GANemotion: Increase Vitality of Characters in Videos by Generative Adversary Networks

Muhammad Hassan*, Yutong Liu†, Linghe Kong†, Ziming Wang*, and Guihai Chen†
*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.
Email: {hassankhan, wangziming1022}@sjtu.edu.cn
†Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.
Email: {isabelleliu, linghe.kong}@sjtu.edu.cn, gchen@cs.sjtu.edu.cn

***Abstract*—Increasing the vitality of facial expression for characters in videos can greatly improve the user experience on entertainments, like in films, animations, news broadcasting, or even a static painting. Recent advances have been made on facial expression migration, especially by GANs for better performance, while most of them cannot be automatically applied in consecutive frames.**

In this paper, we propose *GANemotion*, an automatic facial expression migration architecture for real-time videos. Three main modules are designed for this purpose: audio emotion classification module, facial expression classification module, and facial expression migration module on GANs. The outputs of the first two modules will be the inputs of the third module, where the emotion type acquired from text analysis will be the target generation label and image from facial classification is the object for processing. Expression attention algorithm and standard AU library are specially designed for avoiding distortion and processing on non-real world characters. Experiments have been applied to each module and the whole architecture respectively. They show the feasibility of *GANemotion* with higher classification accuracy, and vivid generation, breaking the limits of non-real world characters and low illumination conditions.

I. Introduction

Recently, an artificial intelligence (AI) robot newsreader is firstly launched in China's state-run Xinhua News Agency, whose name is Anchor [1]. Anchor is modelled on presenter Zhang Zhao and learns from live videos, who can report via social media 24 hours without a break. As it is declared as "world first", this robot newsreader seems not so much satisfactory, with its indifferent emotions and unchanged facial expressions (shown in the first column of Fig. 1). Generally speaking, virtual characters like this robot newsreader are widely applied in daily entertainment, such as film shooting, virtual live shows, AR interactions, and so on. It is necessary to increase the vitality of their facial expressions, for providing better user experiences.

Facial expressions, actually, are the combination and coordination of the actions on facial muscles [2]. To anatomically analyze such muscle movements, Ekman et al. [3] proposed a Facial Action Coding System (FACS) to represent facial expressions in terms of 44 standard Action Units (AUs). Taking the facial expression *surprise* as an example, it can be generally represented by 4 AUs: inner brow raiser (AU-1), outer brow raiser (AU-2), lip part (AU-25), and jaw drops (AU-26) [4], [5].

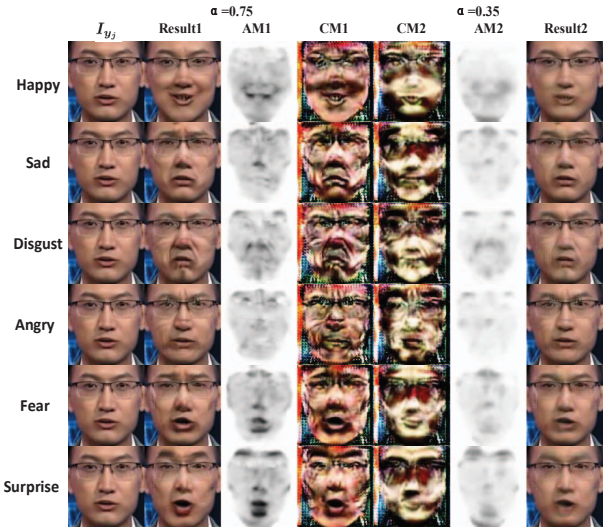

Fig. 1. Facial Expression Generation results for *GANemotion* [AM-Attention Mask, CM-Color Mask, $\alpha = \alpha_1$ in section. IV-C1 represents the degree of AU effects.]

To increase the vitality of facial expressions, for instance, to change a neutral face to a happy one, is actually a classical image migration problem. Early approaches for solving this problem use mass-and-spring models [6] or 2D/3D morphings [7], [8], to artificially approximate skin and muscle movements, but with less reality. Further enhanced, recent advances have been made by GANs to preserve key attributes between original and mapped images, typically like ExprGAN [9], StarGAN [10], and GANimation [2]. However, neither of them can be automatically applied on real-time video processing. That is, their design is mainly applied on single image with manual labeling on target expressions.

Thus, for designing an automatic facial expression migration architecture on real-time video processing, three problems should be solved. Basically, what is the facial expression represented in images? This problem requires an accurate facial expression classification to teach computers on recognizing emotions. Secondly, which facial expression we want to generate/change to? An automatic emotion labeling module should be designed, where we consider that the analysis on audio corresponding to target videos can provide help. It means that a speech recognition algorithm is needed for

978-1-7281-2009-6/$31.00 paper N-20002.pdf

processing speech-to-text transcribed documents. Lastly, how to generate required emotions on talking faces? Especially, two more sub-questions should be further answered: (i) how to add facial expressions without affecting other facial motions (*e.g.*, talking mouth)? (ii) how to generate facial expressions for non-real world characters that the model never learns? As there are several researches leveraging GANs to get better performances, we will design an enhanced model based on GANimation [11], as the third module to solve the above-mentioned problems. For better performance, the accessories on the face and the low illumination condition on images are also considered.

