is obtained in the following way. Let a data set of photographs of subjects making different expressions, each instance is composed of the pair (IiS, aiS), where IiS is the capture of subject i making gesture S, and aiS corresponds to the vector of the 43 action units completed manually from IiS. When an action unit is represented in IiS it takes a value of 1 in the vector. Otherwise it takes a value of 0. Then, as the objective isthat the pseudo-autoencoder finds the transition from win to $wout$ we need the information of those novelties regarding the vectors of action units that correspond to each photograph. The transition from a gesture to S to T, the vector dynamic would be obtained $u_{in}^{s-T} = aiT - aiS$.

Each element k in u_{in}^{s-T}, corresponds to the action unit in position k of aiT and aiS taking a value:

$$\mathbf{u}_{in}[k] = \begin{cases} 1 \text{ if } aiT\ [k] \text{ is activated and as is not activated } aiS \\ 0 \text{ if there is no change in this action unit between } ais[k] \text{ and } aiT\ [k] \\ -1 \text{ if this action unit is deactivated in } aiT\ [k] \end{cases}$$

3. Results

3.1 Training Dataset

The training of the pseudo-autoencoder requires a training dataset with N gestures per person and their corresponding action unit. The dataset [15] captures 230 subjects performing 22 pre-established gestures, and also has the action units of each image in a binary format (0 not detected, 1 detected).

We used a pretreatment where each face image was aligned using MTCNN [16] to a position compatible with the pre-trained model StyleGAN2. Then all the latent vectors wSi of each image were obtained, which added to the action units vector a^{i}s make up the data tuple of subject $\boldsymbol{i}$ performing gesture $\boldsymbol{S}$.

3.2 Pseudo-Autoencoder training

The training set consists of the inputs X and the expected outputs Y , computed using the OSU set and its labels. The input set $X = \{(w11, u1 - 21), (w11, u1 - 31), (w11, u1 - 41), \dots (w22230, u22 - 21230)\}$ and the expected values are, respectively, $Y = \{w21, w31, w41, \dots, w21230\}$.

Then, the sample j of $X(j) = (wSi, uS - Ti)$ and its target is $Y(j) = wTi$, which means that given the input of the latent vector (wSi) with the expression S, the output has to have the expression in the latent vector wTi. This defines the loss function as $Lj = ||wTi - \hat{wTi}||$, where $\hat{wTi}$ is the latent vector estimated by the pseudo-autoencoder having as input $X(j)$.

The model was trained for 150 epochs using an Adam optimizer [17] with a learning rate equal to $10e - 3$ and a weight decay of $10e - 5$. In addition, a learning rate decay with a factor $\gamma = 0.5$ was established every 10 epochs, for a smooth learning towards the higher epochs.

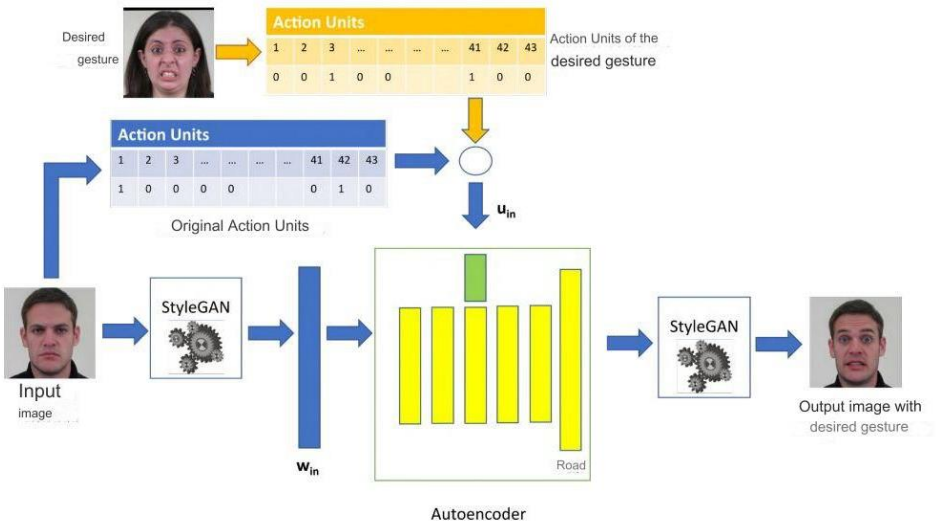

Figure 1. Operation of the pseudo-autoencoder: An input image with its vector of AUs is projected by the model to a latent vector that has the desired expression, defined by the AUs of the image of the woman.

4. Discussion

Given an input image, whose projection we will call ws, and a vector of AUs that describe the desired gesture gt we can generate, thanks to the pseudo autoencoder, the wt associated with gt. We know that images change linearly if we interpolate linearly between two latent spaces[14], we also know that $v = wt - ws$ is true where v is the direction that takes you from ws to wt. This allows us to use the property that $wi = ws + v * \alpha, \alpha \in [0,1]$, where $wi = ws \iff \alpha = 0 \wedge wi = wt \iff \alpha = 1$. Let us see that by modulating this parameter α we can generate intermediate images between the target and the initial image. As an example, in Figure 2 we start of 3 images from the OSU dataset with their respective ws, where for each ws we want to reach the same gesture (g_{happy} that has its wc associated). We use the pseudoautoen coder with its respective AU's to obtain the wc. The transition of images from one expression to another is achieved by $wi = ws + (wc - ws) * \alpha, \alpha \in \{0, 0.25, 0.50, 0.75, 1\}$ with wi being the vector associated with each intermediate image.

