Ensuring identity preservation for motion translation in faces using VICE-GAN

Ronan Oostveen
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
r.oostveen@student.utwente.nl

ABSTRACT

In this paper, we propose further improving the identity-preserving features of the VICE-GAN network. The VICE-GAN network is a network that generates a video of a face expressing different emotions than the input video while preserving the same face. We suggest that using a more robust encoder could achieve these improvements. Another encoder could improve upon identity preservation because the encoder proposed in the original paper performs poorly on faces that did not appear in the training set. Therefore using an encoder that performs better on facial feature extraction on unseen faces, such as FaceNet[21] could also improve the accuracy of the VICE-GAN on unseen face models.

Keywords

Generative adversarial networks, VICE-GAN, Video-to-video translation, Encoder optimization, Facial feature extraction, Identity preservation, FaceNet

1. INTRODUCTION

Over the last five years, Generative adversarial networks (GANs) have become a popular topic in deep learning. Their versatile abilities, such as the unsupervised learning of feature representations and generating novel pictures, have been used for multiple use cases [6, 18, 29].

Within this development, there has also been some focus on video applications for GANs. For instance, Vondrick et al. (2016)[26] proposed a network that would be able to untangle a scene's foreground from the background by learning how to map video clips within the latent space. In response to this, Tulyakov et al. (2018)[25] argued that this mapping to the latent space would unnecessarily over-complicate the problem of video applications for GANs. Moreover, since the mapping within the latent space would also limit the generated video clips to have the same length, it would not accurately represent the real-world scenario. After that, they proposed a Motion and Content decomposed Generative Adversarial Network (MoCoGAN). The MoCoGAN structure, after given sufficient video training data, would be able to learn to untangle the motion from the content automatically. Through this structure, researchers developed new methods for video generation and prediction.[28, 26]

These improvements within the field of video generation allowed us to see that there are possibilities to use these techniques to generate entire datasets, which could eventually improve upon more traditional systems of machine learning systems such as emotion recognition[1, 15]. For this exact purpose, Jayagopal (2021) proposed the VICE-GAN [13] to generate videos that would represent different emotions based on a given input while still maintaining the same identity as the person in the input. This model, however, lacked some robustness and would perform poorly faces it has not seen in its training data.

Here we will try to address the inconsistency the VICE-GAN model has with a new face by implementing a different encoder part for the network. We think this will be an effective solution since we speculate that the problem in the original network originated from the original encoder's inability to extract the unique facial features out of the inputs properly. During testing, results for unseen data on the network would look like those from the network's already trained on data. This morphing could be the encoder's inability to extract the unique features properly and, therefore, only being able to output data that would result in good training results.

Additionally, research into specific optimizations, such as improving identity preservation, will help deepen understanding of the black box within these networks and how they operate.

1.1 Contributions

In this paper, we explore ways to improve the performance of the VICE-GAN network by enforcing better identity consistency. The model proposed in the original paper[13] had a generally good performance in facial emotion translation tasks. However, when testing on faces that were not in the training dataset, the network performance was significantly lower. As also explained earlier, the original network had an encoder that was somewhat rudimentary and only trained for the data for this network; this is the reason for exploring whether a generally more robust encoder could perform better regarding consistency. For this reason, we want to experiment specifically with the encoder and the loss function of the network according to these properties:

- Whether using a more robust encoder to generate feature vectors from input data will result in the network being better able to represent the input data better.
- If there are ways to optimize the triple-loss function

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
28^{th} *Twente Student Conference on IT* Febr. 2^{nd}, 2018, Enschede, The Netherlands.
Copyright 2018, University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

to enforce better identity preservation

When working on these properties, we also need to explore further other variables which go along with these properties. These variables are:

- *Scaling of the encoder input:* when adapting the network to a new encoder, the size of the input data for the encoder should be adapted. The original encoder could take in images of the size 64x64; other encoders such as FaceNet cannot encode images with a resolution lower than 100x100. The encoder should perform optimally
- *Modifying the loss composition:* the triple-loss function is optimized through the ADAM optimizer. For this optimizer, the composition of the triple-loss will affect which losses are of priority. There should exist a good balance between these parameters.
- *Scaling of the consistency loss:* the consistency loss is one of the three losses of the triple-loss function. The scaling of this loss affects how well the network will optimize for facial preservation.

1.2 Research Question

In the VICE-GAN, the encoder was not robust enough to produce similar results between unseen and seen training data. Here we want to analyze whether we would implement an encoder with higher accuracy in identifying faces and whether the VICE-GAN model would perform better on unseen training data. Further, we would also like to explore if there maybe are other measurements that could be used to more accurately measure the identity preserving loss in the model.