Correspondingly, we propose *GANemotion*, an automatic facial expression migration architecture based on GANs, for increasing the vitality of characters in videos. It is combined of three modules: facial expression classification module, audio emotion classification module, and facial expression migration model by GANs. The first module is the basis of this architecture, which outputs the facial expression type in each frame. The second module mainly realizes the automatic labeling by sentiment analysis. The third module accumulates the two outputs from former modules as its input to generate expressions we need for target videos. To avoid the effect on normal facial motions, we design the expression attention algorithm, which applies different coefficients for AUs affecting original facial motions. And we build a standard AU library for generations on non-real world characters. As illustrated in Fig. 1, compared with the original frame in the first column, *GANemotion* can successfully increase the vitality of news broadcaster by generating rich expressions, with realistic generations.

Evaluations further show the feasibility of *GANemotion*. The classification accuracy on the audio module has improved to 56.9% after category clustering. The accuracy for facial expression recognition is 78.5%, outperforming among single models, such as SVM, VGG-16, and ResNet50 [12], and other classification methods, including Shinohara et al. [13], Lyons et al. [14], and Feng [15]. The migration ability of *GANemotion* is also proved to be efficient, even can deal with the limitations of non-real world characters and low illumination conditions, avoiding the effect on original facial motions.

In a word, the contributions of this paper are as follow:

1) We design and implement an automatic facial expression migration architecture, for increasing the vitality of characters in real-time videos, without manual labeling.
2) To increase the migration ability, we propose an expression attention algorithm to eliminate distortion on facial motion, together with the standard AU library for non-real world characters common in entertained videos.
3) Rich work on training, fine-tuning and testing have proved the feasibility of our proposed architecture, which also indicates the potential of GANs applied in the entertainment area.

II. Related Work

Corresponding to three modules in *GANemotion*, a significant amount of researches have been made for each module, including facial expression classification, audio emotion recognition, and facial expression migration by GANs.

A. Facial Expression Classification

As a basis of facial expression migration architecture, processors have to know which expression representing in images. This module requires two main steps: feature extraction and classification, which derives three kinds of approaches, including optical flow based, fiducial points based, and model-based approaches.

Optical flow based approaches describe spatial-temporal facial actions by estimating the direction and magnitude associated with certain AUs [16], [17]. This kind of methods have consistent sensitivity for subtle facial recognition, but flow estimation tends to be noisy [18]. For the finer level of FACS AU, fiducial points based approaches quantify facial movements by measuring the geometrical displacement of facial feature points between two consecutive frames [19], [20]. After the segmentation of facial features (*i.e.*, eyes, eyebrows, mouth, etc.), each point is coded into AUs, which will be further compared against the descriptors in FACS for classification. These are approaches are more robust, but fail to cope with illumination changes and non-rigid facial motions. Further enhanced, model-based approaches define several metrics for extract features, like Gabor wavelet coefficients [15], Hidden Markov Model (HMM) [18], Active Appearance Model (AAM) [21], or more complicated neural networks [22]. The classification next is realized by the comparison of nearest mean vectors. For discrete emotion categories, these model-based approaches are satisfied enough by its higher classification accuracy.

B. Audio Emotion Classification

This module includes two main steps: (i) speech-to-text conversion; (ii) sentiment analysis in text. As for the first step, several speech documents like talks [23], voice mails [24], or broadcast news [25], should be transcribed for further speech recognition [26]. In such transcription, not only streams or words are contained, but also the structural information. It is not effective to simply transcribe all words, because of too much irrelevant information. To achieve better conversion result, generally, recent innovations in speech-to-text transcription system combines three modules: feature extraction front end, acoustic models, and language models [27]. As a typical solution, the IBM Watson Speech to Text service [28] provides better quality of services (QoS) on transcribing more than 10 languages speeches into text, after simple uploading of audio files. In this paper, we also adapt this service for our further sentiment analysis.

The second step requires a pre-trained language representation applied to downstream sentiment analysis, which can also be divided to two kinds: feature-based and fine-tuning methods. The feature-based approach uses specific tasks as