Figure 2. Transition from one expression to another given by a linear interpolation between two latent vectors.

One of the problems we encountered when generating the faces was the loss of identity of the target face after passing through the pseudo-autoencoder. Although the process works very well for modifying gestures, the face loses many of the factors that identify it, as shown in Figure 3. This is because the training dataset only contains the faces and expressions of 230 people, therefore the resulting image imitates the features of that group.

Figure 3. Pseudo-autoencoder bias when projecting real images outside.

5. Conclusion

This work proposes a method for the generation of synthetic images corresponding to a sequence of natural gestures, which can be associated with the person's mood. GAN networks are shown as a complete tool, on which the work can be built and given different imprints. Several points remain as perspectives that we elucidated in the development of the demo, such as the incorporation of other manipulations, pose movements or alterations in age, which would allow enriching the final product. On the other hand, the biases that we detected when projecting an identity by the autoencoder show us the importance of carefully handling the input data.

REFERENCES

[1] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzingand improving the image quality of stylegan," in CVPR, pp. 8110–8119, 2020. https://doi.org/10.48550/arXiv.1912.04958

[2] Abdal, Rameen, et al. "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows." ACM Transactions on Graphics (ToG) 40.3 (2021): 1-21. https://doi.org/10.1145/3447648

[3] Oliva, Aude, and Phillip Isola. "Ganalyze: Toward visual definitions of cognitive image properties." Journal of Vision 20.11 (2020): 297-297. https://doi.org/10.1167/jov.20.11.297

[4] Khosla A, Xiao J, Isola P, Torralba A, Oliva A (2012) Image memorability and visual inception. SIGGRAPH Asia. https://doi.org/10.1145/2407746.2407781

[5] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions inthe gan latent space," in International conference on machine learning, 2020. https://doi.org/10.48550/arXiv.1811.10597

[6] C. Tzelepis, G. Tzimiropoulos, and I. Patras, "Warpedganspace: Finding non-linearrbf paths in gan latent space," in ICCV, 2021. https://doi.org/10.48550/arXiv.2109.13357

[7] Liu, Yunfan, et al. "Towards spatially disentangled manipulation of face images with pre-trained StyleGANs." IEEE Transactions on Circuits and Systems for Video Technology 33.4 (2022): 1725-1739.https://doi.org/10.1109/TCSVT.2022.3213662

[8] Speck, Daniel, et al. "The Importance of Growing Up: Progressive Growing GANs for Image Inpainting." 2023 IEEE International Conference on Development and Learning (ICDL). IEEE, 2023. https://doi.org/10.1109/ICDL55364.2023.10364530

[9] Y. Fan, F. Tian, X. Tan, and H. Cheng, "Facial expression animation throughaction units transfer in latent space," Computer Animation and Virtual Worlds, 2020. https://doi.org/10.1002/cav.1946

[10] Tang, Bingyin, and Fan Feng. "Efficient and expressive high-resolution image synthesis via variational autoencoder-enriched transformers with sparse attention mechanisms." Journal of Electronic Imaging 33.3 (2024): 033002-033002. https://doi.org/10.1117/1.JEI.33.3.033002

[11] Ding, Saisai, et al. "High-resolution dermoscopy image synthesis with conditional generative adversarial networks." Biomedical Signal Processing and Control 64 (2021): 102224. https://doi.org/10.1016/j.bspc.2020.102224

[12] P. Ekman and W. V. Friesen, "Facial action coding system," Environmental Psychology & Nonverbal Behavior, 1978. https://psycnet.apa.org/doi/10.1037/t27734-000

[13] Imamverdiyev, Yadigar N., and Firangiz I. Musayeva. "Analysis of generative adversarial networks." Problems of Information Technology (2022): 22-30. http://doi.org/10.25045/jpit.v13.i1.03

[14] Barzilay, Noa, Tal Berkovitz Shalev, and Raja Giryes. "MISS GAN: A multi-IlluStrator style generative adversarial network for image to illustration translation." Pattern Recognition Letters 151 (2021): 140-147.https://doi.org/10.1016/j.patrec.2021.08.006

[15] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion,"Proceedings of the National Academy of Sciences, vol. 111, no. 15, pp. E1454–E1462, 2014. https://doi.org/10.1073/pnas.1322355111

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment usingmultitask cascaded convolutional networks," IEEE signal processing letters, vol. 23,no. 10, pp. 1499–1503, 2016. https://doi.org/10.1109/LSP.2016.2603342

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization. iclr. 2015. https://doi.org/10.48550/arXiv.1412.6980