2CET-GAN: Pixel-Level GAN Model for Human Facial Expression Transfer

Xiaohang Hu
xiaohang.hu@outlook.com
University of New South Wales
Sydney, NSW, Australia

Nuha Aldausari
n.aldausari@unsw.edu.au
University of New South Wales
Sydney, NSW, Australia

Gelareh Mohammadi
g.mohammadi@unsw.edu.au
University of New South Wales
Sydney, NSW, Australia

ABSTRACT
Recent studies have used GANs to transfer expressions between human faces. However, existing models have some flaws, such as relying on emotion labels, lacking continuous expressions, and failing to capture the expression details. To address these limitations, we propose a novel two-cycle network called 2Cycles Expression Transfer GAN (2CET-GAN), which can learn continuous expression transfer without using emotion labels in an unsupervised fashion. The proposed network learns the transfer between two distributions while preserving identity information. The quantitative and qualitative experiments on two public datasets of emotions (CFEE and RuFID) show our network can generate diverse and high-quality expressions and can generalize to unknown identities. We also compare our methods with other GAN models and show the proposed model generates expressions that are closer to the real distribution and discuss the findings. To the best of our knowledge, we are among the first to successfully use an unsupervised approach to disentangle expression representation from identities at the pixel level. Our code is available at github.com/xiaohanghu/2CET-GAN.

CCS CONCEPTS
• Computing methodologies → Computer vision; Neural networks: Unsupervised learning.

KEYWORDS
Facial expression transfer, Generative Adversarial Network, Unsupervised learning, Pixel-Level Learning

ACM Reference Format:
Xiaohang Hu, Nuha Aldausari, and Gelareh Mohammadi. 2023. 2CET-GAN: Pixel-Level GAN Model for Human Facial Expression Transfer. In Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE ’23), October 29, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3607541.3616822

1 INTRODUCTION
Human facial expression transfer aims to transfer the expression from a reference face to a target face while maintaining the target’s identity, which has exciting applications such as filmmaking, photo editing and entertainment. Figure 1 shows some examples of two emotion datasets. For example, the face located in the second row and second column of Figure 1(a) takes the expression from the left reference face and the identity from the upper target face. This process is also known as face reenactment, according to some studies on videos [9]. The existing studies of human facial expression transfer can be divided into three categories: Three dimensional (3D) model-based methods, landmark/points-based methods, and pixel-based methods.

3D model based methods [10, 16, 27–29] consider the 3D shape of the face contains the expression. They first extract 3D face models from the target and reference faces and then transfer the expression from the reference 3D model to the target 3D model, and finally combine the deformed target 3D model with the surface of the original face to generate the output. These steps can be implemented using mathematical-based methods and partially replaced by deep learning models [16]. The 3D-based methods are well established in movie and entertainment applications [1].

To simplify the 3D-based methods, landmark-based methods [11, 25, 26, 32] were introduced that use fiducial points to represent the facial geometry. Similar to 3D-based methods, expression can be transferred from the reference landmarks to the target landmarks, but the landmarks can not directly blend with the face’s surface since the landmarks do not contain sufficient geometric information. Thus, those methods usually employ a neural network to generate the final output by inputting the target image and the deformed target landmarks. Landmark-based generative models need large-scale datasets, and the intensity/number of points can limit the quality of the generated expressions.