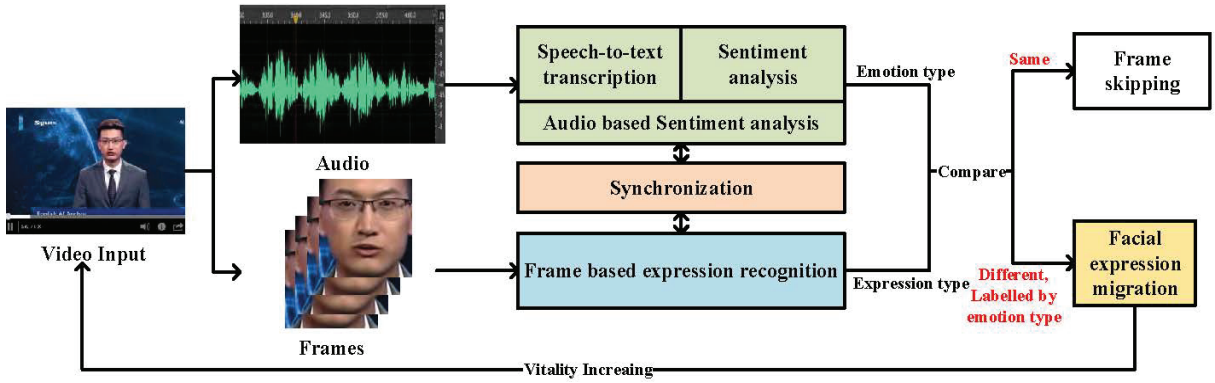

Fig. 2. The framework illustration of *GANemotion*.

additional features [29]. But it has less satisfactory than fine-tuning methods, which is directly trained on downstream tasks by fine-tuning the pre-trained parameters [30]. To solve the unidirectional limitation of fine-tuning method, BERT [31] is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context.

C. Facial Expression Generation by GANs

As a member of generative models, GANs are built on game theory for two parties: generator and discriminator. Generator tends to produce realistic fake samples, while discriminator tries to distinguish it as fake data. The final tradeoff of this game is called Nash Equilibrium, where the best migration result can be acquired [32]. Recent works [9]–[11] show it efficiency in facial expression migration, especially on solving less paired training data and low synthetic face resolution. StarGAN [10] achieves the general image-to-image translation applications, not only on facial expressions, but also on other attributes like hair color or genders. ExprGAN [9] designs an expression controller module to learn an expressive and compact expression code, which can further adjust expression intensity continuously. Considering that these designs can only change a discrete number of facial expressions, GANimation [11] realizes a novel GAN conditioning architecture based on AU annotations. It allows the training on multiple combination of AUs with a fully unsupervised manners. Among state-of-the-art researches, GANimation achieves the best migration result, which is also leveraged in our design.

III. Framework and Problem Formulation

To realize an automatic emotion-text migration architecture, we separate the input video into two parts: audio part and frame part. Synchronized by timestamps t_i (i $\in [1, n]$) in video V, corresponding audio A and frames F will be processed by two flows. Firstly, the audio A will be transcribed to text T_A and then analyzed by sentiment classification algorithm, which outputs the classified emotion type series $(e_{t_1}, e_{t_2}, ..., e_{t_n})$ followed by time sequence $(t_1, t_2, ..., t_n)$. Here, e_{t_i} belongs to the discrete emotion categories K, which is defined as 7 standard emotions: happy, sad, surprise, angry, fear, disgust, and neutral. For another flow, the facial expressions in frames F will be classified by facial expression classification module and acquire expression series $(Fe_{t_1}, Fe_{t_2}, ..., Fe_{t_n})$ followed by same time sequence. The comparison between emotion type in audio and corresponding expression type in frames at the same timestamp leads to the further decision: if these two types are same, then skip the frame; otherwise, this frame will be transferred to the next facial expression migration module, and the emotion type will become its target migration label.

Additionally, a necessary synchronization should be done between the above two modules. In synchronization, we should first set the sampling rate of videos, indicating the number of pictures sampled for facial expression classification module. Considering that people cannot change their expression too fast, we set the sampling rate as 1 frame per second. That is, in one second, there will be one image input into facial expression classification module, and the corresponding one-second audio will be transcribed to text for sentiment analysis. However, as the emotions cannot be analyzed from such a short time in speech, the synchronization module should next consider the changing timestamp of emotions in speech. As emotions can only be detected by complete sentences, such changing timestamp exists in one certain gap between two sentences, which will be given by audio emotion classification model. The synchronization module will limit the output of the audio module as one emotion type at one time, and record the corresponding timestamp to calculate the time durations. This time duration indicates the number of the images should be output and further generated in the next module, when our pre-set sampling rate is 1 fps.

In this migration module, each input frame (RGB image) will be defined as $\boldsymbol{I_{y_j}} \in \mathbb{R}^{h \times w \times 3}$ (h and w is the height and width of the frame), and the facial expressions are encoded by a set of N AUs $\boldsymbol{y_j} = (y_1, y_2, ..., y_n)^T$ ($y_j \in [0, 1]$). The aim of GANs in this module is to learn a mapping M to translate $\boldsymbol{I_{y_j}}$ to an output image $\boldsymbol{I_{y_e}}$ labelled by target AU $\boldsymbol{y_e}$ on another input emotion type e_{t_i}. So this mapping can be represented as $M : (\boldsymbol{I_{y_j}}, \boldsymbol{y_e}) \rightarrow \boldsymbol{I_{y_e}}$. In *GANemotion*, we adapt the generative processing in GANimation [11], where we train M in a fully unsupervised manner with training triples $\{\boldsymbol{I^m_{y_j}}, \boldsymbol{y^m_j}, \boldsymbol{y^m_e}\}^{\mathrm{M}}_{m=1}$. Note that the target AU vectors $\boldsymbol{y_e}$ here are acquired from standard AU libraries and extracted

by input emotion type, rather than random migration, which mainly realizes the automatic labeling in *GANemotion*. The generated emotion picture will be applied back to original video to increase its vitality. The whole framework is shown in Fig. 2.

