VICE-GAN: Video Identity-Consistent Emotion Generative Adversarial Network

Master Thesis

Tarun Narain Jayagopal

University of Twente

t.n.jayagopal@student.utwente.nl

ABSTRACT

We propose the Video Identity-Consistent Emotion Generative Adversarial Network (VICE-GAN) model for video generation. The proposed model is able to generate realistic videos of six emotional expressions while allowing the identity of the individual to be preserved. This was achieved by introducing (i) a pre-trained autoencoder which produces a compressed representation of the individual present in an input video and therefore preserves the content of the video and (ii) a content consistency loss to further enforce identity consistency by extracting and comparing the content representations between the generated and real frames of a video. In addition, we experimented with three variables in order to determine their impact on model performance. Eight model variants were evaluated based on visual quality, emotion generation and identity consistency. Overall, models which were exposed to the test subjects beforehand for a limited number of emotions produced video sequences of higher visual quality and identity consistency when compared to models in which the test subjects were removed from the training data entirely. Using the content representation of the first frame for all subsequent frames in contrast to using a unique representation for each frame appears to benefit identity consistency only. There is also evidence to suggest that freezing autoencoder weights during GAN training results in improvements for visual quality and emotion generation.

KEYWORDS

Generative adversarial networks, Video-to-video translation, Emotion generation

1 INTRODUCTION

Deep generative models, such as generative adversarial networks (GANs), have received an increasing amount of attention for their ability to generate realistic images and videos for various vision applications such as rendering, synthesis, recognition and augmentation. While there has been significant progress in the application of GANs for image tasks such as generation [15], editing [8], translation [23] [13] and super resolution [36], video applications have received relatively less attention. Furthermore, GANs which have been developed for video applications [9] [25] have tended to focus on scenes and long-distance human activities such as actions and poses, illustrating the need to focus on close-up human faces as well.

A common approach that has been employed in recent work [30] [18] [29] [28] towards video generation tasks is to decompose a video into its objects (content) and the actions they perform (motion), after which the latent variables obtained from each sub-space are combined to produce videos. Through this approach, promising efforts have been made primarily towards video-prediction, -generation and video–translation using both paired and unpaired data. Paired data refers to a one-to-one relationship between training examples in a dataset. For example, a dataset containing input examples from domain X would require the same examples with the desired modifications as the expected output in domain Y. However, these studies indicate that there are still several challenges that need to be overcome, namely (a) content consistency throughout the video, (b) generating (uncertain) motion, as well as modeling of spatio-temporal consistency.

Given the myriad of domains in which video generation can be useful, we select emotion generation as a use case to develop a deep video generative model that can generate accurate videos of different emotions being expressed by specific individuals in an unpaired manner. Despite emotion recognition (ER) systems being successful when classifying emotions, recent work [20] [1] indicates that deep-learning emotion-recognition systems can be enhanced and improved further. A major technical challenge that these systems face is the lack of appropriate emotion databases. This could be solved by manually creating a dataset for the task at hand but it is a very expensive and time-consuming approach. GANs in general are geared towards tackling data augumentation problems so it would be interesting to observe how the model would handle the task of emotion generation. It is important to note that while the generative model is being developed specific to the emotion generation, it can be adapted not only to tasks involving human faces but also other objects and domains.

1.1 Our Contributions

In this paper, we propose the VICE-GAN for video translation and focus on the task of transferring emotion-specific facial expressions within the same individual. The VICE-GAN has been adapted to leverage the benefits of the two-step method i.e. the decomposition of content and motion while addressing the two challenges listed above by incorporating the following properties:

(i) Input frames are encoded to form compressed representations of the human face present in the video, which is used to inform the content space of video generation.
(ii) The identity of the selected face is enforced using a content consistency loss to ensure that the same individual is generated when generating a different emotion-specific facial expression.

Additionally, we explore the influence of three experimental variables on model performance based on preliminary experiments:

(i) *Composition of Training Data*
One of the aims of this work is to preserve the identity of the individual present in the video when generating to different emotions. A common approach to evaluating such a model is to select a subset of subjects for testing, and these subjects are not seen by the GAN model during training (**Unseen condition**). However, it is plausible that the model may need to have been trained on the test subjects' faces in order to achieve this goal. Therefore, an alternative approach would be to train the model on the test subjects. Specifically, it can be explored whether exposing the model to the test subjects but only for certain emotions allows it to preserve identity while still being capable of generating to different emotions (**Seen condition**).
(ii) *Content Encoding Method*
Content encoding method refers to the way in which content vectors are used to represent the individual face present in each video before being fed into the generator. One way to achieve this is to apply the autoencoder to each individual frame of a video, thus producing a unique vector for each frame (**All Frames**). Alternatively, the autoencoder can produce a unique vector for the first frame as a reference and this vector is then repeated across all frames (**Single Frame**).
(iii) *Fine-Tuning Autoencoder Weights*
Assuming that the autoencoder is pre-trained, it was hypothesized that autoencoder can be further fine-tuned by unfreezing weights during the training of GAN. Alternatively, the autoencoder weights are frozen and the weights are not updated while training the GAN model.

1.2 Research Questions

Through the development of the VICE-GAN, this paper focuses on the following overarching research question *"Given an input video of an individual expressing a given emotion, to what extent and quality can we generate videos of the same individual expressing different emotions?"* The research question can be formulated in the following sub-questions:

(1) What are the current state-of-the-art models that are applicable for image-to-image and video-to-video generation/translation tasks?
(2) How can we ensure that during video-to-video translation, a network is able to preserve the content of the input video (e.g. the identity/face of an individual expressing a certain emotion)?
   (a) To model the content space, which autoencoder architecture produces the most accurate, high-quality compressed representations of human faces?
   (b) Can identity consistency be additionally enforced through the use of a content consistency loss?
(3) What are the influences of the following *experimental variables* on model performance in terms of (i) visual quality, (ii) accuracy of generated emotions and (iii) identity consistency?
   (a) Allowing the GAN model to train on samples depicting the test subjects for a limited number of emotions *(Seen)*, or removing the test subjects from the training data completely *(Unseen)*
   (b) Content encoding method, where the autoencoder either produces a unique content vector for each input frame, or repeats the content vector obtained for the first frame for all subsequent frames
   (c) Unfreezing or freezing the autoencoder weights while training the GAN model
(4) Which of the model configurations perform best in terms of (a) visual quality and (b) identity consistency and (c) accuracy of the generated emotions?

The remainder of the paper is structured as follows. In **Section 2**, we provide the background on the existing methods that are used for image-to-image and video-to-video generation/translation related tasks. **Section 3** presents the methodological framework that is being used to address the problem. In **Section 4**, we propose the VICE-GAN model as a solution to address the challenges encountered by existing methods. In **Section 5**, we describe the dataset used as well as the experiments and evaluation metrics conducted in the study. The results of the research are explained in **Section 6** followed by a discussion in **Section 7** which addresses the findings and answers the research questions introduce earlier in this section. We further discuss the shortcomings of the model and suggest some improvements. Finally, the paper is concluded in **Section 8**.

2 RELATED WORK

2.1 Image-to-Image translation

Unpaired image-to-image translation approaches have been proposed to address the lack of paired data. Paired data refers to a one-to-one relationship between training examples in a dataset. For example, a dataset containing input examples from domain X would require the same examples with the desired modifications as the expected output in domain Y. CycleGAN [38] uses two generative models and cycle consistency loss to perform regularisation.

Assuming two domains A and B, generator A performs translation from domain B to A while generator B performs translation from domain A to B. The two corresponding discriminator models determine whether the generated data is real or fake, and update the generator models accordingly. Cycle consistency is

centred on the premise that images generated by a given generator can be fed into the other generator to reconstruct the original image.

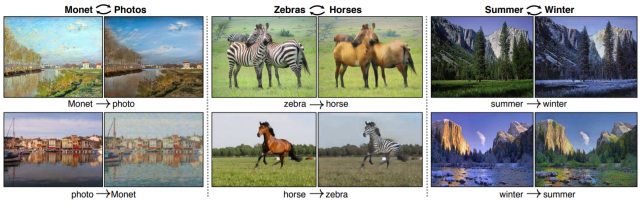

Figure 1: Examples of translation tasks performed by CycleGAN [38]

From left to right: between (a) Monet and natural scenes, (b) horses and zebras, and (c) summer and winter scenery

The corresponding cyclic consistency loss is used to push the generators to be consistent with each other. CycleGAN has therefore successfully been used to perform a variety of translations which can be observed in Figure 1. These findings led to the growth of CycleGAN-inspired models for unpaired translation tasks which will be discussed below.