To better formulate these goals in concrete tasks, we will state them as the following research questions (RQ):

- RQ1: Will implementing the FaceNet[21] encoder improve the performance of the VICE-GAN? How did the performance specifically change?
- RQ2: Are there any other loss measurements that would improve the performance with the new encoder? How can we accurately measure how the GAN affects the encoder's input?

2. RELATED WORK

In sharp contrast to the lack of research in combining video generation and emotion generation within AI, research is abundant in the individual fields. The introduction mentions papers particularly concerning the VICE-GAN network, now we will examine a broader scope of research. Starting with research focused on GANs, followed by the specifics of video-to-video translation, and lastly, more profound literature on identity consistency. For this research, we used multiple means to gather the related literature; the current sources include domain Scopus, Google Scholar, and IEEE.

2.1 Generative Adversarial Networks

GANs were first proposed in 2014 by Goodfellow et al. [9] it is a unique network that consists of both a generator and a discriminator. The generator aims to generate authentic images, while the discriminator tries to discern real images from generated ones. The training of both these networks functions as a min-max optimization where one network tries to minimize while the other tries to maximize; this is represented by alternating gradient upgrades between these networks. In their paper, Goodfellow did show that these networks can still converge to an optimum, given enough capacity and iterations.

(a) Examples of generated video frames using [26].

(b) Examples of generated video frames using [18].

Figure 1: Video's generated by dynamic transfer [4] in (a) from using MoCoGAN [25] input frames (b) using Realistic Dynamic Textures GAN [20] for input frames

This field has already flourished into multiple fields, which include representation learning[18, 29] and recognition networks [1, 8]. Further specifically related to video processing GANs were used for different types of video decomposition [25, 26]. These GANs relate to our use of GANs since we will decompose the emotion and the face of the person in the video for the problem of emotion synthesis. This new MoCoGAN model opened doors to video-to-video synthesis[28], image-to-image translation[11] and video prediction[17, 3].

2.2 Video-to-video

For video-to-video translations there has been previous research focusing a variety of different video manipulations on video super-resolution [23], video style-transfer [7], and video snipping [5]. For video generation, video-to-video synthesis [27] has been proposed as a state-of-the-art technique for generating high-resolution, photo-realistic video. Moreover, video synthesis uses a unique method where the model focuses on matching the conditional video distributions instead.

Expanding on the MoCoGAN framework, a method has been created for transferring the arbitrary dynamics from a video input sequence onto a target image (See Figure 1)[4], thus creating a new video with the dynamics from the video on the target image. The VICE-GAN model[13] is comparable since it aims to translate the dynamics of an emotion on a target video. Newer papers also proposed similar emotion editing using latent space manipulation[24]. Additionally, there is a review of current state-of-the-art GAN models used for human emotion synthesis and how they perform comparatively[10].

2.3 Identity consistency

GANs in recent years are becoming more notorious for getting good results in generating high-quality images of faces (such as generated by StyleGAN [14]). The safety implications for GANs generating faces and the novelty of these faces have been questioned. However, research

on this has shown that GANs trained on face images are still capable of generating new identities and can also be used for data anonymization [19]. This generation of novel items is essential to mention since the identity consistency issue faced in the previous VICE-GAN model since the model would perform poorer on unseen faces, considering whether this model could generate these new faces.
Following this, other research on identity consistency with GANs has also been done. In a specific self-supervised video GAN [12] the discriminator was tasked with learning two things, the representation of the appearance (the identity) and the temporal structure (the motion), and this model showed improvements over the state-of-the-art models.

3. METHODOLOGIES

This section will describe the methodologies proposed for this research. These methods include the modifications we will apply to the original VICE-GAN system described in section 3.1. After this, these methodologies will be used to run the experiments in section 4 and generate the final research results.

3.1 VICE-GAN

Th VICE-GAN network is the framework of this entire project, the structure of the network can be seen in Figure 2. The network exists of 5 different sub-networks, which include: an encoder for the input, a Recurrent Neural Network (RNN) for creating the motion vector, the generator network for the fake video; and then two different discriminators for the output, one for the image and one for the video. The function of each of these sub-networks will now be further elaborated.