IV. Design of *GANemotion*

In this section, we will go for the detail about the design of *GANemotion*, especially for three main modules colored in Fig. 2, that is, audio emotion classification module, facial expression classification module, and facial expression migration module. The outputs of former two modules are the inputs of the third module, so the goal of GAN in the third module is to generate a picture from original expression in frames to emotion type automatically acquired from text analysis.

A. Facial Expression Classification

Accurate classification is the basis of the further modules, which gives the recognition ability for *GANemotion* on facial expressions. Considering that a diverse set of models can derive more accurate analysis than a single model, we use ensemble learning [33] for this classification. To avoid error accumulation from the first model, we take Boosting as the main technique of ensemble learning, which follows sequential steps as below:

1) Create a subset from original dataset, and initial all data points to equal weights;
2) Create a base model on this subset to make predictions on the whole dataset;
3) Incorrectly predicted values will be given higher weights, and another model is next created to predict the whole dataset;
4) The rest can be done in the same manner, until the final model is constructed by the weighted mean of all created models.

In our module, we attempt to combine three state-of-the-art deep learning models: ResNet-20 [12], ResNet-50 [12], and VGG16 [12]. Inspired by this Boosting technique, the first model will extract features by ResNet-20 pre-trained by CIFAR10 dataset [34] and get classification probabilities by SVM classifier, which replaces the final fully connected layer in ResNet by softmax function of SVM. Similarly, the next two modules use ResNet-50 and VGG16 pre-trained both on ImageNet [35] for feature extraction and XGBoost Classifier [36] for classification. All output probabilities of these three models will be transferred as three input features for training the final logistic regression model, as shown in Fig. 3. We test this final model on test dataset, and then take the output with the highest score as the final prediction. Here, the logistic regression is implemented by sklearn library [37].

B. Audio Emotion Classification

This module is seamlessly connected by two sub-modules: speech-to-text transcription and sentiment analysis. It aims to collect the indicated emotions from word representations, which will further become the label of target migration. Taking

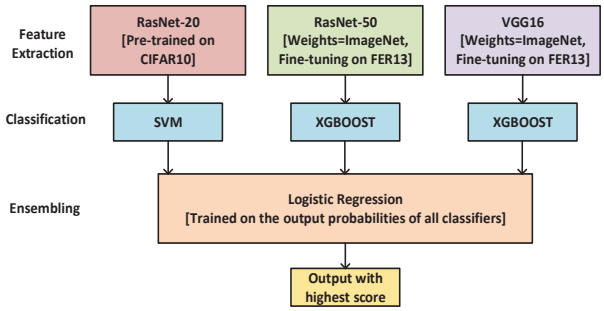

Fig. 3. The framework illustration of facial expression classification module in *GANemotion*.

Anchor at the beginning as an example, we analyze the news and comments he gives to imply any emotions, and then generate corresponding facial expressions to increase his vitality when broadcasting news.

1) Speech-to-text Transcription: As mentioned in Section. II-B, general transcription model has three components: feature extraction front end, acoustic models, and language models. In *GANemotion*, we leverage the IBM Watson Speech to Text API [28] for our audio transcription. Its three components are designed as [38]:

1) **Feature extraction front end:** The features for representing the acoustic signal in audio are 40-dimensional vectors collected from linear discriminant analysis (LDA) projection. It will extract those features at a rate of 100 frames/second from a Hamming windowed speech segmentation of 25ms durations. The segmentation here is obtained by HMM-based segmentation system with a five-state, left-to-right sequence with no skip state. The K-means clustering algorithm is further applied for speaker adaptation.
2) **Acoustic model:** The speaker-independent acoustic model in Watson is trained on the aforementioned features normalized on a per-speaker basis. It leverages continuous density, left-to-right HMMs, and uniform transition probabilities.
3) **Language model:** Three separated four-gram models are constructed with modified Kneser-Ney smoothing [39], including meeting model, scientific proceeding model, and Fisher's video model.

2) Sentiment Analysis: Accurate emotion type analysis is required for further better migration. Latest BERT [31] shows its higher efficiency, especially for solving the unidirectional constraints by its pre-trained manners. This sub-module is also a multi-layer bidirectional transformer encoder based on BERT [31], and we then fine tune it for the compatible of video emotion analysis. The construction of this sub-module followed by pre-training and fine tuning procedures.

The pre-training step leverages the pre-trained $\mathrm{BERT}_{\mathrm{BASE}}$ model, with 12 layers, 768 hidden size, 12 self-attention heads, and 110M parameters in total. It is pre-trained on BooksCorpus (800M words) [40] and English Wikipedia (2,500M words) [41]. The detailed pre-trained procedures can be referred to Jacob et al. [31].