Pixel-based methods [7, 22, 24] directly transfer expressions at the pixel level. These methods employ neural networks to handle the complexities of extraction, transformation, and generation. The removal of the intermediate facial geometry means the model can potentially capture all expression details, such as tears and the colour of the face. Some pixel-based models [5, 30] take the discrete emotion labels as conditional expression codes, such as happy, sad and angry; and the training can be implemented by Conditional Generative Adversarial Network (CGAN) [21] in a supervised manner. The introduction of Cycle Consistency Loss (CycleGAN) [33] removes the requirement of paired data for identity-preserving, which means each input does not need a corresponding true output for training. A further improvement was introduced in [7, 22], which through a “Controller Module”, maps discrete labels to a continuous space for generating continuous expressions. However, almost all existing pixel-based methods directly or indirectly rely on expression labels [7], and they hardly generate high-quality continuous expressions.
In this study, we propose a novel pixel-level GAN model that can transfer continuous expressions from one face to another without needing 3D/2D annotations or pre-trained models. We use only one label to distinguish natural faces from other expressions, unlike most models that use multiple expression labels. Our network contains an encoder and a generator. The encoder extracts diverse and continuous expression code from the reference face while the generator can apply the expression to the target face. The proposed model borrows ideas from CycleGAN[33], InfoGAN[4] and StarGAN v2[6] and combines them in a new fashion to transfer expressions while maintaining identity. The main contributions of this paper can be summed up as:

- We propose a novel two cycles architecture that can learn facial expressions in an unsupervised manner. It only requires labelling neutral faces in a dataset.
- The proposed model can extract and transfer continuous expressions, which can capture diverse details and provides flexibility to applications.
- It provides a GAN-based architecture that can learn the transfer between any two distributions while preserving identity information. This architecture can be further applied to learn the expression transfer on 3D shapes or landmarks.
- We provide qualitative and quantitative evaluations on two human emotion datasets and show our model is superior compared to baseline models.

2 RELATED WORK

In this section, we briefly discuss the related work that we have used or got inspiration from in the proposed model.

GAN. Generative Adversarial Network (GAN) [12] has been widely used in human face synthesis. A generator and a discriminator are the two main components of GANs. In the face synthesis paradigm, a generator synthesizes a human face and the discriminator judges whether the given sample is from the real distribution. The generator competes with the discriminator through the adversarial loss, and they learn from the error in each iteration in an unsupervised manner.

CycleGAN. Cycle-Consistent Adversarial Networks (CycleGAN) solves the identity-preserving problem by introducing cycle consistency loss [33]. It adds an inverse translator $F$ to transfer the generated fake face $G(x)$ back to the original face $x$, and computes the $L_1$ distance between the original face and generated-back face as the cycle consistency loss: $\|x - F(G(x))\|_1$. Thus, the forward generator in CycleGAN keeps the identity information. We employ the cycle consistency loss in both cycles to keep the identity on the generated face.

InfoGAN. Information Maximizing Generative Adversarial Networks (InfoGAN) [4] learns disentangled representations in an unsupervised manner by maximizing the mutual information $I(\hat{c}; G(z, \hat{c}))$ between a random latent code $\hat{c}$ and the observations $x$. The generator $G$ takes a random latent code $\hat{c}$ as an additional input, the discriminator $D$ extracts the code out from the fake image, and then the "info loss" $\|\hat{c} - D(G(z, \hat{c}))\|_1$ is used to maximize the mutual
information. Thus, the latent code \( \hat{c} \) will correspond to some semantic features of the data, such as the pose, colour or expression of the face. InfoGAN successfully discovers emotions on the CelebA [18] face dataset. We employ a continuous multidimensional random latent code \( \hat{c} \) as an additional input to a CycleGAN and maximize \( Q(\hat{c}; x) \) by a code loss \([c - E(G(x, \hat{c}))]\). Our experiment shows the network successfully disentangles the representation as emotion over diverse identities.

StarGAN v2 [6] also adopts a CycleGAN- and InfoGAN-based architecture, and it utilizes the diversity-sensitive loss [19, 31] to ‘push’ the generator to explore the instance space widely to generate diverse styles. The experiment of StarGAN v2 shows that the diversity-sensitive loss is essential for style diversity, even though the discriminator receives diverse real instances. The visual quality of the images generated by StarGAN v2 is quite high, and the FID [14] reaches 23.8 on the CelebA-HQ test set [15]. We adopted the same Residual Network-based modules from StarGAN v2 [19], but instead of using a mapping network to transfer the latent code to the style code, we use the latent code as the expression code directly.

3 PROPOSED METHOD

We consider the difference between a neutral face and an emotional face of the same identity to be expression. Based on this, the aim of our proposed network is to learn a unified and disentangled expression representation while excluding identity information.