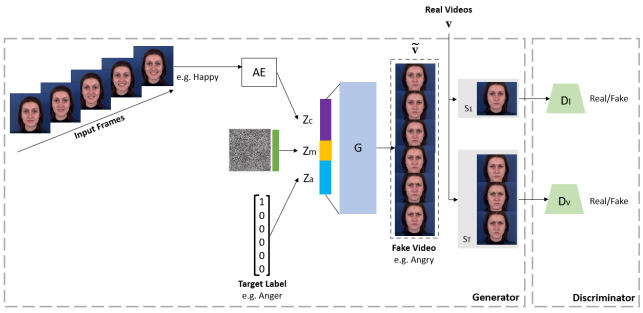

Figure 2: The structure of the VICE-GAN network [13]

3.1.1 The encoder

Within the VICE-GAN structure, the encoder creates a feature vector from a certain input video, represented by the formula below.

$$Z_c = I_e(X)$$

here I_e is the encoder, which takes in X the vector representation of the input video; this generates the vector Z_c, which contains all the previous frames and their feature vector.

3.1.2 RNN for motion creation

In this network, the motion must be changed between the frames, but the new motion still needs to be related to the previous frame. For this purpose, an RNN is used since these networks specialize in generating sequences with correlation.

$$Z_M = R_m(\epsilon)$$

here the R_m represents the RNN which takes as input a vector ϵ which is sampled from a normal distribution; this generates the eventual motion vector Z_M.

3.1.3 The generator

As seen in Figure 2 the generator is the central part of this GAN. This generator receives input from the previously discussed feature vector Z_c, motion vector Z_M, and a new one-hot vector Z_A which represent the target emotion. From these input vectors it will attempt to create a video sequence that should adequately represent the feature vector's identity and the motion of the emotion. The generator is then trained using a triple loss function comprised of the content-consistency loss, adversarial loss, and category loss. This training process will be elaborated further in section 3.1.5.

3.1.4 The discriminators

Two types of discriminators are utilized for this network; these are an image discriminator and a video discriminator.

The image discriminator is based on a Convolutional Neural Network (CNN). Its goal is to discriminate whether a frame from a video is real or generated. The feedback from this image discriminator has then added the loss of the generator to improve its performance.

The video discriminators utilize spatial-temporal modeling; this models how the video frames vary over space and time, then it aims also to discriminate whether this spatial-temporal frame is real or generated. This creates feedback for the motion of the generated video and is used to train the R_m RNN network. This discriminator also trains to learn the different categories of emotions, which it then aims to label correctly.

3.1.5 The training

For the training of the generator, the VICE-GAN uses a triple-loss function for the generator. This triple loss consists of three separate losses and is used to calculate the combined loss for the generator. These separate losses include the $L_{discriminator}$, $L_{consistency}$, and $L_{category}$. These three losses are combined and then used for the backward propagation with the ADAM optimizer to update the weights properly.

Now to further elaborate on each one of these losses. The $L_{discriminator}$ is the loss that optimizes the discriminators. This loss motivates the generator to generate realistic data, which will trick the discriminator into thinking it is real data. While the discriminator is corrected on how to differentiate real and fake data properly. This process is the same min-max optimization discussed in section 2.1.

The $L_{consistency}$ loss compares the similarity between the actual data and the network's output. This loss ensures consistency is encouraged within GAN. There are different ways to calculate how different these vectors are. For this the original paper uses Mean Squared Error, but in section 3.3 we will also propose Cosine distance.

Lastly, the $L_{category}$ loss is used to condition the generator into creating a video that will be correctly categorized. Here the video discriminator will aim to learn how to discern the categories from the training data. At the same time, the generator will generate the videos with the categories to be recognized by the discriminator.

This training process is iterated on for roughly 50.000 to 60.000 generations to keep optimizing the generator, discriminators, and RRN continuously.

3.2 Encoder

For the newly proposed method, we propose using the FaceNet[21] encoder for this network. The original encoder in the network was an auto-encoder that would take in the video data and then generate a feature vector of 50 features. The accuracy of this encoder in the VICE-GAN[13] paper was 96.1%, but this was achieved on a limited amount of specialized data. We think that using a more generalized encoder which has proven to have an accuracy of 99.63% on the huge diverse Labeled Faces in the Wild dataset[16].

For our networks, we want to replace the old encoder with the new encoder; this causes two problems. The first is that FaceNet can only accept images above a resolution of 100x100, and the original training images are 64x64 in resolution. Furthermore, the second issue is that the generator only takes in 50 feature vector items from the previous encoder.

To solve the problem with the generator input, we will first scale up the intake of the generator to 512 feature vector items, after which we will have to train a new network. For the problem with the encoder input, we have a couple of choices: we could either scale up the data to a resolution of 100x100 to keep it as close to 64x64 as possible or up to 160x160 since the FaceNet encoder was initially trained on data with this resolution. After testing both resolutions on a data set of 50 images, we had an accuracy of 88.6% on the 64x64 resolution, while we had an accuracy of 98.9% on the same images scaled to 160x160. This was the motivation to scale up the images for the encoder to 160x160; for this bi-linear scaling was used.

3.3 Other loss metrics

For the consistency loss criterion, the original was Mean Squared Distance, and this makes sense since it will aim to make the vectors as similar as possible. For this we would like to aim for similarity, and since we do not specifically mind if the data is not close in the Euclidean space, we want the data to be at a similar angle to each other. This is because facial recognition systems such as FaceNet will still recognize the data as the same. Therefore our choice is Cosine similarity because this is already very similar to Mean Squared Error, and limitations on time would currently restrict us from trying ArcFace.

Because we are using a different encoder and loss criterion, the size of the loss might differ. This is an issue since the loss is composed of multiple loss functions, and when the consistency loss becomes smaller it loses priority. This is why we scale the consistency losses to get the variable as similar as it was in the original network; this also affects the loss composition of the network.

4. EXPERIMENTS

In this section, the experimental setup will be discussed along with the dataset and hyper parameters related to this setup; this will be followed by an overview of the experiments performed and how the experiments were evaluated. Further technical details are as follows:

- **dataset:** The MUG Facial Expression Database was used for this experiment [2]. We used data from this dataset consisting of 50 individuals expressing five different emotions. In the pre-processing, all videos with less than 64 frames were removed, and the frame resolution was scaled to 64x64 pixels to comply with the MoCoGAN framework standards. After the pre-processing, the total video count was 603. Of these 603 videos 548 were used for regular training and testing, while 55 were kept separate for testing on unseen individuals. The data was given an 80-20 train-test split during the training.
- **parameters:** The batch-size for the experiment was 32 for both videos and images, the ADAM optimizer was used with a learning rate of α=0.00002, and for the momentum the β_1=0.5 and β_2=0.999

The training was done using the CTIT cluster of University Twente utilizing either a single TITAN X 12GB GPU or Quadro RTX6000 24GB GPU.

4.1 GAN modification

Four different GAN models have been trained to observe the effects of different hyper-parameters while changing the model. The models differ in: the encoder used, consistency loss criterion, the scale of the consistency loss, and the composition of the generator loss. These model types are specified in Table 1

Table 1: Overview of GAN variants used

Model	*Encoder*	*Cons. loss*	*Scale*	*Loss composition (cons., cat., discr.)*
1	VICE-GAN	MSE	1	(63%, 34%, 3%)
2	FaceNet	MSE	10000	(78%, 20%, 2%)
3	FaceNet	MSE	15000	(78%, 19%, 3%)
4	FaceNet	COS	2000	(57%, 39%, 4%)

The different encoders and consistency losses used for these models affected the loss composition. This effect was partly mitigated by scaling the consistency loss.

Model 1 is a reference model to the original VICE-GAN model. This model uses the original encoder proposed[13], and the consistency loss for this model was at the start of training 44. Model 2 uses the FaceNet encoder, with a loss consisting of the Mean Squared Error scaled by 10000. If this variable did not scale the consistency loss, it would have become diminutive. Scaling it by 10000 prevents this resulting in the eventual consistency loss for model 2 being 31. Model 3 is identical to Model 2 except for the scaling variable being 15000, resulting in a consistency loss of 54. The reason for including both these models is to observe the effect of scaling the consistency loss differently. Model 4 used the cosine similarity as loss criteria. For model 4, the consistency loss was scaled up by 2000, which resulted in it 13 at the start of training (a model with a consistency loss closer to Model 1 would perform differently).

4.2 Evaluation

In every experiment, all variables not discussed in section 4.1 are kept identical to the original baseline. All data presented was gathered from testing results of the training networks unless specified otherwise.

4.2.1 Data used

Since it is important to observe both the performance of the models on faces it has seen and the performance on faces unseen, we have devised two data groups: the seen and unseen data group. The seen group consists of data

containing the same individuals used in training the network. While the unseen group consists of data from two individuals on whom the network had not trained before.

From testing the GAN variations on these data groups, two batches were gathered consisting of 32 videos. One of these batches is the video data of the expected result, while the other batch is the video data developed by the generator.

4.2.2 *Evaluation methods*

To eventually answer the research questions we will first answer the following sub-questions using results generated from the different networks:

- How similar are the facial identities?
- Is the same face recognized?
- Are the same emotions recognized?

These sub-questions are used to bridge the gap between the research questions and the results. To answer the sub-questions the lightweight face recognition and facial attribute recognition framework DeepFace[22] was used.

Using this framework the first question will be answered by observing the distributions of the cosine distance results of the varying networks. This will give us insight since a lower cosine distance would mean that the network is better at creating similar facial identities.

For answering the second questions we will see if these cosine distances are below certain thresholds. If these cosine distances are below this threshold we can conclude that the faces are actually recognized as the same faces.

And lastly, the third question will be answered by the emotion analysis function of the DeepFace framework. This function will observe for every video which emotion is most likely expressed. After which we will compare if this emotion is recognized as the same for both data groups.

4.2.3 *DeepFace models*

DeepFace has eight models from which we can choose which one we would like to use for calculating the facial similarity. From these models we would like to use the one which is most precise with our specific data, for this task we will use a ROC-curve. A ROC-curve is a graphical plot that shows the precision of a binary classifier using different thresholds. In figure 3 you can see the complete ROC-curve for our data.

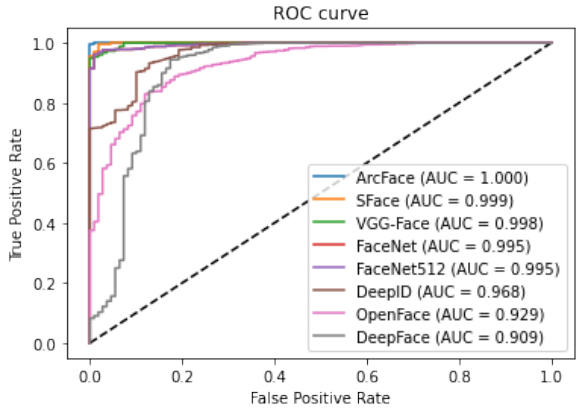

Figure 3: The constructed ROC-curve

In the legend of figure 3 you can see the Area Under the Curve (AUC), a higher AUC is linked to a better precision. In the ROC curve the best case is to have line the closest to the top left as possible which is represented by a higher AUC value. In this graph you can see that ArcFace is thereby the best performing model with a AUC of 1.000 (rounded up from 0.99995). Therefore we will utilize ArcFace for the DeepFace framework to calculate facial similarities for the results.

From the ROC-curve a more fine-tuned threshold has also been calculated for classifying our data.

$$Threshold = 0.36993933946974566$$

This new threshold will have a better accuracy correctly identifying similar faces than the original threshold, which is 0.68 for ArcFace. In our results we will use both thresholds to observe the difference in the quality of the data generated by the networks.

5. RESULTS AND DISCUSSION

In this section, the results from the experiments will be displayed and discussed. Then other limitations and potential areas for further research will be disclosed. Just like the original VICE-GAN model some artifact were still observed in the generated videos on seen data. In figure 4 you can see a visualisation of the expected result versus the generated video of each model. You see that model 4 had the biggest deviation while model 1, 2, and 3 all still made recognisable videos. The emotion expression also seems more accurate in model 1 and 3.

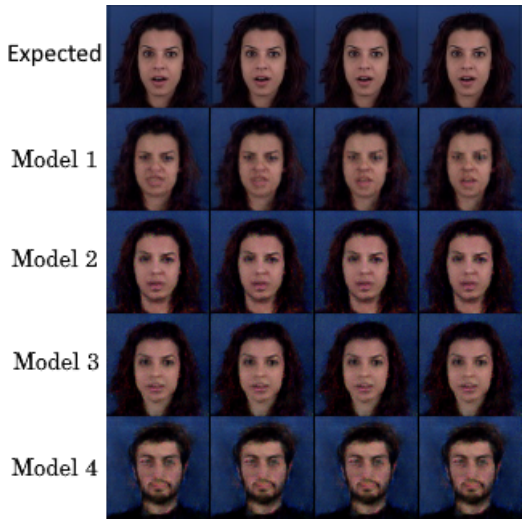

Figure 4: Representation of the same seen video by different models

Figure 5 is a similar representation but for unseen data. Here you can see that all images do not look recognisable. However model 3 seemed to do better with maintaining facial features such as the nose, mouth, and beard. Model 4 seemed to have a lot of disfigurements here which could be an exaggeration of the models in ability to also present normal data properly as well. Further model 1 and 2 seem to generate the correct emotion expression while still having a very different facial features.

Follow this each evaluation sub-question from section 4.2.2 will be assessed.

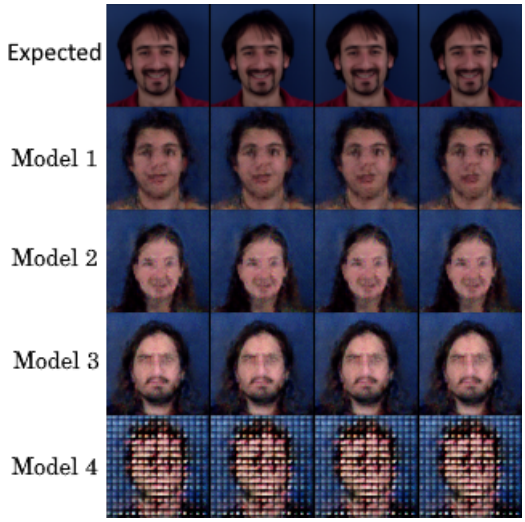

Figure 5: Representation of the same unseen video by different models

5.1 How similar are the facial identities?

The facial similarity was evaluated using DeepFace combined with the ArcFace model. From this we generated the distribution of the cosine distances for both unseen and seen individuals in Figure 6. For every model you can see the distribution of cosine distances, with seen data on the left side and unseen data on the right side, which is represented by how wide the graph is at that point. The combined average cosine distance per model is presented by the white dot in the middle.

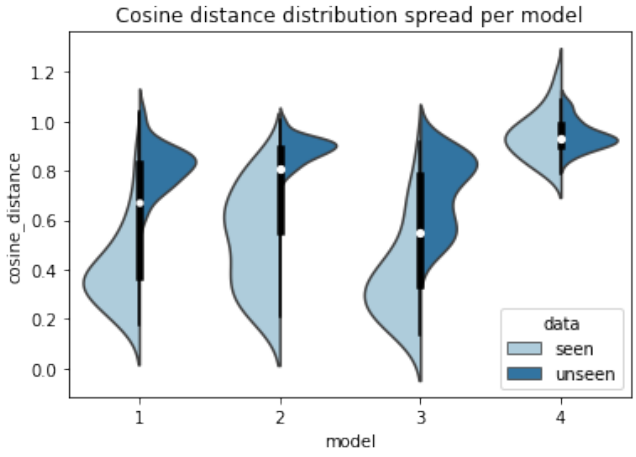

Figure 6: Constructed ROC-curve zoomed in top-left

In this graph a lower combined average means represents a better performing network. Therefore we can see that model 3 was the best performing with models 1, 2, and 4 all perform respectfully more worse. Looking at the distributions of the data we can also see that model 3 on seen had a slightly lower distribution than model 1 on seen data. This means that model 3 generally generated higher similarity videos than model 1. Between model 1 and model 3 we can observe an even bigger difference on unseen data. Here you can see that the distribution for model 3 goes much lower in cosine distance than model 1. This means that model 3 is better at identity preservation for unseen data, while model 1 is equally bad for the identity preservation of all unseen data.

The result of model 2 were significantly worse than model 3, which is surprising since the networks only differ in scaling. This insinuates that scaling can be a big influence in the identity preservation of the models. Model 2 for the seen data has a very wide distribution spread which means that the model had a very varying performance on this data. For unseen data this model has a very narrow distribution of equally high cosine distances this means that the model is equally bad at the identity preservation for all unseen data. This model performs worse than the original VICE-GAN model despite using the FaceNet encoder.

Model 4 is the worst performing model, it is difficult to state whether this is caused by the cosine similarity for the consistency loss or because the scale and loss composition. Early stages of the research showed that having a consistency loss function which is not scaled would result videos virtually no identity consistency. Looking at the composition of the loss function for network 4 in Table 1 it also shows that the consistency loss is off a lower priority than in the other networks. Based on the distribution in these results it seems like cosine similarity might be a good criterion, but it should not be seen as conclusive for the difference in loss composition and scale.

As a conclusion for this question, it seems that using FaceNet as an encoder can result in more similar facial identity representation for both seen and unseen data. However this is very depended on the scale of the consistency loss. Using a cosine similarity loss generates very poor results, but these results might also be cause by the scale or loss composition.

5.2 Is the same face recognized?

By checking whether the average cosine distance is lower than the pre-determined thresholds we can observe if the faces can be identified as the same person. The accuracy is then calculated from the amount of videos recognized divided by the amount of total videos. For the standard threshold for ArcFace the accuracy is presented in figure 7.

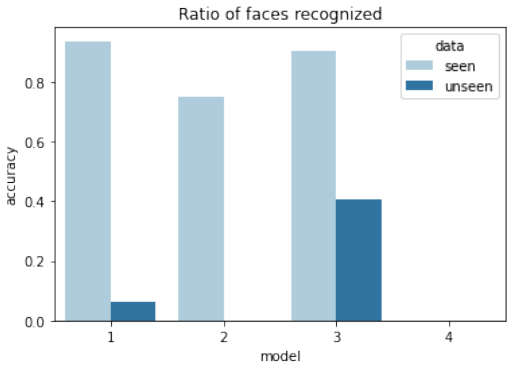

Figure 7: The ratio of faces recognized in the video with a threshold of 0.68

Here you can see that model 1 has a very high accuracy on seen data, but a very low accuracy on unseen data. Comparatively model 3 has a slightly lower accuracy on the seen data, but a much higher accuracy on unseen data. This means that for seen data model 3 was slightly worse than model 1 in generating similar results. But for unseen data model 3 was significantly better than model 1

at generating similar results. Model 2 and model 4 both perform significantly worse, which was expected based on the results of section 5.1.

It is important to also compare this accuracy to the accuracy of the fine-tuned threshold calculated earlier. Using this threshold the accuracy of the models is shown in figure 8.

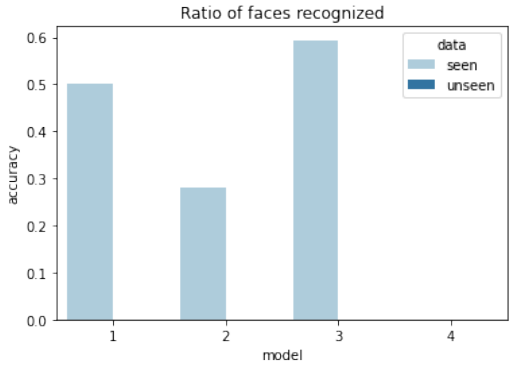

Figure 8: The ratio of faces recognized in the video with a threshold of 0.37

In this graph, model 3 now has a higher accuracy than model 1 on seen images. This is because model 3 generates more highly similar videos than model 1. With these results every model has a 0% accuracy on the unseen data, which means that the videos generates for unseen data are still not similar enough for the fine-tuned threshold.

As conclusion, models 1 and 3 also have the highest amount of faces recognized below the thresholds. Model 1 performs better than model 3 generating seen data below the default threshold of 0.68. While model 3 performs better than model 1 on seen data with a more fine-tuned threshold. Model 3 also greatly out performs model 1 on unseen data on the default thresholds. And both models 2 and 4 perform significantly worse than the model 1 and 3. Even though model 2 manages to generate some recognisable faces.

5.2.1 *Is the same emotion recognized?*

Using DeepFace it is also possible to observe the emotion in a video. These results can be observed in Image 9.

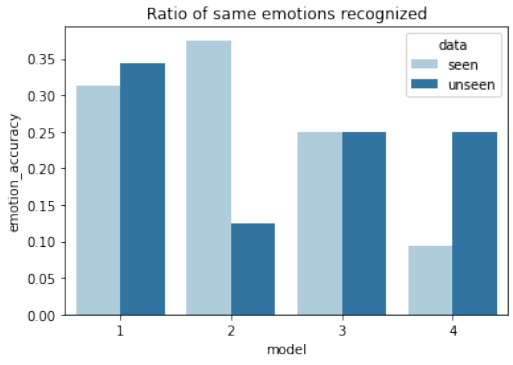

Figure 9: The ratio of videos with the same emotion expressed.

Here model 2 was the best at expressing the same emotions on seen data. Model 1 performs both accurate on seen and unseen data, which indicates a general consistency with emotions representation. Model 3 has identical results on the emotion expression which indicates also consistency in emotion representation but at a lower accuracy than model 1. Surprisingly model 4 performs better at unseen data than seen data, the reason for this could be that model 4 does the incorrect learning for the emotion representation. But it could also be a display of limitations of this metric on a 16 frame video.

5.3 Discussions and Outlook

5.3.1 *Research Question 1*

To address research question 1, we can state that using the FaceNet encoder can improve the performance of the VICE-GAN. However, other variables, such as the scaling and loss composition, can play a big part in what extent. In our results, model 3 performed better than model 1 on unseen data. While models 2 and 4, using FaceNet, performed worse than model 1.

The increase in performance from model 3 to model 1 was for unseen data 32% with a low threshold. For seen data, model 3 performed 10% than model 1 with a fine-tuned threshold, but for the low threshold, model 1 still performed a little better.

5.3.2 *Research Question 2*

In these experiments, whether other loss measurements could improve the performance with FaceNet is not proven. Using cosine distance model 4 with FaceNet performed the worst out of all previous models, but we cannot conclude if this was for the loss measurement or the other variables involved. Using other loss measurements such as ArcFace, the model could perform better, but this is not proven in this experiment.

To better measure the GANs effect on the encoder input, we can look at the results from the ROC-curve analysis. Here we can see that ArcFace is the best method for differentiating fake faces from real ones; using this metric to measure the GANs affect, the model could potentially train better.

5.3.3 *Shortcomings*

The biggest issue for these experiments is the data limitation, for data with unseen individuals, we only had two different individuals. This therefore limits the credibility of those experiments since we do not know if models perform better or worse because the individuals in the data look similar to the individuals the network trained on. Using more individuals for the training data could also increase the networks robustness on different faces.

Another major weak points within these experiments is the image resolution, using 64x64 images decreases the accuracy of the facial identifying algorithms. When testing different resolutions scaling up the images for the FaceNet encoder it performed with a 10% higher accuracy on images that were 160x160 compared to 100x100 images. This could also insinuate that the network could perform better on 160x160 resolution images, but this is inconclusive and requires further research.

Lastly, the MoCoGAN model only trains on 16 frames, and in our experiments these frames are selected based on every first frame. So most of the time only 16 very early frames are selected, also with the MUG data not every video starts expressing the emotion in these 16 frames. Not being able to correctly identify these emotions could limit the emotion generating capabilities of the GAN.

5.3.4 Future work

This paper has showed that using a different encoder could provide better results on the original VICE-GAN structure. However due to fluctuations in the scaling and loss composition it is not easy to discern to what extend. Future research in this area could focus more elaborately on different scaling and loss compositions, which could help identify the optimal values for these variables. Combining this with ArcFace to calculate the consistency loss could provide even better results.

6. CONCLUSION

This paper has presented and experimented with three new adaptations of the VICE-GAN network, using FaceNet as the encoder. The results showed that using this encoder can improve the network's performance using the correct variables.

This was achieved by creating different models using the VICE-GAN model and FaceNet encoder but changing the variables for each model. The models using a consistency loss calculated from the Mean Squared Error difference performed better than the model which used the Cosine similarity. Using a different scale for the Cosine model could potentially improve the performance. Within these Mean Squared Error models, the higher scaled model for the consistency loss had the best accuracy and outperformed the original VICE-GAN model. There was also the unexpected benefit that the higher scaled Mean Squared FaceNet model seemed to produce more high-quality videos compared to all the other models.

A discussion of potential future work and improvements is also discussed, where the results for improving on the data and observing the full effects of scaling the consistency loss could be investigated. While the performance in this research is better, more considerable improvements for these models seem plausible using methods such as ArcFace for either the encoding or the loss consistency.

7. REFERENCES

[1] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(02):52–58, 2021.

[2] N. Aifanti, C. Papachristou, and A. Delopoulos. The mug facial expression database. pages 1 – 4, 05 2010.

[3] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

[4] W. J. Baddar, G. Gu, S. Lee, and Y. M. Ro. Dynamics transfer GAN: generating video by transferring arbitrary temporal dynamics from a source video to a single target image. *CoRR*, abs/1712.03534, 2017.

[5] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: Robust video object cutout using localized classifiers. *ACM Trans. Graph.*, 28(3), jul 2009.

[6] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Cvae-gan: Fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[7] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua. Coherent online video style transfer. *CoRR*, abs/1703.09211, 2017.

[8] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[10] N. Hajarolasvadi, H. Demirel, M. Arjona Ramírez, and W. Beccaro. Generative adversarial networks in human emotion synthesis: A review. *IEEE Access*, 8, 12 2020.

[11] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[12] S. Hyun, J. Kim, and J.-P. Heo. Self-supervised video gans: Learning for appearance consistency and motion coherency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10826–10835, June 2021.

[13] T. N. Jayagopal. Vice-gan: Video identity-consistent emotion generative adversarial network, August 2021.

[14] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] T. L. B. Khanh, S.-H. Kim, G. Lee, H.-J. Yang, and E.-T. Baek. Korean video dataset for emotion recognition in the wild. *Multimedia Tools and Applications*, 80(6):9479–9492, 2021.

[16] E. G. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. 2016.

[17] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

[18] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun. Marta gans: Unsupervised representation learning for remote sensing image classification. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2092–2096, 2017.

[19] R. T. Marriott, S. Madiouni, S. Romdhani, S. Gentric, and L. Chen. An assessment of gans for identity-related applications. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020.

[20] K. Olszewski, Z. Li, C. Yang, Y. Zhou, R. Yu, Z. Huang, S. Xiang, S. Saito, P. Kohli, and H. Li. Realistic dynamic facial textures from a single image using gans. pages 5439–5448, 10 2017.

[21] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[22] S. I. Serengil and A. Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.

[23] E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):531–545, 2005.

[24] V. Strizhkova, Y. Wang, D. Anghelone, D. Yang, A. Dantcheva, and F. Brémond. Emotion editing in

head reenactment videos using latent space manipulation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[25] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[26] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

[27] T. Wang, M. Liu, J. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *CoRR*, abs/1808.06601, 2018.

[28] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[29] W. Yang, C. Hui, Z. Chen, J.-H. Xue, and Q. Liao. Fv-gan: Finger vein representation using generative adversarial networks. *IEEE Transactions on Information Forensics and Security*, 14(9):2512–2524, 2019.