TABLE I
THE CORRESPONDENCE OF COMBINATION ON EMOTION CATEGORIES

Original categories	Combined categories	Standard AUs
happiness	happy	6+12
relief		
fun		
love		
anger	angry	4+5+7+23
hate		
surprise	surprise	1+2+25+26
enthusiasm		
neutral	neutral	R12A+R14A
empty		
sadness	sad	1+4+15
boredom	disgust	9+15+16
worry	fear	1+2+4+5+7+20+26

In the fine-tuning step, novel parameters for fine-tuning are added on classification layer $\mathrm{W} \in \mathbb{R}^{K \times H}$, where H is the hidden size, and K is the number of classifier labels, also mentioned as 7 emotion types in Section. III. The input dataset for this step is Kaggle [42], which has 30K tweets and 13 categories. In pre-processing of this dataset, links and At someone (label "@") are replaced by simple notations, and 13 categories are combined to 7 standard types for increasing the classification accuracy. The correspondence in combination is shown in Table. I. The pre-processed data will be input as $C \in \mathbb{R}^H$, and a standard softmax function $P = \mathrm{softmax}(\mathrm{CW}^{\mathrm{T}})$ calculates the label probabilities $P \in \mathbb{R}^K$.

C. Facial Expression Generation by GANs

This module is designed based on GANimation [11], which is the latest bidirectional migration architecture. A generator $G(\boldsymbol{I_{y_j}}|\boldsymbol{y_e})$ is trained realistically to transform the facial expression in $\boldsymbol{I_{y_j}}$ to the emotion type $\boldsymbol{y_e}$. At the meantime, the discriminator $D(\boldsymbol{I_{y_e}})$ tends to classify the generated picture as fake, according to the probability of the overlapping patches to be real. In video migration, there are two challenges should be solved: Firstly, how to avoid irrelevant face section (*e.g.*, a talking mouth) on the migration of emotions? Secondly, for non-real world data distributions which are not previously known by the model, how to accurately generate corresponding expressions? The following designs specially deal with these two challenges.

1) Expression Attention: Inspired by the attention mask idea in GANimation [11], in *GANemotion*, we assign different weights between expression-related AUs in mouth range and other AUs. The algorithm to get the generator condition $\boldsymbol{y_c}$ is shown in Algorithm.1. The parameter α_1, α_2 represents different degrees of activation on target AUs involved in expressions. In this algorithm, we only consider the effect on mouth motion, which we believe it can be extended to general scenarios, like hair swing or eyes blinking. If the traversed AU is associated with mouth motion, lower weight will be multiplied in front of this AU values to decrease its effect on the mouth, and vice versa. After returning, the generator will migrate expressions from $\boldsymbol{I_{y_j}}$ to $\boldsymbol{I_{y_e}}$, with condition of $\boldsymbol{y_c}$. The actual process of GANs in this module can be represented in Fig. 4. This method is robust to light conditions and non-real world data, where the $\boldsymbol{y_e}$ is consisted by standard AU expressions. The next subsection will explain this part in detail. $\boldsymbol{I_{y_e}} = G(\boldsymbol{I_{y_j}}|\boldsymbol{y_c})$

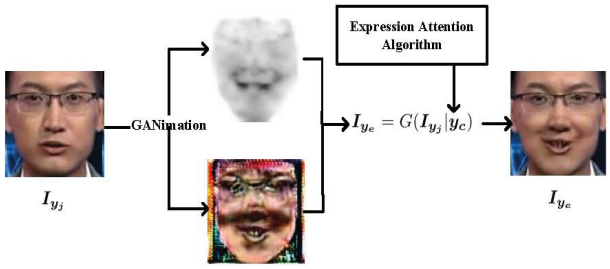

Fig. 4. Facial Expression Generation by GANs in *GANemotion*

Algorithm 1: Expression Attention Algorithm

Input: α_1,α_2,$\boldsymbol{y_j}$,$\boldsymbol{y_e}$
Output: Generator condition $\boldsymbol{y_c}$

$\boldsymbol{y_c} \leftarrow \boldsymbol{y_j}$;
for $\mathrm{AU_i}$ *in* ALL_AVAILABLE_AU **do**
  if $\mathrm{AU_i}$ *in* $\boldsymbol{y_e}$ **then**
    if $\mathrm{AU_i}$ *is in mouth range* **then**
      $\boldsymbol{y_c}[i] = (1-\alpha_1) \times \boldsymbol{y_c}[i] + \alpha_1 \times \boldsymbol{y_e}[i]$;
    else
      $\boldsymbol{y_c}[i] = (1-\alpha_2) \times \boldsymbol{y_c}[i] + \alpha_2 \times \boldsymbol{y_e}[i]$;
    end
  end
end
return $\boldsymbol{y_c}$;

2) Edition on Standard AUs: For non-real world data distributions, where the model is lacking in pre-knowledge, it is difficult to find suitable AUs based on emotion requests. Again taking the Anchor as an example, all videos collected are represented by neutral facial expressions because of its "indifference". We cannot find corresponding happy or sad AU representations for its videos. To solve this problem, we pre-define standard AUs for 7 emotion types, which are analyzed as more than 60% probability to show in general facial expression processing [43]. Summarized in Table. the third column of I, this standard library can help to generate accurate facial expressions for virtual characters.