3.1 Modules

Our network contains three modules: an encoder (E), a generator (G), and a discriminator (D). We use \( x_c \) to represent the emotional face and \( x_n \) to represent the neutral face. The expression code is a \( d \)-dimensional vector \( c \in \mathbb{R}^d \), and we set \( \hat{0} \) as the expression code for the neutral face.

Encoder. The encoder \( E \) takes a real or fake emotional face as input and outputs an expression code \( \hat{c} = E(x_c) \) (The hat means the vector is produced by the network). The training goal of the encoder is to extract an expression code and it does not contain any identity information.

Generator. The generator \( G \) takes an emotional face \( x_c \) or neutral face \( x_n \) and an expression code \( c \) as input and outputs a fake face \( \hat{x}_c = G(x_c, c) \) or \( \hat{x}_n = G(x_c, \hat{0}) \). Note that the generator can not transfer an emotional face to another emotional face directly. This design simplifies the task of the generator since it only needs to learn two one-to-many mappings (one neural expression to many emotional expressions and vice versa) rather than a many-to-many mapping (many emotional expressions to many emotional expressions) over each identity.

Discriminator. Our discriminator takes a real or fake face and a face type label (0 for natural face; 1 for emotional face) as input and outputs whether the input is a real image or fake one, such as \( \hat{y} = D(x_c, 1) \).

3.2 Structure

The structure of our model is a combination of CycleGAN and InfoGAN, which results in a different new architecture, and we use some components of starGAN. As Figure 2 shows, it consists of 2 cycles, where the encoder, generator and discriminator are shared among these two cycles:

(a) Neutral to Emotional to Neutral cycle (N to E to N). As Figure 2(a) shows, this cycle transfers a neutral face to an emotional face and then transfers it back to the neutral face for the same identity. In the forward flow, a multi-dimensional expression code \( \hat{c} \) is sampled from a uniform or Gaussian distribution. The discriminator \( D \) will receive real emotional faces randomly from the dataset, so the generator \( G \) has to transfer the neutral face to an emotional face to compete with the discriminator, which is achieved through an adversarial loss

\[
L_{adv, e} = \mathbb{E}_{x_c}[\log(D(x_c, 1))] + \mathbb{E}_{x_n, \hat{c}}[\log(1 - D(G(x_n, \hat{c}), 1))].
\]

In the bottom backflow, similar to CycleGAN, the fake face is transferred back to the original face at the pixel level. The backflow forces the fake face to be identity-preserving by comparing it with the input face. We take the pixel level \( L_1 \) distance as the cycle consistency loss

\[
L_{cyc, n} = \mathbb{E}_{x_n, \hat{c}}[|x_n - G(G(x_n, \hat{c}), \hat{0})|].
\]

Note that if the generator \( G \) interprets the expression code part as identity features, the network will fail on the cycle consistency loss. In the upper backflow, the encoder \( E \) extracts the expression code from the fake emotional face. Similar to InfoGAN, we compare the extracted code with the random expression code to maximise the mutual information, and thus the code will correspond to the expression on the generated face, this forms the code loss

\[
L_{code} = \mathbb{E}_{x_n}[\|c - E(G(x_n, \hat{c}))\|].
\]

Similar to StarGAN v2, we use the diversity sensitive loss [19, 31] to let the generator \( G \) generate different expressions for the same identity, which is achieved by pushing two fake emotional faces of the same person away from each other. The difference between expressions is minor compared to the difference between identities (average Euclidean distance ratio is around 1 : 2), this further makes the diversity-sensitive loss indispensable.

\[
L_{ds} = \mathbb{E}_{x_c, \hat{c}_1, \hat{c}_2}[|G(x_c, \hat{c}_1) - G(x_c, \hat{c}_2)|].
\]