To overcome the challenge associated with unpaired multi-domain transfer, StarGAN [11] proposes a single generator-discriminator network that can learn multiple mappings simultaneously, resulting in both efficiency and flexibility. This architecture and its subsequent version [12] have been successfully applied to a range of image-to-image translation tasks involving human faces such as facial attribute transfer, facial expression synthesis tasks and gender swapping Figure 2. Other methods such as RelGAN [35], AttGAN [16] and attribute-guided conditional cycleGAN [24] have also been found to be remarkably flexible for a variety of image-based translations such as change in gender, hair colour and facial emotion whilst addressing multi-domain image translation.

Figure 2: Examples of emotion-specific facial expressions generated using StarGAN [11]

2.2 Video generation approaches

A natural extension of image generation or image-to-image translation tasks is addressing videos, although this has proven to be a challenging topic. An intuitive approach to video translation and generation is to apply image-to-image translation methods on each frame. However, such methods result in a lack of continuity between frames, resulting in unrealistic motion and temporal artifacts such as distortions and flickering [5] [25]. In other words, the goal is to prevent perceptual mode collapse by considering both spatial and temporal constraints. Unsupervised video representation learning tasks can be classified as prediction, generation or translation.

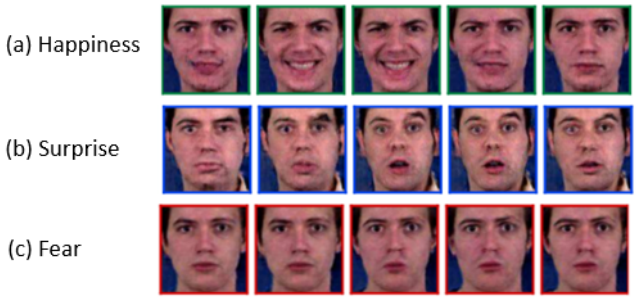

Figure 3: Examples of facial expressions generated using MoCoGAN [29]

Video prediction refers to the inference of subsequent frames of a video given a single or several input frames acting as context. A dual motion GAN was found to be successful for predicting future frames of various natural scenes [22], which enforced generated frames to be consistent with real input on the basis of pixel-wise flows in the video, and additionally addressed motion uncertainty using a probabilistic motion encoder.

Video generation generally requires that the model is able to generate desired videos without providing any input. A two-step method is often used here, which assumes that videos comprise of objects (content) performing actions (motion), and by combining latent variables from each sub-space, a sequence of spatio-temporal consistent frames can be generated. For example, temporal generative adversarial nets (TGAN) [26] use a temporal generator to produce a set of latent variables, and these which are fed into an image generator to produce a video where the number of variables is equal to the number of frames. Cai and colleagues [7] proposed a GAN model that is able to flexibly switch between video prediction and generation. This was done by using a similar two-step framework in which human-skeleton pose sequences are first generated, producing the *motion* and then translated into images to produce human action videos.

Perhaps the most representative of the two-step approach for video generation is the motion and content decomposed generative adversarial network (MoCoGAN) [29]. The MoCoGAN architecture has been used to generate short, motion-consistent videos depicting different human actions and shape motions and facial expressions (Figure 3). Additionally, it is capable of learning how to generate videos belonging to more than one category through the use of one-hot vectors. The framework is based on the assumption that video frames can be represented by a latent space of images, which can be further decomposed into, and therefore reproduced by selectively sampling from, the content and motion subspaces. Sampling a single point in the content subspace and multiple trajectories in the motion subspace can generate the same object performing different motions, and vice versa. However, MoCoGAN generates random content instead of using input videos and is not currently capable of performing tasks in which videos must be generated while preserving the identity of the individual.

Video-to-video translation has the goal of transforming a video from source domain A to the style of target domain B. Inspired by CycleGAN and Pix2Pix, Recycle-GAN [4] was proposed as an unsupervised video retargeting method that translates content from one domain to another but preserves the style (or motion) from the source domain. In addition to cycle loss, the authors also implement recycle loss, which refers to updated cycle loss values across domains and over time as well as a recurrent loss, which is produced by a recurrent temporal predictor that is trained to predict future samples given past samples. The spatio-temporal 3D translator was proposed [25] to improve video-to-video translation by addressing the semantic inconsistencies and temporal artifacts that tend to be observed in the above approaches. Based on a conditional GAN, this method treats inputs and outputs as three-dimensional tensors such that the network takes in a volumetric image from domain A and produces a corresponding volume of the same shape in domain B. This is done with the help of a recurrent generator which consists of an image generator, a flow estimator and a fusion block. Similar to the CycleGAN, it uses two generator-discriminator pairs with the addition of cycle consistency loss. A similar approach is also outlined by [9] in their Motion-guided CycleGAN (Mocycle-GAN), which addresses the imposition of spatio-temporal constraints by explicitly modeling the motion across frames using optical flow.

3 METHODOLOGICAL FRAMEWORK

3.1 MoCoGAN

This work will build on the work conducted by Tulyakov and colleagues [29], namely the Motion Decomposed Generative Adversarial Network (MoCoGAN) for video generation. The MoCoGAN architecture is able to generate videos depicting different motions such as facial expressions, human actions and shape motions, and can do so across multiple categories of motion by conditioning the GAN using one-hot vectors. For this purpose, three inputs are fed into the MoCoGAN generator: (a) content, (b) motion and (c) categories. Both content and motion vectors are generated from random noise, while category labels are provided if more than one category of motion is present.

While the model is able to successfully generate short video sequences of individuals expressing different emotions, there are several challenges that can be addressed. First, MoCoGAN is a video generation method and uses noise to generate the face of a random individual displaying the required emotion – this means that it is currently not capable of generating videos in which the identity of a specific individual must be preserved. The authors discuss an extension of their model for image-to-video translation but the corresponding code was not made available to the public unlike the previous model. Additionally, preliminary experiments using the MoCoGAN indicated that the identity of the randomly generated individual was sometimes lost or distorted across certain frames of the videos produced, such that another individual appeared entirely, and artifacts were observed in the form of facial features (such as facial hair and hairstyles) that were added or removed.

3.2 Enforcing Identity Consistency

This work aims to address the task of identity consistent-video generation by building upon the MoCoGAN architecture. Videos can be thought of as a human face (content) that makes a given facial expression (motion) that corresponds to a particular emotion (category). The proposed video generator therefore requires that the content, which in this case refers to the human faces depicting one of six emotions, be fixed for each generated video. This is to ensure that the identity of the individual is preserved throughout the video across the different generated emotions and reduces the likelihood of artifacts. The main goal of the proposed model is to therefore enforce identity consistency.

3.2.1 Autoencoder.

For this, an autoencoder is trained and used to inform the content sub-space of the generated videos. More specifically, input frames are encoded to form compressed representations of the human face present in the video and used to produce the *content*. Autoencoders are unsupervised deep learning models that are used for the task of representation learning. They can be trained to encode or compress data, and then reconstruct it back from the encoded representation such it resembles the original as closely as possible. One of the main advantages of the autoencoders is the ability to capture low-dimensional features by learning to ignore the noise in the data, making them suitable for dimension reduction tasks and thus are commonly used in applications such as anomaly detection [10], image denoising [31] [32] and image reconstruction [37].

3.2.2 Content Consistency Loss.

The above-mentioned autoencoder is used to encode the content representation present in each video. In addition, a *content consistency loss* is proposed to enforce the *content*, or identity of the selected face to ensure that the same individual is generated regardless of the emotion that is being generated. This is implemented by applying the pre-trained autoencoder on pairs of real and generated frames after which the resulting encoded representations are mapped onto each other for consistency.

4 PROPOSED METHOD

The goal of this framework is to flexibly generate videos representing different emotion categories given an input while retaining the identity of the individual in the video. We make similar assumptions to that of MoCoGAN [29] and adopt the premise that in a latent space of images Z_I, each vector z represents an image while a video is represented by $[z^{(1)}, z^{(2)}, ..., z^{(K)}]$ containing K frames. In order to disentangle motion and content from a video, Z_I is further decomposed into a content Z_C and motion subspace Z_M. Below, we discuss our proposed method for how an autoencoder is used to produce content representations for each frame in the generated videos. The rest of this section describes the architecture, training and implementation of the

proposed VICE-GAN network, which comprises of a generative adversarial network, autoencoder and recurrent neural network.

4.1 Utilizing Autoencoders for Encoding Content

4.1.1 *Overview*.

In order to preserve the identity or *content* throughout the video across the different generated emotions or *motion*, an autoencoder was used to produce the content representations for the input frames of a video.

4.1.2 *Implementation*.

Autoencoders were either standard (convolutional) [27] or variational [21], and each type of autoencoder had three varying architectures increasing in size (small, medium and large). In the standard autoencoder, the encoding network produces a single value of an encoding dimension for every incoming observation. In contrast, the encoder of the variational autoencoder provides a probability distribution for each observation in the latent space, giving the network the added benefit of learning latent representations with disentangled factors [17]. The standard and variational autoencoder contains the same number of convolutions per architecture in order to make the results comparable. Six autoencoder models were trained based on the configurations in Table 1.

	Small		*Medium*		*Large*	
Type	***#Conv***	***#Deconv***	***#Conv***	***#Deconv***	***#Conv***	***#Deconv***
Standard	2	2	4	3	5	4
Variational	2	2	4	3	5	4

Table 1: Autoencoder Model Configurations

4.1.3 *Training details*.

The small standard autoencoder (SAE) contains two convolution layers in the encoder block, followed by two max pooling layers for downsampling. The linear layer succeeding this is responsible for encoding the compression. The decoder block contains another linear layer which decompresses the features from the previous layer, which is followed by two fractionally-strided convolutions for upsampling. The medium and large SAEs were investigated to see if there was an improvement on reconstruction performance (see Table 1). The same configurations were implemented for the variational autoencoder (VAE) with the exception being that the linear layers were replaced with a mean and standard deviation layer allowing it to sample across a continuous space based on the data it has learned.

The models were trained on 6 emotions from the MUG Facial [2] dataset, and a train-test split of 80-20 is performed on the dataset. All models were trained from scratch for 50 epochs using the adam optimizer with a learning rate of 0.005 and a batch size of 32. The SAE and VAE models use a reconstruction loss function to measure the error between the original and the reconstructed data. Additionally, the VAE models also use a regularization term namely the Kullback-Leibler divergence [21] which forces the encoder layer to distribute close to normal distributions and thereby allowing the model to create more general latent spaces. Mean square error is the reconstruction loss function which is used during the training of all the models.

The network details of the autoencoders can be found in the Appendix A.

4.2 Proposed Approach: VICE-GAN

Towards achieving our goal, we propose a framework (Figure 4) that consists of the following 5 sub-networks: (i) Autoencoder I_e, that encodes each input frame and produces the content representation Z_C, (ii) Recurrent neural network R_m, that generates a set of motion vectors which represent the motion dynamics Z_M in a video, (ii) Generator G_I, that accepts Z_C (content), Z_M (motion) and Z_A (category) as inputs and generates the video sequences, (ii) Image discriminator D_I that determines whether a generated image is real or fake, and (iii) Video discriminator D_V, that determine whether a set of frames in a video are real or fake and in addition evaluates the authenticity of the generated category-specific motion.

4.3 Network Architecture

4.3.1 *Autoencoder for encoding content*.

As we are dealing with visual data, the autoencoder is tasked with encoding K frames which represent the content aspect of a video.

$$Z_C = I_e(X)$$

where, X is an input video that contains a set of input frames $[x_1, x_2, ...x_K]$ and I_e is the trained autoencoder which encodes K frames belonging to a video X.

Two content encoding schemes were introduced for producing the content vectors in a video. In the **Single Frame** scheme, the autoencoder takes the first frame of the video and produces a representation for it ($z_c^{(1)}$) and then fixes the same representation for K frames in a video.

While in the **All frames** scheme, the autoencoder produces a individual content vector for each frame in a video. In other words, Z_C contains a set of content vectors $[z_c^{(1)}, ..., z_c^{(K)}]$ that represent the respective frames of a video.

4.3.2 *Recurrent Neural Network for modelling motion*.

As the identity of the individual remains the same in a video with only motion changing between the frames, it is important to model this change between frames. RNNs [19] are useful for modelling sequence of data such that each sample in the sequence is dependent on, or correlated with, the previous one.

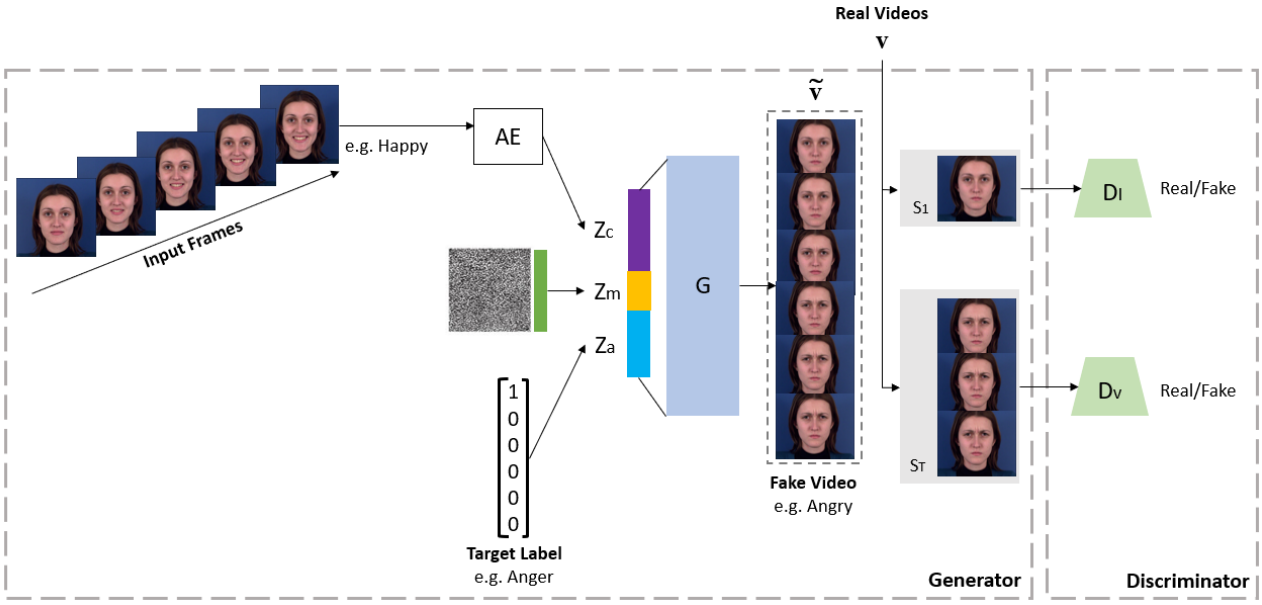

Figure 4: Model pipeline for video generation

*This figure illustrates the model pipeline for video generation. On the left are a sequence of extracted and concatenated frames of an individual from one of six emotion categories obtained from the MUG facial database e.g. happiness. These frames are fed into the selected autoencoder model **AE** to produce a 50-dimensional representation of the human face and constitutes the **content Zc**. An additional recurrent neural network is used to transform random noise into a sequence of correlated variables that represent the **motion Zm**, or in this case facial expression, which the content will be performing Finally, the input is augmented using a one-hot vector encoded variable **category Za** that represents the target emotion category. The above components are concatenated and fed into a 2D decoder architecture that generates a sequence of frames depicting the same individual expressing the target emotion e.g. anger. The image discriminator randomly samples single frames from real and generated videos, while the video discriminator randomly samples T consecutive frames.*

$$Z_M = R_m(\epsilon)$$

where, ϵ is a vector that is sampled from a gaussian distribution, R_m is the recurrent neural network which generates motion vectors from the noise vectors and Z_M is the motion representation or space that contains a set of motion vectors.

The recurrent neural network R_m is a one-layer GRU network which is responsible for generating the vectors $[z_m^{(1)}, ..., z_m^{(K)}]$ in Z_M which constitutes the motion representation in a video. Similar to [29], noise is injected at every iteration to model uncertainty of the ensuing motion at each step.

4.3.3 *Image and Video Discriminators.*

The network uses two types of discriminators - an image discriminator D_I and a video discriminator D_V. D_I is based on a standard CNN architecture that provides criticism to the G_I based on randomly-sampled individual images or frames. The purpose of D_I is to determine whether a frame is sampled from a set of real or fake videos. Based on the findings of [29], it was found that the addition of D_I improved the overall training of the GAN model since focusing on stationary appearances is relatively easier.

D_V is of a spatio-temporal type architecture that samples the frames from a video clip in order to determine if the set of frames was sampled from the real or fake videos. D_V penalizes the motion aspect of the video and sends the feedback back to R_m. In addition, the D_V also attempts to learn the different categories present in the training data. By doing so, it generates category labels for generated videos which are then compared to the original labels to enforce accurate category-specific generations.

4.3.4 *Generator.*

The generator model G_I is fed two components, the content vectors z_c in Z_C and the motion vectors z_m in Z_M in order to capture the dynamics of a video. Additionally, a categorical one hot vector is also added so that the generator can produce video-specific emotions. The goal of G_I is to produces realistic generations based on the criticisms provided by the discriminators D_I and D_V. The generator is of a decoder type architecture so by concatenating the z_c, z_m and z_a and providing this as input to G_I, it will attempt to generate a video sequence. The model has a generator network composed of 4 transposed convolution layers for upsampling. Batch normalization is used in the generator network.

During training, we experimented with two types of content encoding schemes in the generator. In the first scheme i.e., single frame method, we fix the content once and repeat it K times and this is shown in the following equation.

$$\left[\begin{bmatrix} z_a \\ z_m^{(1)} \\ z_c \end{bmatrix} .... \begin{bmatrix} z_a \\ z_m^{(K)} \\ z_c \end{bmatrix}\right]$$

Alternatively in the second scheme i.e., all frame method, we produce independent content vectors for K frames and this is represented by the below equation.

$$\left[\begin{bmatrix} z_a \\ z_m^{(1)} \\ z_c^{(1)} \end{bmatrix} .... \begin{bmatrix} z_a \\ z_m^{(K)} \\ z_c^{(K)} \end{bmatrix}\right]$$

The network configurations of the above sub-networks can be found in the Appendix A.

4.3.5 ***Objective functions***.

Full Objective loss.

The full objective loss function contains an adversarial loss L_{adv}, a content-consistency loss $L_{content}$ and a category loss $L_{category}$. The loss can be formulated as follows:

$$L_{obj} = L_{adv}(G_I, D_I, D_V, R_m) + L_{content}(G_I) + L_{category}(G_I, D_V)$$

Adversarial loss.

In order to generate videos which are difficult to distinguish from the real videos, an adversarial loss [14] is adopted. The adversarial loss generally, refers to the simultaneous optimization of the two networks namely, the generator and the discriminator. The generator is encouraged to generate realistic data that can fool the discriminator while the discriminator seeks to distinguish the real data from the generated data. The training of the generator and discriminator networks is achieved via a min-max manner.

The adversarial objective for our model $L_{adv}(G, D_I, D_V, R_m)$ can be expressed as follows,

$$\mathbb{E}_{v\sim p_v}[-\log D_I(G_I(Z_C, Z_M))] + \mathbb{E}_{v\sim p_v}[-\log(1 - D_I(G_I(Z_C, Z_M)))] + \mathbb{E}_{\tilde{v}\sim p_{\tilde{v}}}[-\log D_V(G_I(Z_C, Z_M))] + \mathbb{E}_{\tilde{v}\sim p_{\tilde{v}}}[-\log(1 - D_V(G_I(Z_C, Z_M)))]$$

Here the first and second terms in the loss function encourage the image discriminator D_I to classify the individual frames from the real v and fake $\tilde{v}$ videos. Based on the third and fourth terms, the video discriminator D_V is encouraged to distinguish a consecutive set of frames from v and $\tilde{v}$. The generator G_I and recurrent neural network R_m attempt to produce realistic video sequences based on the second and fourth terms in the equation.

Content-consistency loss.

It was initially hypothesized that the addition of a reconstruction loss between the real and fake videos might be sufficient enough to achieve identity consistency. However, this approach was found to be unsuccessful. This may have arisen from the model not being able to differentiate between the content and motion aspects of the video when measuring the reconstruction loss.

To address this, a content consistency loss was proposed. As the generated videos contained motion and content, it was theorized that reproducing the content representations of the real and fake videos and computing the loss between the two representations would be able to enforce the identity when generating to different domains.

To carry this out, we leverage the autoencoder, which produces a content representation for each frame for all pairs of real $z_c(v)$ and fake $z_c(\tilde{v})$ videos. The loss is computed as the mean squared error between each pair of real and fake feature representations.

$$L_{content} = \mathbb{E}[z_c(v) - z_c(\tilde{v})]$$

where z_c is a content representation produced by the pretrained autoencoder.

Category loss.

To model the categorical dynamics (aspects) of a video, the input to the generator is conditioned with a categorical random variable z_a, where each category is represented by a one hot vector, and the dimensionality of this vector is equal to the number of categories present. The addition of the one-hot vector allows the model to perform multi-domain generation with a single network, and more specifically, allows the generator to generate videos corresponding to the six different emotions present in the dataset. Since the frames in a given video belong to the same category, we keep the realization fixed for all frames in that video.

The category loss is represented as follows, $L_{category}(G_I, D_V)$. D_V attempts to learn the categories from the training video while G_I tries to generate categories that are recognizable from the video discriminator.

5 EXPERIMENTS

The following section describes the utilized dataset and experimental set-up, followed by an overview of the planned experiments and the evaluation measures that will be used in the study.

5.1 Dataset

The MUG Facial Expression Database [2] consists of videos from 86 Caucasian subjects (35 women and 51 men), although only data from 52 participants is available to authorized users. Subjects were seated in front of a blue background and on a chair

in front of a single camera (examples can be seen in Figure 5). Videos were captured at 19 frames per second. Subjects were requested to perform six basic expressions corresponding to the following emotions: anger, happiness, disgust, fear, sadness and surprise. The video sequences start and end at neutral state and follow the onset, apex, offset temporal pattern. In addition, for each subject a short video sequence depicting the neutral state was recorded. The distribution of data with respect to the number of videos and participants is shown in Figure 6

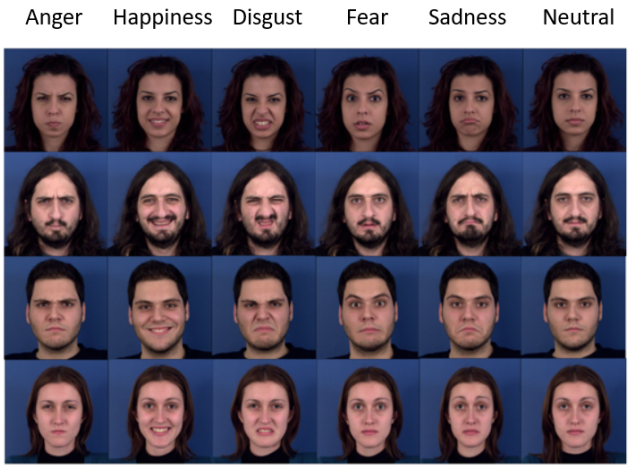

Figure 5: Examples of participants from MUG Database displaying the six emotions [2]

5.1.1 Pre-processing. All frames were extracted from each video, and were resized to 64x64 for practical reasons. Videos with less than 64 frames were removed in order to comply with the selected frame-sampling method from [29]. The resulting dataset comprised of 50 participants, and a total of 548 videos.

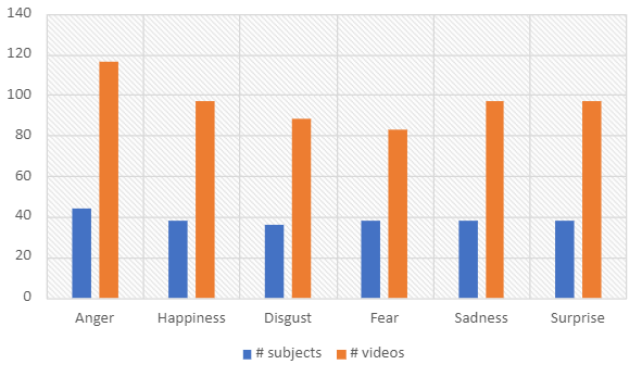

Figure 6: MUG Facial Database

The graph shows (a) number of subjects (blue) and (b) number of videos (orange) present for each emotion category

5.2 Experimental Setup

The training of the GAN models occurs on a single TITAN X 12GB GPU machine located in the CTIT cluster at the University of Twente. In all the experiments, the image and video batch were set to the size of 32. The adam solver was used for training, with a learning rate of 0.0002. The β_1 and β_2 were equal to 0.5 and 0.999 respectively. The models were saved every 10,000 steps in order to observe the generation of the video sequences.

5.3 Experimental Overview

Four GAN variants are described below in Table 2, which differ along two variables: the content encoding method and fine-tuning of the autoencoder. All models were also trained on the basis a third variable, namely composition of training data, giving rise to eight model variants which are compared using the evaluation measures listed in section 4.2.

Model	*Loss functions*	*Content Encoding*	*Fine-Tuning*
1	Adv+Category+Content-consistency	All Frames	✓
2	Adv+Category+Content-consistency	All Frames	-
3	Adv+Category+Content-consistency	Single Frame	✓
4	Adv+Category+Content-consistency	Single Frame	-

Table 2: Overview of VICE-GAN variants

5.3.1 Composition of Training Data.

First, given a 80-20 train-test split, we devised two approaches to testing the model. First, all videos pertaining to the test subjects were removed from the training data such that during testing, the model would be seeing these faces for the first time **(Unseen)**. In the second strategy, videos of the test subjects were included in the training data but only for two out of the six emotions. Therefore, the model would have seen these faces before but would not have seen emotion-specific training data corresponding to the remaining four emotions which will be generated during testing **(Seen)**. On the basis of preliminary experiments, it was hypothesised that (a) the model may need to see an individual during training in order to reproduce that individual, and (b) that allowing the model to be trained on atleast two emotion-specific samples would allow the individual's identity to be reproduced effectively while still being able to flexibly generate unseen emotions for those individuals.

5.3.2 Content Encoding Method.

Using the autoencoder model, content can be encoded in two ways: (1) all 16 frames of a given video are encoded to produce sixteen content vectors **(All Frames)** and (2) only the first frame of a given video is encoded to produce a single content vector which is then repeated sixteen times to produce sixteen content vectors **(Single Frame)**. While the first approach may provide some variation between the content vectors, the second approach may keep the content more constant and increase content consistency across frames.

5.3.3 Fine-Tuning Autoencoder Weights.

Third, we hypothesise that allowing the pre-trained autoencoder to continue training alongside the GAN may result in better results regarding identity consistency. While the weights of the autoencoder are normally frozen during GAN training, the alternative is that the weights will be updated according to, and could benefit from, the generator loss.

5.4 Evaluation

	SSIM	*MSE*	*ACD*	*CAS - M*	*CAS - H*
Quality	✓	✓	-	-	-
Identity	-	-	✓	-	-
Emotion	-	-	-	✓	✓

Table 3: Overview of evaluation measures

5.4.1 Structural Similarity Index Measure.

The single-scale structural similarity index measure (SSIM) [34] is a well-characterized perceptual similarity measure that aims to discount aspects of an image that are not important for human perception. It compares corresponding pixels and their neighborhoods in two images using three quantities i.e., luminance, contrast and structure. The three quantities are combined to form the SSIM score.

To evaluate the autoencoder, the SSIM score was computed to measure the similarity between the real and reconstructed images. A high SSIM score indicates that the reconstructed image was similar to the real images, and by extension, that the compressed representation produced by the encoder was of good quality.

To evaluate the frames generated by VICE-GAN, the SSIM score was computed to measure the similarity between the real and generated frames. A high SSIM score indicates that the generated frames are similar to the actual frames, suggesting that the generated frames are of high quality.

5.4.2 Mean-squared error.

Mean-square error (MSE) is another measure which is widely used to assess image similarity [6]. It is calculated as the average of the squared differences between the actual and predicted target values.

To evaluate the autoencoder, MSE was computed to measure the similarity between the real and reconstructed images. A low MSE indicates that reconstructed image was similar to the real images, and by extension, that the compressed representation produced by the encoder was of good quality.

To evaluate the frames generated by VICE-GAN, the MSE was used to calcuate the error between the real and generated frames in a video. A low MSE indicates that the generated frames are more likely to resemble the real frames, suggesting that the generated frames are of high quality.

5.4.3 Average Content Distance.

Average Content Distance (ACD) proposed by [29] measures the content consistency of the videos, and refers to the average L2 distance among all consecutive frames in a video. A feature vector is produced for each frame in a generated video using the pretrained autoencoder, and then the ACD is computed using an average pairwise L2 distance of the per-frame vectors in a video. A reference ACD is also computed on the real videos that correspond to the test subjects, which allows a direct comparison between the reference and generated frames.

A smaller ACD means that the generated frames in a video are perceptually more similar, and vice versa. As the result, the identity of the individual is more likely to be preserved between frames.

5.4.4 Classification Accuracy Score - Machine.

A pre-trained classifier was obtained from [3], which consists of a convolutional neural network that was trained on the FER-2013 in-the-wild emotion dataset. The model achieved 66% classification accuracy. The classifier was used to categorize the generated videos into the six emotion classes (anger, happiness, disgust, fear, sadness, and surprise), and compare the inferred labels with the labels of real data.

5.4.5 Classification Accuracy Score - Human ratings.

10 videos were generated for each emotion, which resulted in 60 videos per model. Only videos generated in the *Seen* condition were selected as it performed superior to the *Unseen* condition for both visual quality and emotion generation, and produced videos of sufficient quality to be rated by humans. Participants were shown a random set of 60 video sequences and were asked to indicate which emotion they would assign to each sequence. Participant responses were compared with the true labels and percentage accuracies were computed for each emotion and were averaged to produce an overall score for each model.

6 RESULTS

In this section, the autoencoder selection is presented followed by an in-depth evaluation of the VICE-GAN. The generated video sequences are evaluated along (a) visual quality, (b) identity consistency and (c) emotion generation quality, and are interpreted both quantitatively and qualitatively.

6.1 Autoencoder Selection

Table 4 shows MSE and SSIM scores for the small-, medium- and large-variants across standard and variational autoencoder models. The largest standard autoencoder was found to perform best, as indicated by the highest MSE and SSIM scores.

Type	*Small*		*Medium*		*Large*	
	MSE	SSIM	MSE	SSIM	MSE	SSIM
Standard	0.001156	0.909136	0.000984	0.937786	**0.00086**	**0.961358**
Variational	0.002331	0.872967	0.001905	0.9062063	0.001277	0.938729

Table 4: Performance of autoencoder variants

6.2 VICE-GAN Evaluation

6.2.1 *Visual Quality.*

Quantitative

Quantitatively, visual quality was evaluated using structural similarity index metric (SSIM) and mean squared error (MSE). Figure 7 and Figure 8 show the SSIM and MSE values obtained using the four models.

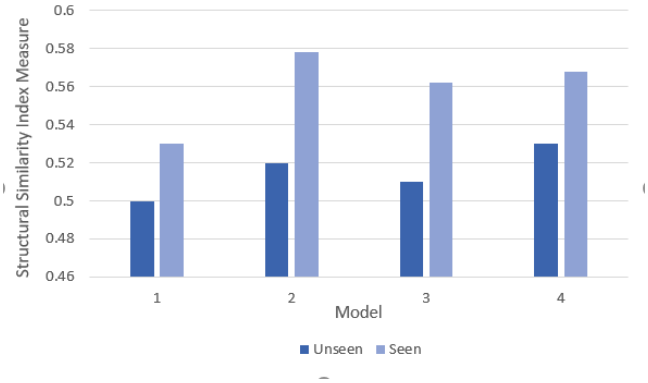

Figure 7: Structural Similarity Index Measure

Comparing the performance of the models based on composition of training data, it can be seen that the *Seen* condition resulted in higher SSIM and lower MSE values, indicating higher similarity between the real and generated frames, and therefore increased visual quality.

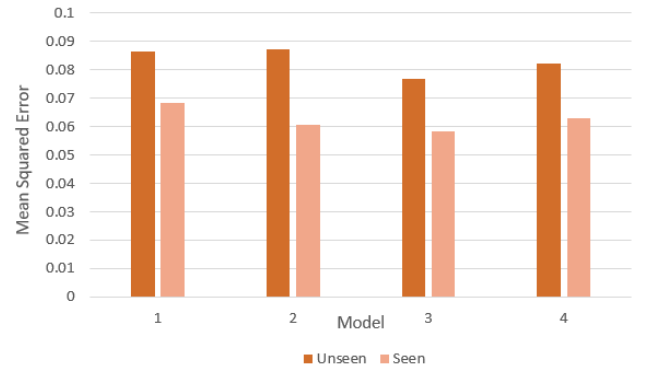

Figure 8: Mean Squared Error

Given that the models in the *Seen* condition were more likely to produce higher quality videos, the results were further analyzed within this strategy. Considering the SSIM scores (Figure 7), Model 2 resulted in the highest value, indicating higher quality, while Model 1 resulted in the lowest value, indicating lower quality. The scores for Models 3 and 4 were comparable. Overall, Models 2 and 4 indicated higher visual quality than Models 1 and 3. These pairs of models differed in the way the autoencoder weights were frozen, with the latter having the weights not frozen.

Considering MSE scores (Figure 8), Model 3 resulted in the lowest value, indicating higher quality, while Model 1 resulted in the highest value, indicating lower quality. That said, the performance of Models 2, 3 and 4 were comparable. As MSE is computed pixel-by-pixel and tends to be insensitive to differences in internal structure, it is to be expected that the values across the models are similar to one another and high - this is because in each video, all frames represent the same individual with only slight changes to their facial expressions. In contrast, SSIM has been shown to be more meaningful when applied to images and videos as it measures perceptual similarity by modelling similarity as a combination of structure, luminance and contrast.

Qualitative

Figure 9 shows examples of video sequences for a selected individual who is expressing the same emotion (Happiness) in the *Seen* condition, and illustrates the visual quality obtained using the four model variants, allowing us to make qualitative observations. In Models 1 and 3, the individual's face is distorted, and therefore the facial features which allow us to detect emotions are unclear. Model 2 appears to have produced frames in which the visual quality achieved is moderate despite the ambiguous emotion which resembles disgust. In contrast, Model 4 resulted in high visual quality, which allows us to identify the individual and the corresponding emotion clearly.

Figure 9: Comparison of generated video sequences across model variants

6.2.2 *Identity Consistency.*

Quantitative

Figure 10 shows the average content distance (ACD) for the four model variants in comparison to the reference (computed using training data) across the two types of training data composition. In the *Seen* condition, the ACD values are higher than in the *Unseen* condition, indicating good identity consistency between frames. Looking at the models in the *Seen* condition, Models 1 and 2 were found to exhibit the highest ACD values, indicating low identity consistency between frames. In contrast, Models 3 and 4 resulted in the lowest ACD values, indicating high identity consistency between frames, with Model 4 performing the closest to the reference. These pairs of models differed in their content encoding method, with the latter encoding the first frame and using it as the content vector for all subsequent frames.

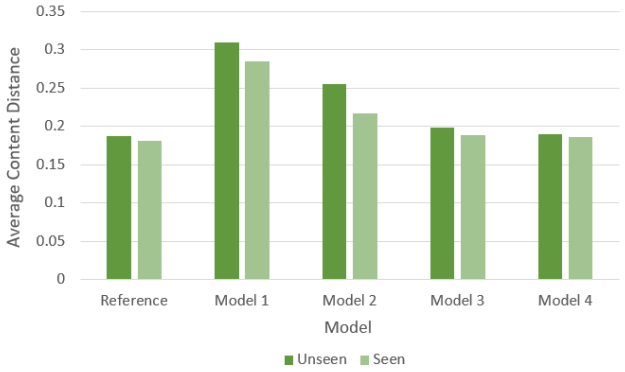

Figure 10: Average Content Distance

Qualitative

Figure 11 shows examples of several video sequences generated in the *Unseen* and *Seen* conditions. To reiterate, in the *Unseen* condition, the test subjects were removed completely from the training data, whereas in the *Seen* condition, test subjects were retained in the training data but only for two random emotions such that the remaining emotions were generated during evaluation.

It can be seen that in the *Unseen* condition, the correct emotion is captured but the identity of the individual has been lost when compared to the original video. Instead, the model appears to have retrieved the individual that it has encountered before which is most similar to the input received. This phenomenon is not observed in the *Seen* condition, which suggests that the identity of the individual is more likely to be preserved if the model has seen the individual before, even if it has not seen them make the same emotion expression. This is consistent with the findings presented in the previous section, in which the *Seen* condition resulted in improved identity consistency.

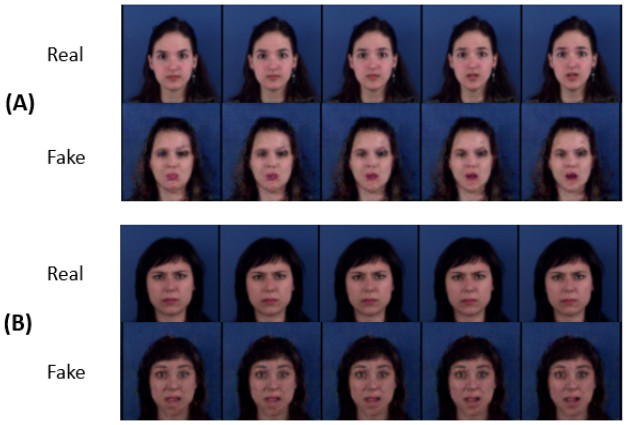

Figure 11: Comparison of generated video sequences based on composition of training data - (A) Unseen condition (B) Seen condition

6.2.3 *Emotion Generation.*

Emotion generation was evaluated using two types of classification accuracy scores (CAS): using a pre-trained classifier (CAS-M) and using human subjective ratings (CAS-H).

Quantitative

Figure 12 shows overall CAS scores for the four models achieved using the pre-trained classifier on the generated video sequences. Based on Figure 12, Models 1 and 4 resulted in the highest and comparable CAS scores in the *Unseen* condition while Model 2 resulted in the highest CAS score *Seen* condition. However, the differences in performance on the basis of training data composition were not noteworthy. CAS scores obtained for each emotion across the four models in the *Seen* condition can be found in Appendix A.

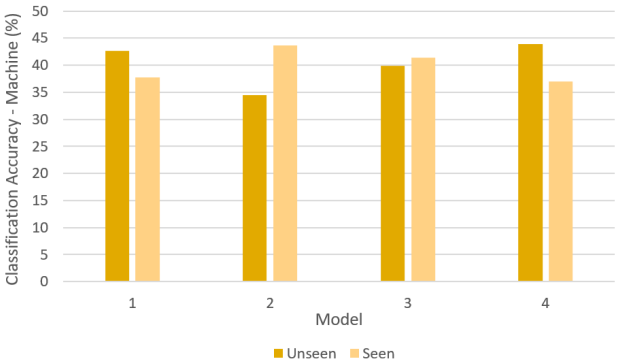

Figure 12: Classification Accuracy Score (CAS-M) - Using a Pre-Trained Classifier

Qualitative

Figure 13 shows overall CAS scores for the four models obtained using subjective ratings provided by human participants who were shown video sequences generated in the *Unseen* condition (in dark blue). Model 2 outperformed the other models with

the highest CAS-H score with regards to generating accurate emotions, followed by Model 4 while Models 1 and 3 resulted in noticeably lower CAS scores. These pairs of models differed in the way the autoencoder weights were frozen, with Models 2 and 4 having the weights not frozen. CAS scores obtained for each emotion across the four models in the *Seen* condition can be found in Appendix A.

Figure 13 also shows the comparison between CAS scores obtained using the pre-trained classifier (CAS-M) and using human subjective ratings (CAS-H). Looking at CAS-H scores, Model 2 resulted in the highest CAS scores, and therefore performed better than the other models with regard to emotion generation. Comparing CAS-M and CAS-H scores, it can be seen that the CAS scores were on average higher when using human ratings when compared to the pre-trained classifier.

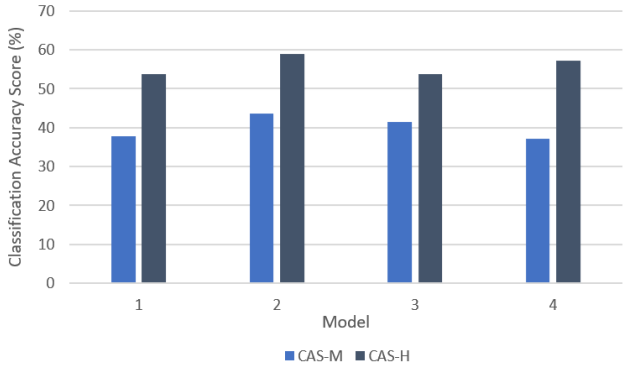

Figure 13: Classification Accuracy Score - Comparison between Pre-Trained Classifier (CAS-M) and Human Ratings (CAS-H)

This suggests that the generator is indeed performing at an adequate level, and that the low CAS scores obtained using the pre-trained classifier can be explained by the quality of the classifier itself. Therefore, the CAS-H scores, and therefore the use of human ratings, can be considered more robust and valid as an evaluation measure in this context.

6.2.4 *Holistic Qualitative Evaluation.*

Figure 14 presents several short video sequences from different model variants in the *Seen* condition in order to provide examples of good and poor generations produced by VICE-GAN.

The generated frames are shown as falling along two dimensions: visual quality, which refers to clarity and presence of distortions, and emotion generation, which refers to the discernibility and accuracy of emotional expressions. In the top-left corner are instances of low visual quality but discernable and correct emotional expressions - this is in contrast to the examples presented below that, in which low visual quality is also accompanied by poor emotion generation. On the right are examples of frames characterized by high visual quality, although the instances in the bottom-right corner show that it is possible to obtain clear frames without the appropriate emotion.

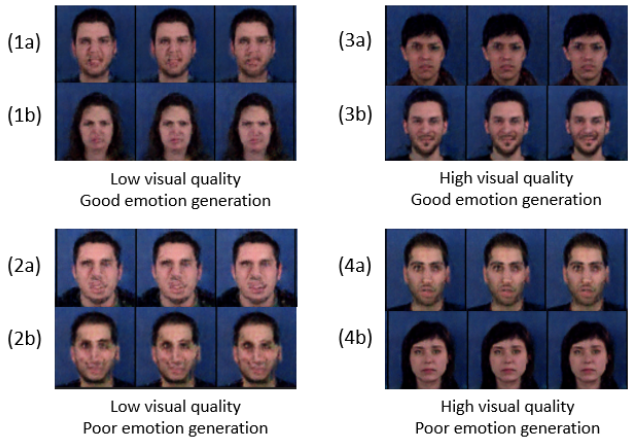

Figure 14: Examples of low and high quality generated video sequences

While these examples can be found across all model variants, frames of low visual quality were often produced by Model 1, while Models 2, 3 and 4 were overall found to generate frames of higher visual quality and discernible emotions. These qualitative observations are also consistent with the quantitative measures discussed earlier in this section.

The results of this section are summarized below (Table 5), where the influence of the three chosen experimental variables is shown as either present (✓) or absent (-) on the three evaluation aspects - visual uality, identity consistency and emotion generation.

Experimental Variable	*Visual Quality*	*Identity Consistency*	*Emotion Generation*
Training Data Composition	✓	✓	-
Content Encoding Method	-	✓	-
Autoencoder Fine-Tuning	✓	-	✓

Table 5: Overview of VICE-GAN Results based on Experimental Variable

7 DISCUSSION

The present study sought to develop a deep generative model that is applicable for generating video sequences of individuals expressing different emotions whilst retaining the individual's identity. Throughout this thesis, We have discussed the different methodological and experimental questions and the previous section presents the results. In this section, these results will be interpreted and discussed, and by doing so, each research question presented at the start of this thesis will be answered.

RQ2a: *To model the content sub-space, which autoencoder architecture produces the most accurate, high-quality compressedrepresentations of human faces?*

Six autoencoder variants were trained on the MUG facial dataset and evaluated using mean squared error (MSE) and structural similarity index measure (SSIM). For the purpose of selecting an autoencoder, the standard autoencoder with the largest architecture was found to outperform the other variants and was used in the subsequent experiments to model the content sub-space. That said, the difference in performance between the variants was sufficiently small that a smaller architecture may have also been used for computational efficiency. While resource constraints limited experimentation to one autoencoder, the variational autoencoder may have been useful for encoding content as it models the distribution of faces it has been trained on, therefore being able to generalize better to unseen individuals and by extension, preserving identity more effectively.

RQ2b: *Can identity consistency be additionally enforced through the use of a content consistency loss?*

Preliminary experiments indicated that encoding the content representation of the input frames alone was insufficient to enforce identity consistency. The addition of a regular reconstruction loss function computed using real and generated sequences did not additionally improve the results. The loss function was refined by applying the autoencoder to the real and generated sequences to produce content representations which were then used to compute content consistency loss. The addition of the refined loss function was successful in reproducing the identity of the test subjects, and was therefore implemented in the final models. Identity consistency was evaluated using average content distance (ACD) and qualitative observations, and it was found that individual identity was preserved to a large degree across models, and on average, they performed comparably to the reference videos. Thus, the addition of the content consistency loss was instrumental for this purpose.

RQ1(a): *How do the following experimental variables influence the generated video sequences: (1) composition of training data?*

It was observed that in the *Seen* condition, the models were able to produce higher quality video sequences. Moreover, the composition of training data influenced identity consistency, such that *Seen* condition, the models were able to reproduce the identity of the individual more successfully. To reiterate, the *Seen* condition allowed the model to see examples of the test subjects during training but only for two emotions while the remaining four emotions were generated during testing. This suggests that identity consistency is better achieved if the model has seen the individual before. This could be a limitation of the standard autoencoder model used in this study, and as mentioned previously, a variational autoencoder can be investigated in future work in order to circumvent the issue of identity consistency for individuals that the model has not yet seen.

RQ2(b): *How do the following experimental variables influence the generated video sequences: (2) content encoding method*

In order to determine the influence of content encoding method, Models 1 and 2 are compared to models 3 and 4. For both visual quality and emotion generation, Model 2 consistently performed well but in contrast, Model 1 generally performed poorly. On the other hand, Models 3 and 4 generally performed comparably for visual quality but Model 4 outperformed Model 3 for emotion generation using human ratings. In line with our hypothesis, Models 3 and 4 outperformed Models 1 and 2 with regards to identity consistency. These pairs of models differed in their content encoding method, with Models 3 and 4 using the content encoding method in which the first frame is encoded and the resulting content vector is used for all subsequent frames of the video. If unique content vectors are produced for each frame, it is possible that the identity is not preserved across all frames, leading to a loss of identity in the generated video. As expected, fixing the identity by using only the content vector produced for the first frame may have allowed increased control over reproducing the identity as desired.

RQ3(c): *How do the following experimental variables influence the generated video sequences: (3) fine-tuning the autoencoder*

In order to determine the influence of freezing autoencoder weights during GAN training, Models 1 and 3 are compared to Models 2 and 4. Model 2, and to an extent, Model 4, outperformed Models 1 and 3 on the basis of visual quality and emotion generation. This suggests that utilizing the pre-trained autoencoder without fine-tuning it during GAN training resulted in higher quality video sequences and more accurate emotion expressions. In other words, allowing the autoencoder to continue training alongside the GAN provided no measurable benefit and may have even negatively impacted the results. It was initially hypothesized that fine-tuning the autoencoder may allow it to benefit from the generator loss and produce better representations. However, negative transfer may have occurred due to the category loss of the generator that was back-propagated to the autoencoder as it was not designed to learn multiple emotion categories.

RQ3(d): *Which of the model configurations perform best in terms of (a) visual quality and (b) identity consistency and (c) accuracy of the generated emotions?*

The model (Model 2), which was configured to use a unique content vector for each frame and to freeze autoencoder weights, seemed to exhibit superior performance for both visual quality and emotion generation as seen by SSIM, MSE and CAS scores. The model (Model 4), which was configured to repeat the first content vector for all frames and to freeze autoencoder weights, performed closely to Model 2 for visual quality and emotion generation, but outperformed it for identity consistency.

7.1 Limitations and Future Work

It was observed that the current models had difficulty generalizing to fully unfamiliar or unseen individuals, and this is an area which requires further work. One potential solution is the use of software such as OpenFace, which can be used to map out facial landmarks and generate features which can then be used to train the autoencoder and generator instead of the raw videos.

This may aid in more accurate emotion generation but could also improve the model's ability to generalize to unseen faces.

Another limitation that was observed was the lack of motion, or changes between frames in several video sequences. One way to improve this would be to focus on pre-processing the training data more effectively. For example, a metric such as SSIM can be used to discard frames with the least change in a pairwise manner, and a video classifier can be used to validate the training data and only select video takes that accurately represent each emotion. Another approach could be to include a temporal generator or predictor and implement an additional motion consistency loss that enforces dynamic changes between frames [4]. It is also plausible that a one-step approach to videos may be more suitable for producing better results, in contrast to the two-step approach that was taken in this project. Specifically, instead of decomposing video into content and motion, the two can be considered simultaneously using a spatiotemporal 3D generator as suggested by Wang and colleagues [33].

8 CONCLUSION

We presented the VICE-GAN model for generating identity-consistent emotion-specific videos. With an appropriate training strategy, VICE-GAN is able to retain the identity of a chosen individual when producing videos belonging to different emotion categories by fixing the content subspace.

This was achieved through the introduction of a trained autoencoder and an additional content consistency loss in order to enforce the mapping between the content present in the input and generated videos. Each frame is defined by a vector containing two parts: content and motion. Specifically, the content refers to the compressed representation produced by the autoencoder on the input video, and the motion is a series of correlated vectors produced by a recurrent neural network that models motion dynamics across frames. Additionally, the generator is conditioned on one of six emotions during training to enable multi-domain generation with a single generator-discriminator network.

Our mixed-method evaluation supports the ability of VICE-GAN to produce videos of desired individuals that are of good quality and can accurately capture six different emotions. Importantly, this work also investigated the role of three experimental variables on model performance to inform subsequent studies. Overall, model variants resulted in higher visual quality and increased identity consistency if they were trained on videos containing the test subjects for two emotions than if the models were not exposed to these individuals at all. It was also found that freezing autoencoder weights during GAN training often produced higher quality videos and improved emotion generation while repeating the content vector obtained for the first reference frame across all frames instead of producing a unique content representation for each frame resulted in improved identity consistency.

Avenues for future work are discussed, with an emphasis on improving pre-processing techniques and efforts to model temporal dynamics, or motion more effectively. While emotion generation was chosen as the application for the VICE-GAN in this study, it is important to note that the model can be adapted to other domain-specific generation tasks as well.

REFERENCES

[1] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends* 2, 02 (2021), 52–58.

[2] N. Aifanti, C. Papachristou, and A. Delopoulos. 2010. The MUG facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. 1–4.

[3] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2017. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557* (2017).

[4] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*. 119–135.

[5] Dina Bashkirova, Ben Usman, and Kate Saenko. 2018. Unsupervised video-to-video translation. *arXiv preprint arXiv:1806.03698* (2018).

[6] Ali Borji. 2018. Pros and Cons of GAN Evaluation Measures. arXiv:1802.03446 [cs.CV]

[7] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. 2018. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 366–382.

[8] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. 2018. Attention-gan for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 164–180.

[9] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. 2019. Mocycle-gan: Unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 647–655.

[10] Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. 2018. Autoencoder-based network anomaly detection. In *2018 Wireless Telecommunications Symposium (WTS)*. IEEE, 1–5.

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 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*. 8789–8797.

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8188–8197.

[13] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Babu Chinnam. 2020. Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Transactions on Multimedia* 23 (2020), 391–401.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[15] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. 2018. GAN-based synthetic brain MR image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 734–738.

[16] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* 28, 11 (2019), 5464–5478.

[17] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. 2016. Early Visual Concept Learning with Unsupervised Deep Learning. arXiv:1606.05579 [stat.ML]

[18] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. *arXiv preprint arXiv:1806.04166* (2018).

[19] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Trinh Le Ba Khanh, Soo-Hyung Kim, Gueesang Lee, Hyung-Jeong Yang, and Eu-Tteum Baek. 2021. Korean video dataset for emotion recognition in the wild. *Multimedia Tools and Applications* 80, 6 (2021), 9479–9492.

[21] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]

[22] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. 2017. Dual motion GAN for future-flow embedded video prediction. In *proceedings of the IEEE international conference on computer vision*. 1744–1752.

[23] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Conditional image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5524–5532.

[24] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. 2018. Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European conference on computer vision (ECCV)*. 282–297.

[25] Kwanyong Park, Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. 2019. Preserving semantic and temporal consistency for unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1248–1257.

[26] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. 2017. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*. 2830–2839.

[27] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61 (Jan 2015), 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

[28] Sameerah Talafha, Banafsheh Rekabdar, Chinwe Pamela Ekenna, and Christos Mousas. 2020. Attentional adversarial variational video generation via decomposing motion and content. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. IEEE, 45–52.

[29] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. MoCoGAN: Decomposing motion and content for video generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1526–1535.

[30] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. 2017. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033* (2017).

[31] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. 1096–1103.

[32] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, 12 (2010).

[33] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. 2020. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1160–1169.

[34] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[35] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. 2019. Relgan: Multi-domain image-to-image translation via relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*. 5914–5922.

[36] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. 2019. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE). *IEEE transactions on medical imaging* 39, 1 (2019), 188–203.

[37] Jin Zheng and Lihui Peng. 2018. An autoencoder-based image reconstruction for electrical capacitance tomography. *IEEE Sensors Journal* 18, 13 (2018), 5464–5474.

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

A APPENDIX

A.1 Autoencoder network configurations

	Encoder
0	Conv-N32, K3, S2, P1, BN, ReLU
1	Conv-N64, K3, S2, P1, BN, ReLU
2	Conv-N128, K3, S2, P1, BN, ReLU
3	Conv-N256, K3, S2, P1, BN, ReLU
4	Conv-N512, K3, S2, P1, BN, ReLU
	Fully-connected Linear Layer (N=50)
	Decoder
0	Deconv-N256, K3, S2, P1, BN, ReLU
1	Deconv-N128, K3, S2, P1, BN, ReLU
2	Deconv-N64, K3, S2, P1, BN, ReLU
3	Deconv-N32, K3, S2, P1, BN, ReLU

Table 6: Autoencoder - Large architecture

	Encoder
0	Conv-N32, K4, S4, P1, BN, ReLU
1	Conv-N64, K4, S4, P1, BN, ReLU
2	Conv-N128, K3, S3, P1, BN, ReLU
3	Conv-N256, K3, S3, P1, BN, ReLU
	Fully-connected Linear Layer (N=50)
	Decoder
0	Deconv-N128, K3, S2, P1, BN, ReLU
1	Deconv-N64, K3, S2, P1, BN, ReLU
2	Deconv-N32, K3, S2, P1, BN, ReLU

Table 7: Autoencoder - Medium architecture

	Encoder
0	Conv-N16, K3, S1, P1, BN, ReLU
1	Conv-N4, K3, S2, P1, BN, ReLU
2	MaxPool, K2, S2, P1
	Fully-connected Linear Layer (N=50)
	Decoder
0	Deconv-N4, K2, S2, P1, BN, ReLU
1	Deconv-N16, K2, S2, P1, BN, ReLU
2	Deconv-N64, K2, S2, P1, BN, Sigmoid

Table 8: Autoencoder - Small architecture

Note: The Variational autoencoder configurations are the same as that of the standard autoencoder with the only exception being that the linear layer is replaced with a mean and standard deviation layers.

A.2 VICE-GAN network configuration

	Generator G_I
0	DeConv-N512, K6, S0, P0, BN, LeakyReLU
1	DeConv-N256, K4, S2, P1, BN, LeakyReLU
2	DeConv-N128, K4, S2, P1, BN, LeakyReLU
3	DeConv-N64, K4, S2, P1, BN, LeakyReLU
4	DeConv-N3, K4, S2, P1, BN, LeakyReLU
	Image Discriminator D_I
0	Conv-N64, K4, S2, P1, BN, LeakyReLu
1	Conv-N128, K4, S2, P1, BN, LeakyReLu
2	Conv-N256, K4, S2, P1, BN, LeakyReLu
3	Conv-N1, K4, S2, P1, Sigmoid
	Video Discriminator D_V
0	Conv3D-N64, K4, S1, P0, BN, LeakyReLu
1	Conv3D-N128, K4, S1, P0, BN, LeakyReLu
2	Conv3D-N256, K4, S1, P0, BN, LeakyReLu
3	Conv3D-N1, K4, S1, P0, Sigmoid

Table 9: VICE-GAN architecture

A.3 VICE-GAN: Supplementary Results

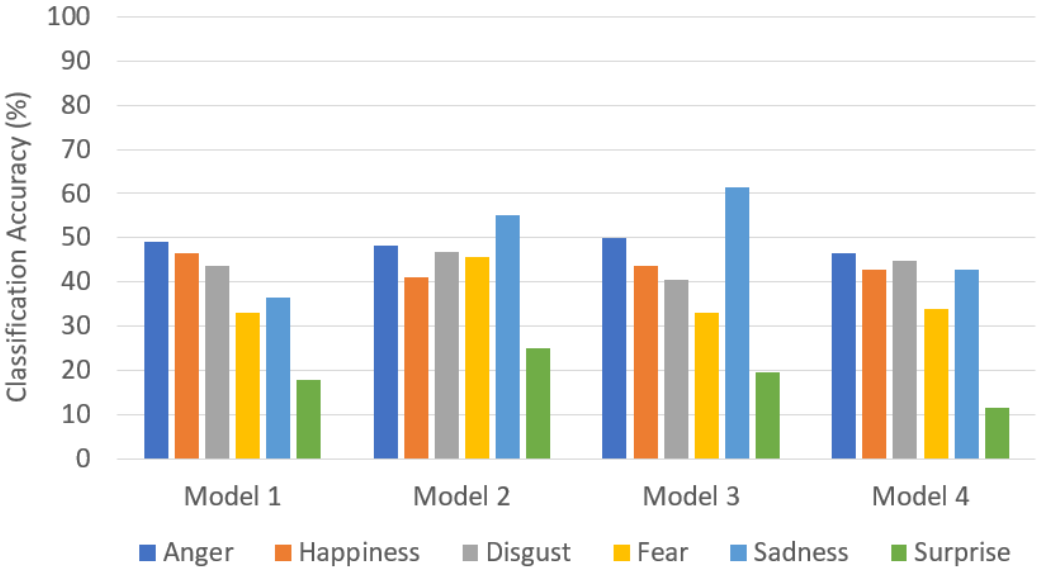

Figure 15: Emotion-Specific Classification Accuracy Score (CAS-M) - Using a Pre-Trained Classifier

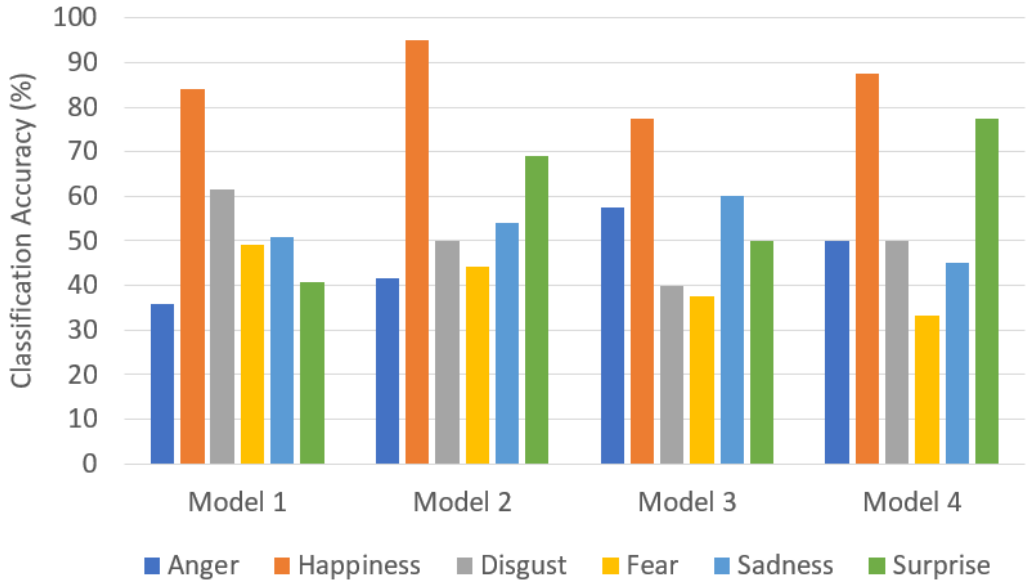

Figure 16: Emotion-Specific Classification Accuracy Score (CAS-H) - Using Human Ratings