V. IMPLEMENTATION DETAILS

As *GANemotion* is composed by three main modules, where each of them contains a deep neural network, they are trained separately.

Training on audio emotion classification module. We leverage the pre-trained BERT model and fine-tuning it with Kaggle dataset, which has 30K tweets divided to 28K training set and 2K testing set. The categories are summarized from 13 to 7. We use Adam optimizer [44] with several necessary parameters: $b_1 = 0.9$, $b_2 = 0.999$, error $= 1\mathrm{e}^{-6}$, and weight decay$= 0.01$.

TABLE II
EVALUATION AND COMPARISON BETWEEN OUR CLASSIFICATION MODULE WITH THREE OTHER BENCHMARKS.

Reference	Method	Problem	Accuracy
Shinohara et al. [13]	HLAC+ Fisher weight mapping	Manual face cropping	69.4%
Lyons et al. [14]	Gabor wavelet+ PCA+ LDA	Manual fidual points selection	75%
Feng [15]	LBP+ Coarse-to-Fine classification	Face cropping by position of pupils	77%
GANemotion	Ensemble learning	Complicated deep learning structure	**78.5%**

Training on facial expression classification module. The stacked model in this module is pre-trained on Cifar10 dataset [34] with fine-tuning on FER2013 dataset [45]. In Cifar10 dataset, the total 60K 32×32 colour images in 10 classes are grouped by 50K training set and 10K testing set. In FER2013 dataset, the total 35K datapoints will be separated to 28K training set and 7717 points used for testing. Similar to the above module, we still use Adam optimizer, but with 0.0005 learning rate, 200 epochs, and 32 batch size.

Training on facial expression migration module. Our GANs model is modified based on GANimation, where all parameters are kept the same, except the evaluated α. The model is trained on EmotioNet dataset [43], with 191,160 in 200K for training and the rest for testing. The batch size is set as 25, and the epoch is 30. The AUs in this module can be obtained by OpenFace API [46].

For the final test, three modules will structurally connect and take one input image (one frame in the video) with one audio at the same timestamp as inputs. The audio will firstly pass the speech-to-text transcription on the IBM Watson Python SDK [47]. And the resulting texts together with frames will be parallelled put into two classification modules. The final generated frame will be applied back to the original video depending on its timestamp. All of these modules are coded in Python and running on Ubuntu 16.04.5 LTS with single NVIDIA GTX 1080 Ti GPU.

VI. PERFORMANCE EVALUATION

This section will provide a comprehensive evaluation of *GANemotion*. As three main modules have different designs, their performances are respectively shown in each following subsections. For evaluation on facial expression classification, we show the classification accuracy compared with three benchmarks: Shinohara et al. [13], Lyons et al. [14], and Feng [15]. The audio emotion classification module is evaluated also on its accuracy. Furthermore, for the migration module, we globally test its behavior on two extreme conditions: (i) non-real world characters; (ii) low illumination condition, and shows its migration results by figures. Note that for GANs, the most obvious performance can be seen from the quality of generated images, as long as they have higher resolution. So most of the results we present in this section is shown by images rather than evaluation indexes.

A. Evaluation on Facial Expression Classification Module

The evaluation on facial expression classification module is compared with three benchmarks:

1) Shinohara et al. [13]: This paper designs a Higher-order Local Auto-Correlation (HLAC) features with Fisher weight maps for classification. But it requires a manual face cropping.
2) Lyons et al. [14]: They propose the combination between Wavelet+PCA+LDA, where fiducial points are manually selected.
3) Feng [15]: This paper uses Local Binary Pattern (LBP) features with coarse-to-fine classification. It crops faces by the position of pupils.

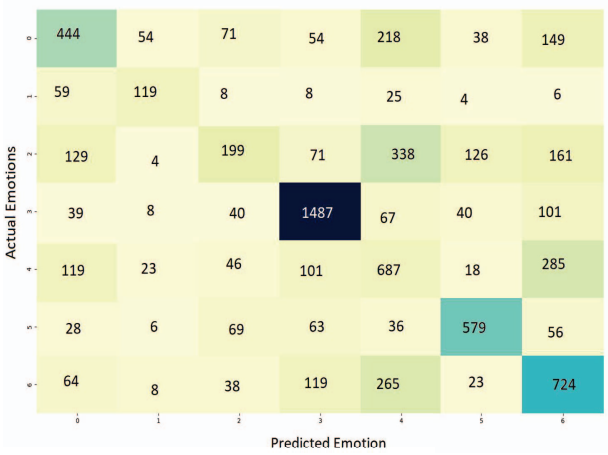

Fig. 5. Confusion matrix for facial expression classification module.

TABLE III
COMPARISONS BETWEEN ENSEMBLE LEARNING WITH SINGLE MODELS

Model	Accuracy	Precision	Recall
ResNet20+SVM	57%	54.8%	55.4%
VGG-16+XGBoost	59%	70.2%	69.8%
ResNet50+XGBoost	67%	79.2%	74.9%
Logistic Regression (Ensembled)	**78.5%**	**84.4%**	**83.2%**

Our module can automatically extract features by neural networks, and give better accuracy by ensemble learning. The compared results are shown in Table. II. Additionally, we also compare the classification ability with the composed single model summarized in Table. III. Our module outperforms in all these two comparisons, implying its ability for further classification. The corresponding confusion matrix is also provided in Fig. 5.

B. Evaluation on Audio Emotion Classification Module

This subsection shows the result of pre-trained BERT_Base model fine-tuning results by Kaggle dataset, with 13 original categories and 7 combined categories. The classification accuracy on 13 categories is 37.2%, while the result on 7 classes is 56.9%. It is obvious that the combined categories are beneficial

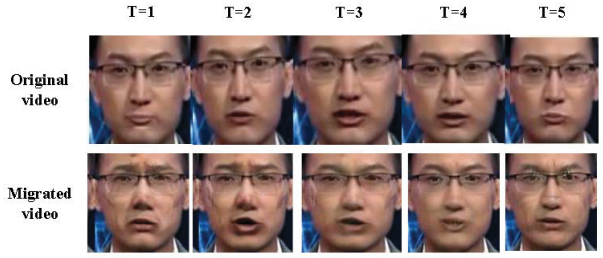

Fig. 6. Evaluation results for *GANemotion* on consecutive frames

for accuracy enhancement, which have **around 20% progress**. It is also worth noting that the fine-tuning results on 11 NLP tasks proposed by its original paper [31], including MNLI, QQP, QNLI, SST-2, CoLA, STS-B, MRPC, RTE, WNLI, are all above 70%, with 81.9% average accuracy. But the sentiment task in *GANemotion* is 7-class emotion detection based on real-time audio, whose scenario is more complicated than these 11 tasks. The hidden meanings behind of sentences are complicated. Considering about a funny example, a girl possibly will express her "happiness" to her boyfriend even she is angry when they are in a quarrel.

C. Evaluation on Migration Module

As shown in Fig. 1, migration module based on GANs can successfully add more vivid facial expressions for original pictures. Two migration results are made to check the effect of α in Algorithm.1. Here, $\alpha = \alpha_1 + \alpha_2$. It proves that if α is high, the target AUs will have more effect on original pictures to generate facial expressions on a large extent, and vice versa. It is useful for customizing the requirements on facial expressions and corresponding scenarios. For example, cartoon producers will require higher α for exaggerating characters, while news reporter should be more conservative with lower α settings. Additionally, from the result of Fig. 1, accessories on faces like glasses will not disturb the final migration result. Further proved by Fig. 6, we apply *GANemotion* into a shortcut of videos for Anchor. The migrated facial expressions are corresponding to the content of his speech, which can successfully increase his vitality on news broadcasting.

In addition to Fig. 1 and Fig. 6, Fig. 7 proves the feasibility of *GANemotion* on processing virtual characters. As their data distributions are not known by GANs before, we use the standard AU library for further processing, where their AUs can be extracted by OpenFace API [46]. The migration results follow the same rules on different effects of α.

VII. Conclusion

The intuitive way for vitality enhancement is the addition of facial expression, to change an indifferent face to more vivid expressions. Although recent researches have been solved this problem based on GANs, they cannot automatically acquire the labeling for target generations. In this paper, we propose *GANemotion*, to build a more convenient emotion migration model on GANs. The audio emotion classification provides

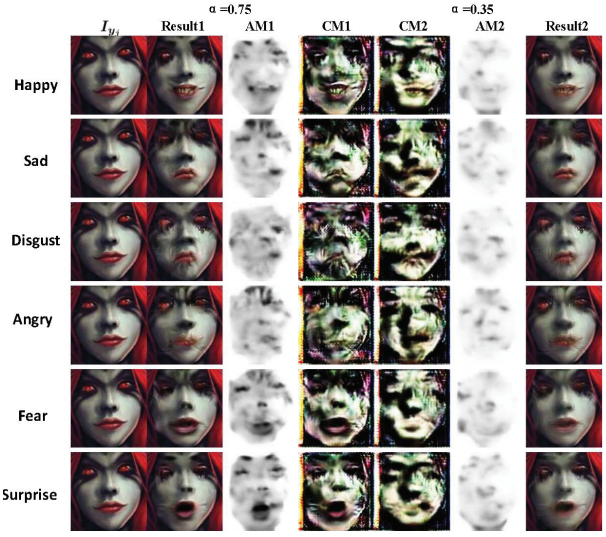

Fig. 7. Evaluation results for *GANemotion* on virtual character [AM-Attention Mask, CM-Color Mask, α represents the degree of AU effects.]

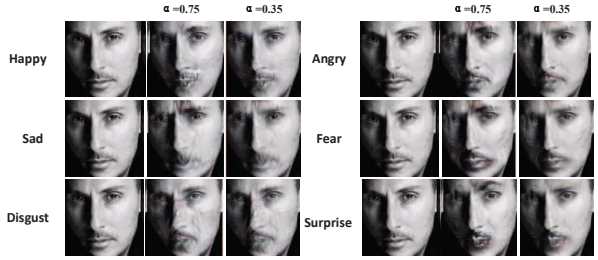

Fig. 8. Evaluation results for *GANemotion* on consecutive frames

emotion type as migration labels. With the basis of facial expression classification, the migration model by GANs can overcome the limitations on non-real world characters, and low illumination conditions, with guaranteeing the normal mouth behaviour. Rich training, fine-tuning, and customized design in *GANemotion* make it as an effective migration architecture. Evaluation results have shown its power on vitality enhancement, breaking the limits of non-real world characters and low illumination conditions.

For further enhancement of our work, we propose two possible future directions for research improvement. The first one is to apply face pose adjustment for standardizing the input images. As we leverage standard AUs with fixed positions, if the face in an image is in a side pose, the GANs cannot generate correct results. To this end, we tend to add standardize step on images before putting into GANs module. A lot of methods [48], [49] have been proposed for this face pose adjustment, which can help us in this direction.

The second one is the improvement on AU libraries used in the migration module. As shown in Fig. 8, the migration result reveals visible artifacts because of the leverage of standard AU libraries. We tend to research on AU prediction corresponding to different characters to realize more smooth and natural facial expression migration.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China 2018YFB1004703, in part by China NSF grant 61672349, 61672353.

REFERENCES

[1] L. Handley, The 'world's first' A.I. news anchor has gone live in China, Website, https://www.cnbc.com/2018/11/09/the-worlds-first-ai-news-anchor-has-gone-live-in-china.html (2018).

[2] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, F. Moreno-Noguer, Ganimation: Anatomically-aware facial animation from a single image, in: ECCV, Munich, Germany, 2018, pp. 835–851.

[3] F. W. Ekman P., Facial action coding system: A technique for the measurement of facial movement., Consulting Psychologists Press.

[4] C. L. Huang, Y. M. Huang, Facial expression recognition using model-based feature extraction and action parameters classification, Journal of Visual Communication & Image Representation 8 (3) (1997) 278–290.

[5] D. Shichuan, T. Yong, A. M. Martinez, Compound facial expressions of emotion, Proc Natl Acad Sci U S A 111 (15) (2014) E1454.

[6] M. A. Fischler, R. A. Elschlager, The representation and matching of pictorial structures, IEEE Trans Computers C 22 (1) (1973) 67–92.

[7] Z. Liu, S. Ying, Z. Zhang, Expressive expression mapping with ratio images, in: Conference on Computer Graphics & Interactive Techniques, 2016.

[8] Y. Fei, J. Wang, E. Shechtman, L. D. Bourdev, D. N. Metaxas, Expression flow for 3d-aware face component transfer, Acm Transactions on Graphics 30 (4) (2011) 1–10.

[9] H. Ding, K. Sricharan, R. Chellappa, Exprgan: Facial expression editing with controllable expression intensity, in: AAAI, New Orleans, Louisiana, USA.

[10] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: IEEE CVPR, Salt Lake City, UT, USA, 2018, pp. 8789–8797.

[11] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, F. Moreno-Noguer, Ganimation: Anatomically-aware facial animation from a single image, in: ECCV, Munich, Germany, 2018.

[12] Keras, https://github.com/fchollet/keras, 2015.

[13] Y. Shinohara, N. Otsu, Facial expression recognition using fisher weight maps, in: IEEE FGR, Seoul, Korea, 2004, pp. 499–504.

[14] M. J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999) 1357–1362.

[15] X. Feng, Facial expression recognition based on local binary patterns and coarse-to-fine classification, in: 2004 International Conference on Computer and Information Technology (CIT 2004), 14-16 September 2004, Wuhan, China, 2004, pp. 178–183.

[16] I. A. Essa, A. Pentland, A vision system for observing and extracting facial action parameters, in: CVPR, Seattle, WA, USA, 1994, pp. 76–83.

[17] Y. Yacoob, L. S. Davis, Recognizing Human Facial Expressions From Long Image Sequences Using Optical Flow, 1996.

[18] J. J. Lien, T. Kanade, J. F. Cohn, C. Li, Detection, tracking, and classification of action units in facial expression, Robotics and Autonomous Systems 31 (3) (2000) 131–146.

[19] C. Huang, Y. Huang, Facial expression recognition using model-based feature extraction and action parameters classification, J. Visual Communication and Image Representation 8 (3) (1997) 278–290.

[20] Z. Hammal, L. Couvreur, A. Caplier, M. Rombaut, Facial expression classification: An approach based on the fusion of facial deformations using the transferable belief model, Int. J. