Let's put a smile on your face

Marcis Teodors Upenieks[1,2], Evalds Urtans[1,2]

[1] *Riga Technical University, Kipsalas iela 6A, Riga, LV-1048, Latvia*

[1] *asya.ai, Pils iela 17, Ventspils, LV-3601, Latvia*

Abstract

This study compares three methods for the facial expression transfer task using adversarial generative networks. The facial expression transfer task aims to transfer one human facial emotion to another emotion without affecting the identifying features of the human face. The image-to-image method uses the whole image for style transfer, and the novel face-to-face and parts-to-parts methods use only segmented features of the face to do the style transfer. The results show that the image-to-image method achieves a precision of 69.7% and an FID score of 21.67, face-to-face achieves a precision of 78.6% and an FID score of 17.6, and finally the parts-to-parts method achieves a precision of 97.8% and an FID score of 17.37 to transfer emotion from neutral to happiness in a photo of a face.

Keywords

Deep Learning, Generative Adversarial network, Emotions, Style transfer, Facial Expressions, Semantic segmentation

1. Introduction

In recent years, generative adversarial networks have gained popularity among generative machine learning models. At first, they were used for simple image generation tasks [1], [2], [3], [4], but later started to be used for style transfer tasks [5], [6]. The last five years have seen an increase in the number of research papers published towards style transfer and utilization of generative adversarial networks. Additionally, publicly available facial expression and emotion data sets have increased. The image style transfer problem deals with the transfer of style between two image domains, for example, zebras, to horses, or vice versa [6]. In this work this principle is used to transfer facial expressions. Face emotion and expression style transfer is useful for a number of tasks like augmenting existing unbalanced facial datasets, video processing where only one particular emotional state is accessible, and video game avatar creation, where you could use reference image of yourself and transfer facial expression to in-game avatar. It could also be used to improve face re-identification systems to augment target face with different facial expressions. Many other use cases can be found where human-machine interaction is important. For example, emotion style transfer can be useful for emotion expression in virtual assistents or as a filter in remote video meetings. As a result, it improves

the quality of remote meetings, customer service, or other services where computer-generated avatars are used. This paper compares three methods: image-to-image, face-to-face, and parts-to-parts on face emotion and expression style transfer using generative adversarial networks. The first method uses the entire photo from the dataset to achieve style transfer, but the second and third methods, which are new and original work, use face segmentation to extract specific face regions like eyes, eyebrows, nose, and mouth. The second method uses all segments together, and the third method uses each segmented part separately.

2. Related work

The Generative Adversarial Network (GAN) [7] was proposed as a framework for generative models using adversarial training, where two networks are used. Generator to create images and a discriminator to classify generated images as real or fake. Deep Convolutional Generative Adversarial Networks [2] were proposed primarily for image data and have been shown to be better than the original GAN, since it leverages Convolutional Neural Networks to encode and decode images, which are known to be powerfull feature extractors. Style transfer task between two different unpaired images from different domains using GANs was introduced with CycleGAN [6]. Style transfer for images uses the style of the objective to transfer the source image to the target image, for example, to transfer a photo of neutral emotion in the source dataset to the smiling emotion photo in the target dataset [5], [8]. GAN models have been shown to be unstable to train and hence to improve the convergence of GAN during training, Wasserstein GAN was proposed [4]. Gradient clipping, which was used in Wasserstein GAN, was found to be ineffective and sometimes even degraded the quality and diversity of the generated images, and therefore, instead of clipping the gradients, it is more effective to implement a gradient penalty term [9]. Recently, even more improvements have been made in the style transfer task by PatchGAN [6] and StarGAN [8]. GAN models have been applied for the emotion style transfer task as listed in previous survey papers [10], but this paper introduces robust experimental metrics to measure the effectiveness of existing methods and novel using two datasets.

3. Methodology

The image-to-image method uses original images from the dataset that include facial expression, but can also include background artifacts that can degrade the performance of the style transfer task. This paper proposes a novel face-to-face method that only segments the face region from the original image and uses it for the style-transfer task. Similarly, novel parts-to-parts method segments face regions that express emotions, e.g., mouth and eye regions, and use them to train models. For both methods, a pre-trained face segmentation model is used to extract specific face parts from the original image. The semantic segmentation model was trained on CelebAMask-HQ [11] using the DeepLabV3 [12] model. Three methods for face segmentation are shown in Figure 1.
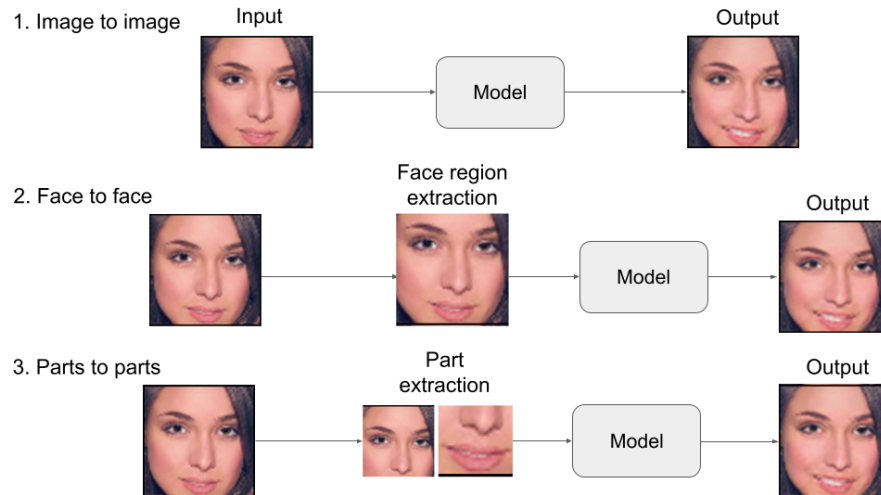
Figure 1: Segmentation methods used for style transfer inputs.

For face-to-face and parts-to-parts methods, it is necessary to apply a blending algorithm to place the generated parts back in to the original image and reduce the changes in brightness around the edges. To achieve blending, the mean pixel brightness value is calculated for the background and part itself and adjusted with the original image, and then the linear interpolation of the pixel values between each edge of the image is performed. The visual representation of the edge blend algorithm is shown in Figure 3, and the brightness adjustment algorithm is visualized in Figure 2.
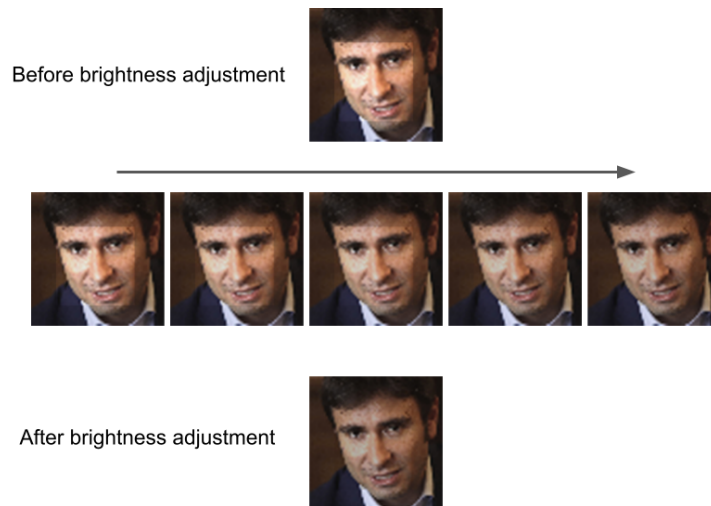

Figure 2: The brightness adjustment algorithm.

Figure 3: The linear interpolation blending algrithm.

3.1. Datasets

Two data sets used in this study are AffectNet [13] and FERPlus [14]. Both datasets contain photos of emotional facial expressions, but FERPlus has grayscale images, while AffectNet uses 3-channel color images. AffectNet consists of 440 thousand labeled samples, and FERPlus has only 35.8 thousand samples. From each dataset, 6000 thousand random samples were taken for training and 2048 samples for testing. This was chosen for the minimal requirements for the calculation of the FID score using limited hardware resources. Both datasets were resized to 64x64 pixels and filtering procedures were applied using segmentation of parts of the face, and only those images that contain all the necessary facial features were left in the datasets.

3.2. Model

StarGAN [8] and CycleGAN [6] models were used to compare three methods for the transfer of facial emotions. The main idea of CycleGAN lies in the cyclic reconstruction loss training procedure, where two generator models are used to convert source image to target and then back to source image, where the loss is calculated, hence cyclic. While Cycle-GAN requires two generator and disindicator models per training, Star-GAN only requires one of each because of its use of class label information in each training step.

For both models, Wasserstein loss is used, and gradient clipping is replaced with gradient penalty. Adam optimizer is used with beta values (0.5, 0.99), learning rate is 1e-4 similar to that used in Wasserstein GAN paper. Models are trained with batch size of 64 for 5000 epochs with generator being trained every 5th iteration allowing discriminator to learn differences between two domain images and give usefull feedback to generator model. For CycleGAN model cycle loss lambda of value 10 is used and identity loss lambda value is 5. StarGAN lambda value for reconstruction is 5. The number of parameters for the generator model in StarGAN is 7.2M and in CycleGAN 3M while for the discriminator model the number of parameters is 11.1M for

StarGAN and 3M for CycleGAN. To help solve the balance between discriminator and generator model training, a Gaussian noise is applied at the start of the training before all convolutional layers in the discriminator model with a decay rate of 0.01 and a standard deviation of 0.1.

Layers used in generator and discriminator for Cycle-GAN and Star-GAN are following - for image size reduction convolution layers are used and for upsampling image dimensions Bilinear upsampling functions are used. After downsampling, residual skip connections [15] are used. For the discriminator, convolution downsampling layers are used that reduce image dimensions by a factor of two after every downsample. Instance normalization [16] is used for generator networks and LayerNorm [17] for discriminator networks with LeakyReLU [2] as activation function in both models. Finally, the last discriminator output layer has the PatchGAN convolution configuration [18]. Please refer to the according model papers for exact layer parameters used [6], [8].

3.3. Metrics

Emotion accuracy was acquired using ResNet-based facial expression classifier pretrained on the full FERPlus dataset that is used to predict the emotion class after style transfer is performed. Frechet Inception Distance (FID) [19] is used to evaluate the quality and diversity of the generated image against the original image data set. To calculate FID 2048 generated images and 2048 original images are taken from same domain. FID uses features extracted from the last layer of Inception V3 architecture where the number of features is 2048 and since FERPlus has classes where the number of samples is smaller, 2048 was the smallest usable number, since a smaller value would allow imaginary parts to appear when calculated FID [19].

4. Experiments

Experiments were done using neutral to hapiness and hapiness to sadness emotion samples from the AffectNet dataset. These exact emotions were chosen because of their balanced sample count in the dataset and because these are significantly different facial expressions. The purpose of the study is not to compare all possible combinations, but just to evaluate three methods using robust metrics. Table 1 shows the best results for neutral-to-happiness emotion style transfer task from all three methods. In addition, different emotion transfer task for happiness-to-sadness have been tested using image-to-image and face-to-face methods as shown in Table 2.

Table 1
Neutral-to-Happiness emotion style transfer results.

Method	Model	FID	Accuracy
Image-to-image	Star-GAN	21.67	0.68
Image-to-image	Cycle-GAN	42.5	0.64
Face-to-face	Star-GAN	17.7	0.78
Face-to-face	Cycle-GAN	19.27	0.51
Parts-to-parts	Star-GAN	**17.37**	**0.97**
Parts-to-parts	Cycle-GAN	24.32	0.90

Table 2
Happiness-to-Sadness Neutral emotion style transfer results.

Method	Model	FID	Accuracy
Image-to-image	Star-GAN	24.3	**0.79**
Image-to-image	Cycle-GAN	49.65	0.76
Face-to-face	Star-GAN	**21.7**	0.77
Face-to-face	Cycle-GAN	26.4	0.77
Parts-to-parts	Star-GAN	21.9	0.78
Parts-to-parts	Cycle-GAN	25.8	0.75

The best qualitative visual results for the neutral-to-happiness transfer style using the face-to-face method are shown in Figure 4 and parts-to-parts method in Figure 5. The transfer of emotion style to the face using different combinations of target emotions is shown in Figure 6.

Figure 4: Neutral-to-Happiness emotion style transfer results using face-to-face method and AffectNet dataset.
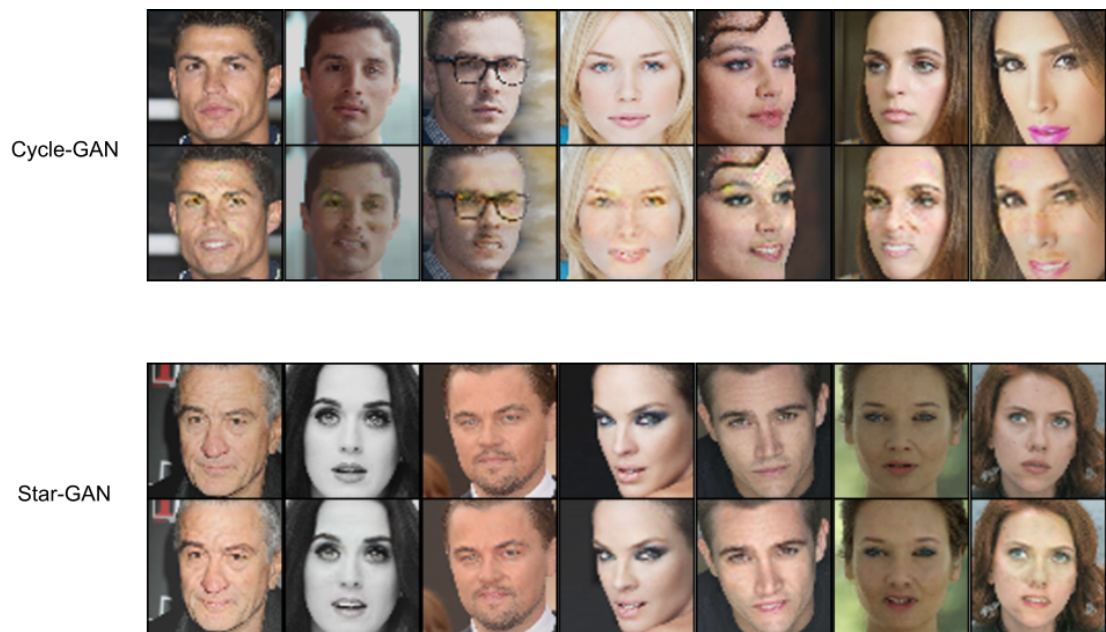

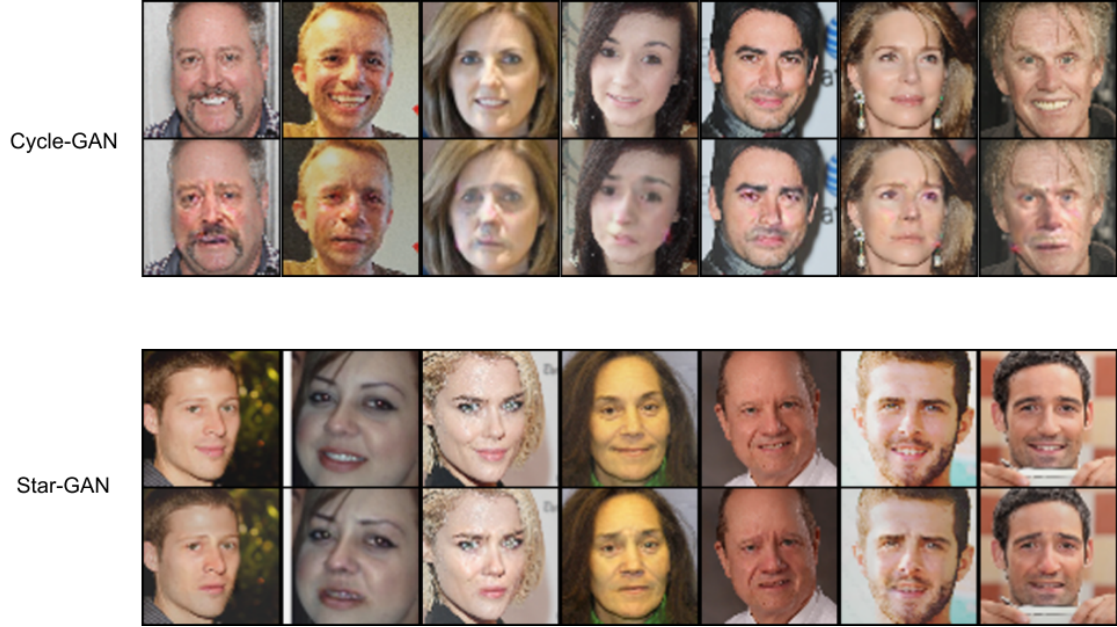**Figure 5:** Neutral-to-Happiness emotion style transfer results using parts-to-parts method and AffectNet dataset.

Figure 6: Happiness-to-Sadness emotion style transfer results using face-to-face method and AffectNet dataset.

5. Further research

The potential research direction could be the use of Adaptive Instance Normalization in these style transfer models [5]. In addition, recent advances in vision models and deep learning model architectures should be considered for testing. Vision transformers have gained popularity among image recognition models, which is also an interesting addition to GAN [20]. Difussion-based generative models could also be added to improve the quality of generated images and reduce computation requirements [21]. Using face-to-face and parts-to-parts methods requires the blending step to put the generated parts of the image back into the original source image. Currenlty, this is a separate method from the model architecture, but it could be made part of the model itself with a learnable blending procedure.

6. Conclusions

Novel pre-processing and post-processing methods were introduced. Pre-processing methods constituted segmentation for face-to-face and parts-to-parts methods, and post-processing constituted blending methods of generated parts with original image. All three face emotion style transfer methods, image-to-image face-to-face and part-to-parts, were tested, compared, and evaluated using two generative adversarial networks and the happiness-to-neutral emotion transfer task. The best quantitative results show that the part-to-part method achieves the highest accuracy of 97% and the highest FID score of 17.37, while the face-to-face method

produces more natural results of the transfer of emotion style, as seen in qualitative examples.

Acknowledgments

References

[1] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, ArXiv abs/1809.11096 (2019).

[2] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, CoRR abs/1511.06434 (2016).

[3] M. Mirza, S. Osindero, Conditional generative adversarial nets, ArXiv abs/1411.1784 (2014).

[4] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, ArXiv abs/1701.07875 (2017).

[5] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4396–4405.

[6] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2242–2251.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014.

[8] Y. Choi, M.-J. Choi, M. S. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 8789–8797.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, in: NIPS, 2017.

[10] N. A. N. M. Noor, N. M. Suaib, Facial expression transfer using generative adversarial network : A review, IOP Conference Series: Materials Science and Engineering 864 (2020).

[11] C.-H. Lee, Z. Liu, L. Wu, P. Luo, Maskgan: Towards diverse and interactive facial image manipulation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[12] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, ArXiv abs/1706.05587 (2017).

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

[14] E. Barsoum, C. Zhang, C. Canton-Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, Proceedings of the 18th ACM International Conference on Multimodal Interaction (2016).

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778.

[16] D. Ulyanov, A. Vedaldi, V. S. Lempitsky, Instance normalization: The missing ingredient for fast stylization, ArXiv abs/1607.08022 (2016).

[17] J. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, ArXiv abs/1607.06450 (2016).

[18] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 5967–5976.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: NIPS, 2017.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv abs/2010.11929 (2021).

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.