\( \hat{c}_1 \) and \( \hat{c}_2 \) are sampled independently. The full objective of the N to E to N cycle is

\[
\min_{E, G} \max_D \lambda_{adv, e} L_{adv, e} + \lambda_{cyc, n} L_{cyc, n} + \lambda_{code} L_{code} - \lambda_{ds} L_{ds},
\]

where \( \lambda_{adv, e}, \lambda_{cyc, n}, \lambda_{code} \) and \( \lambda_{ds} \) are weights for each term. We observed that the generator learns to generate identities at first and then gradually learn to generate diverse expressions. We employed a dynamic code loss weight \( \lambda_{code} \), which will gradually increase from 0 to a constant. Thus the effect of the code loss in the total loss will align with the learning of the generator. In this cycle, we do not use the expression code of the real reference face but only the random code \( \hat{c} \), which is for reducing bias, since it prevents the network from paying attention to expression codes from the training samples.

(b) Emotional to Neutral to Emotional Cycle (E to N to E). As Figure 2(b) shows, this cycle transfers an emotional face to a neutral face and then transfers it back for the same identity. The
4 EXPERIMENTS AND RESULTS

4.1 Datasets

We have used two datasets to train and test the proposed model. 

CFEE: The Compound Facial Expressions of Emotions Database (CFEE) [8] is used as the main evaluation dataset in our experiments. CFEE contains 230 individuals of different races, each individual has 26 expressions (including the neutral expression), and all faces are in a frontal pose. We mix all emotional faces as the emotional face group and all neutral faces as the neutral face group to build our dataset. Then the training and test set are split at the identity level: 222 individuals for the training set; 8 individuals for the test set. The identity-level splitting is for generalization validation.

RaFD: Radboud Faces Database (RaFD) [17] contains 67 individuals, each with 8 expressions. The photos are taken from 5 camera angles, and each expression was shown with three gaze directions. We use three camera angles: 45°, 90° and 135° in our experiment since the expressions on 0° and 180° are not clear. We take neutral faces from all three camera angles in the frontal gaze direction as the neutral face group and the rest of the images as the emotional face group. We also split the data at the identity level: 61 individuals for the training set; 6 individuals for the test set.

4.2 Training

For both datasets, input images are resized to 128 x 128 resolution with random cropping and horizontal flipping. The random expression code is sampled from a 32 dimensional uniform distribution $U(-0.5, 0.5)^{32}$. We used the uniform distribution instead of the Gaussian distribution to provide an equal chance to train extreme expressions, and the analysis of the components (Section 4.6) showed uniform distribution works better. The code loss weight $\lambda_{code}$ is set to 0 at step 1 and then linearly increases to 1.0 at step 5000. The diversity sensitive loss $\lambda_{ds}$ is set to 1.0 at step 1, and then linearly decreases to 0 at step 10000. The Encoder will convert the image to grayscale before computation. More details of hyperparameters can be found at the code base. The network took 3 days to train on a single Tesla V100-16GB for 100K iterations (steps) on each dataset. We compute the Fréchet Inception Distance (FID, lower is better), [14] on the test set after each 5000 step. The lowest point on the CFEE test set steps 65000, which was used in further analysis.
Figure 3: Qualitative comparison of GANimation, StarGAN v2 and our approach on the CFEE training set. The first row is real target faces. On the second row, two real reference faces are selected for each of the four emotions: happy, angry, sadly surprised and sadly disgusted. All three methods can generate differences within the same emotion, which means they can capture expression details.

4.3 Baseline Models

We adopt GANimation and StarGAN v2 as two baseline models. Both of these are pixel-level GAN models that can perform expression transfer tasks (Note that we avoid comparing with 3D-based methods such as Neural Emotion Director[23] as we aim at improving the methods in the pixel domain). Both models are trained with their original implementations\(^1\)\(^2\). Both models can transfer an emotional face to another emotion directly. The main drawback of these two models is that they are both supervised and require emotion labels. In our approach, we address this issue by making the model unsupervised. In addition, we can control the intensity of the generated emotion.

GANimation adopts a Conditional GAN (CGAN) [21] architecture, which takes Action Units (AUs) as the condition. AUs describe continuous anatomical facial movements, which is a precise and meaningful expression representation. GANimation relies on third-party tools to extract AUs from the image, such as OpenFace\(^3\) [2]. GANimation also employs a cycle consistency loss to preserve identity. We trained GANimation on the CFEE dataset for 750 epochs.

\(^1\)https://github.com/albertpumarola/GANimation
\(^2\)https://github.com/clovaai/stargan-v2
\(^3\)https://github.com/TadasBaltrusaitis/OpenFace

StarGAN v2 learns the cross-domain style transfer in a supervised manner. It adds a mapping network to map latent codes to style codes for different domains, which is combined with a diversity sensitivity loss for generating diverse styles within the same domain. We train StarGAN v2 on the CFEE dataset by taking 26 expression labels as 26 domains (the training on two domains, neutral face domain and emotional face domain, failed to produce diverse expressions. We believe this is mainly because StarGAN v2 cannot effectively learn the difference if the distributions of the two domains are close to each other.).

4.4 Qualitative evaluation

Our expression transfer results on the CFEE and RafD datasets are shown in Figure 1. We observed that our method could transfer expressions over different races on the CFEE dataset and over different angles and ages on the RafD dataset. And as shown by the child in the third row and third column of RafD, it can transfer gaze directions as well. The result also shows deep learning-based methods can derive unknown regions, such as teeth.

Qualitative comparison. In Figure 3 a qualitative comparison of GANimation, StarGAN v2 and our approach is demonstrated. The output of GANimation is not stable, especially when the input
expression and output expression differ greatly; the face on the third column shows this issue clearly. StarGAN v2 performs poorly on identity-preserving compared to our approach, and although it retains the overall style (hair, skin colour, etc.), the details of the face characteristics are altered. One possible reason for our higher identity-preserving quality is that our method focuses on converting natural faces to emotional faces and vice versa, avoiding learning the many-to-many conversion mapping done by GANimation and StarGAN v2. In addition, StarGAN v2 consumes computations in unnecessary translations from and to the same expression. All three methods can capture diverse expression details; the differences between fake faces of the same emotion show that. StarGAN v2 employs a mapping network to map the latent code to the expression code while we take the latent code as the expression code directly; our experiment shows the mapping network can not help to produce diverse expressions in our architecture.

4.5 Quantitative comparison

For quantitative analysis, we generated two fake image sets by GANimation, StarGAN v2, and our network to compute Fréchet Inception Distance (FID, lower is better) [14] and Inception Score (IS, higher is better) [3]. The two fake image sets are: (1) fake neutral faces generated by emotional faces \(x_n = G(x_e)\); and (2) fake emotional faces generated by neutral faces and reference faces \(x_e = G(x_n, E(x_n))\). The result of the FID score comparison is shown in Table 1, which indicates the images generated by our approach are closer to the real distribution than StarGAN v2. AUs, as a precise expression representation, help GANimation achieve a better FID score on fake emotional faces. But the IS, which implies the quality in unnecessary translations from and to the same expression. All three methods can capture diverse expression details; the differences between fake faces of the same emotion show that. StarGAN v2 employs a mapping network to map the latent code to the expression code while we take the latent code as the expression code directly; our experiment shows the mapping network can not help to produce diverse expressions in our architecture.

| Method         | FID | IS |
|----------------|-----|----|
|                | N   | E  | N   | E  |
| GANimation     | 51.5| 29.1| 1.48| 1.49|
| StarGAN v2     | 55.4| 35.4| 1.54| 1.54|
| Ours           | 44.4| 31.1| 1.49| 1.55|
| Real images    | -   | -  | 1.62| 1.61|

Table 1: FID and IS comparison of GANimation, StarGAN v2 and ours on the CFEE test set. The error of FID is around ±0.3, and the error of IS is around ±0.03. N: fake neutral face; E: fake emotional face generated by reference face. The gap of FIDs between the neutral face and emotional face is because the neutral face set is much smaller than the emotional face set, which is a known bias of FID score.

expression and output expression differ greatly; the face on the third column shows this issue clearly. StarGAN v2 performs poorly on identity-preserving compared to our approach, and although it retains the overall style (hair, skin colour, etc.), the details of the face characteristics are altered. One possible reason for our higher identity-preserving quality is that our method focuses on converting natural faces to emotional faces and vice versa, avoiding learning the many-to-many conversion mapping done by GANimation and StarGAN v2. In addition, StarGAN v2 consumes computations in unnecessary translations from and to the same expression. All three methods can capture diverse expression details; the differences between fake faces of the same emotion show that. StarGAN v2 employs a mapping network to map the latent code to the expression code while we take the latent code as the expression code directly; our experiment shows the mapping network can not help to produce diverse expressions in our architecture.

4.5 Quantitative comparison

For quantitative analysis, we generated two fake image sets by GANimation, StarGAN v2, and our network to compute Fréchet Inception Distance (FID, lower is better) [14] and Inception Score (IS, higher is better) [3]. The two fake image sets are: (1) fake neutral faces generated by emotional faces \(x_n = G(x_e)\); and (2) fake emotional faces generated by neutral faces and reference faces \(x_e = G(x_n, E(x_n))\). The result of the FID score comparison is shown in Table 1, which indicates the images generated by our approach are closer to the real distribution than StarGAN v2. AUs, as a precise expression representation, help GANimation achieve a better FID score on fake emotional faces. But the IS, which implies the quality in unnecessary translations from and to the same expression. All three methods can capture diverse expression details; the differences between fake faces of the same emotion show that. StarGAN v2 employs a mapping network to map the latent code to the expression code while we take the latent code as the expression code directly; our experiment shows the mapping network can not help to produce diverse expressions in our architecture.

| Method         | Identity Preserving Score | Expression Transfer Score |
|----------------|---------------------------|---------------------------|
| GANimation     | 4.62 (0.87)               | 3.56 (1.54)               |
| StarGAN v2     | 4.64 (1.06)               | 3.42 (1.62)               |
| Ours           | 4.82 (0.64)               | 3.37 (1.69)               |

Table 2: Human evaluation of GANimation, StarGAN v2 and ours on the CFEE dataset. Each score is in the range [0, 5]. The number in the bracket is the standard deviation, which implies all models perform more stable on identity-preserving compared to expression transfer.

Human evaluation. We further evaluated the generated samples based on the CFEE dataset by 25 workers on Amazon Mechanical Turk (AMT). We focused on two aspects which are identity preservation quality and expression transfer quality. We asked the workers to evaluate each sample using a 5-point Likert-scales from "Definitely not" to "Definitely yes". We randomly selected three images for each of the 25 emotions as reference faces and randomly selected one target face for each reference face. Then, we used GANimation, StarGAN v2 and our model to generate samples for each reference and target pair. The total number of generated samples is 225. We shuffled these samples along with 10 extra quality control samples. The purpose of the quality control questions is to assess the workers’ attention to the task, and we filtered out six workers whose response rates to the quality control samples were very low. For each question, we removed outliers using a statistical outlier analysis by removing samples with a z-score above 3 or below −3 and then averaged the remaining scores as the final score. This ensures that those evaluations that are far away from the rest are excluded. As shown in Table 2, the results are consistent with our observations. Our model achieved the highest identity preservation score among the three models; the expression transfer score is close to the supervised learning model StarGAN v2. The AUs-guided model GANimation is more accurate in expression transfer but over-distorted in some samples. It is important to note that the difference across the three models in both identity preservation and emotion transfer are not statistically significant, which suggests that overall they all perform similarly. However, more evaluation is required to check whether there are emotion-specific differences across the three models or not. For example, one model might have more difficulty in generating high-quality images for a smiley face.

4.6 Components analysis

To demonstrate the effect of different components of the model, we cumulatively change a component in our final model to analyse the impact on the generated results. The FID and IS are computed after removing the many-to-many conversion mapping done by GANimation and StarGAN v2. AU s, as a precise expression representation, help GANimation achieve a better FID score on fake emotional faces. But the IS, which implies the quality of generated images, for GANimation is lower than StarGAN v2 and our approach.

**Human evaluation.** We further evaluated the generated samples based on the CFEE dataset by 25 workers on Amazon Mechanical Turk (AMT). We focused on two aspects which are identity preservation quality and expression transfer quality. We asked the workers to evaluate each sample using a 5-point Likert-scales from "Definitely not" to "Definitely yes". We randomly selected three images for each of the 25 emotions as reference faces and randomly selected one target face for each reference face. Then, we used GANimation, StarGAN v2 and our model to generate samples for each

FIDs show that each component makes generated images closer to the true distribution on the test set, especially the grayscale encoder component.
Figure 4: Qualitative evaluation of generalization on the CFEE and RaFD test set. The first row is real target faces, the second row is real reference faces, and the last row is our outputs. None of the identities appeared during the training phase. It shows our network is able to generalize to unknown identities.

| Component                | FID   | IS    |
|-------------------------|-------|-------|
| final model             | 32.4  | 1.53  |
| - dynamic loss weight   | 33.5  | 1.54  |
| - grayscale encoder     | 37.3  | 1.57  |
| - uniform distribution  | 38.0  | 1.49  |

Table 3: Components analysis. Both FID and IS are computed after each modification.

4.7 Generalization of unknown identities
Some samples of the generated fake faces from the CFEE and RaFD test sets are provided in Figure 4. None of these identities was presented during the training. We observed that our network can generalize to unknown identities to some extent for expression extraction and generation. The examples generated from the test set did not achieve the same level of identity-preserving as in the training set, especially on the RaFD dataset. One reason could be the CFEE training set contains only 214 individuals, and the RaFD training set has only 61 individuals, which is minuscule compared to the whole population.

5 DISCUSSION AND LIMITATION
The combination of random expression codes and diversity-sensitive loss facilitates the network to explore the entire instance space in order to produce diverse expressions. Explicitly converting the many-to-many mapping problem to two one-to-many mapping problems helps reduce the learning burden. The network learns the differences between two groups, which could include details such as posture, gaze direction, tears, etc.

The nature of our model is to learn the transfer of deformation. We speculate this is one of the reasons that our model achieves a lower expression transfer score on some of the faces because, although two faces being close on deformation are usually expressing the same emotion, there are cases that this is not true. We used very neat training sets to train our model; for the CFEE training set, each identity in the neutral face group has 25 corresponding expressions in the emotional face group. If the identities appear in the neutral face group then do not appear in the emotional face group, the training will fail in identity-preserving. To train a general and diverse network, there is a need for a rich dataset for capturing all poses, races, ages, lighting conditions and backgrounds etc.; this is a general issue with all pixel-level expression generation models. For evaluation, third-party emotion recognition algorithms can be used to produce more accurate analyses.

6 ETHICS
The proposed method could be used to generate fake expressions of individuals, which can be used for malicious purposes, including bullying or humiliating individuals and to impact political campaigns. These malicious applications can raise moral, ethical, and legal issues in society, such as invasion of privacy and distortion of democracy. The use of fake detecting technology on content distribution platforms can limit the spread of malicious fake images. Public laws can help regulate the use of such fake generation techniques. More discussion can be found at [20].

7 CONCLUSION
In this paper, we propose a novel GAN-based network to transfer expressions between human faces, which aims to address two major problems: unsupervised learning of expression transfer and generating continuous and diverse expressions. We combine elements of CycleGAN [33] and InfoGAN [4] in a novel way to solve the identity-preserving problem and the unsupervised learning problem. The experiment shows our network can be trained without
emotion labels and can transfer continuous and detailed expressions on CFEE and RaFD datasets. The evaluation shows the generated expressions by our approach are closer to the real distribution and have a higher identity preservation quality compared to GANimation and StarGAN v2, and yet its expression transfer quality is comparable to them. To the best of our knowledge, we are among the first to successfully use an unsupervised method to learn the expression representation and transformation. For further research, removing the requirement of neutral face labels and testing the network on more diverse situations, such as videos, can be explored.

REFERENCES

[1] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. 2009. Creating a photoreal digital actor: The digital emily project. In 2009 Conference for Visual Media Production. IEEE, 176–187.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG) 2018. IEEE, 59–66.

[3] Shane Barratt and Rishi Sharma. 2018. A note on the inception score. arXiv preprint arXiv:1801.01973 (2018).

[4] Xi-Chen, Yan Duan, Jeremy Peck, Hongyi Chen, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing adversarial nets. Advances in neural information processing systems 29 (2016).

[5] Yunjie 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.

[6] Yunjie Choi, Youngjung Um, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7] Hui Ding, Kumar Sricharan, and Ram Chellappa. 2018. Exqrnn: Facial expression editing with controllable expression intensity. In Proceedings of the AAAI conference on artificial intelligence, Vol. 32.

[8] Shuchuan Du, Yong Tao, and Alexi M Martinez. 2014. Compound facial expressions of emotion. Proceedings of the national academy of sciences 111, 15 (2014), E1454–E1462.

[9] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. 2014. Automatic face reenactment. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4217–4224.

[10] Pablo Garrido, Michael Zollhofer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Perez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. ACM Transactions on Graphics (TOG) 35, 3 (2016), 1–15.

[11] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-removing the requirement of neutral face labels and testing the network for diverse image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 1429–1437.

[12] Edvinas Mekyskas, Julija Kalpokiene, Paulius Jurys, and Aidas Lidanduskas. 2020. Regulating deep fakes: legal and ethical considerations. Journal of Intellectual Property Law & Practice 15, 1 (2020), 24–31.

[13] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014).

[14] Xin Ning, Shaoshui Xu, Xinyi Zong, Weijuan Tian, Linjun Sun, and Xiaodi Deng. 2020. Emotongan: facial expression synthesis based on pre-trained generator. In Journal of Physics: Conference Series, Vol. 1518. IOP Publishing, 012031.

[15] Albert Pumarola, Antonio Agudo, Alexi M Martinez, Alberto Sanfeliu, and Francesco Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In Proceedings of the European conference on computer vision (ECCV). 818–833.

[16] Edvinas Mekyskas, Julija Kalpokiene, Paulius Jurys, and Aidas Lidanduskas. 2020. Regulating deep fakes: legal and ethical considerations. Journal of Intellectual Property Law & Practice 15, 1 (2020), 24–31.

[17] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014).

[18] Xin Ning, Shaoshui Xu, Xinyi Zong, Weijuan Tian, Linjun Sun, and Xiaodi Deng. 2020. Emotongan: facial expression synthesis based on pre-trained generator. In Journal of Physics: Conference Series, Vol. 1518. IOP Publishing, 012031.

[19] Albert Pumarola, Antonio Agudo, Alexi M Martinez, Alberto Sanfeliu, and Francesco Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In Proceedings of the European conference on computer vision (ECCV). 818–833.

[20] Edvinas Mekyskas, Julija Kalpokiene, Paulius Jurys, and Aidas Lidanduskas. 2020. Regulating deep fakes: legal and ethical considerations. Journal of Intellectual Property Law & Practice 15, 1 (2020), 24–31.

[21] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014).

[22] Xin Ning, Shaoshui Xu, Xinyi Zong, Weijuan Tian, Linjun Sun, and Xiaodi Deng. 2020. Emotongan: facial expression synthesis based on pre-trained generator. In Journal of Physics: Conference Series, Vol. 1518. IOP Publishing, 012031.

[23] Albert Pumarola, Antonio Agudo, Alexi M Martinez, Alberto Sanfeliu, and Francesco Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In Proceedings of the European conference on computer vision (ECCV). 818–833.

[24] Edvinas Mekyskas, Julija Kalpokiene, Paulius Jurys, and Aidas Lidanduskas. 2020. Regulating deep fakes: legal and ethical considerations. Journal of Intellectual Property Law & Practice 15, 1 (2020), 24–31.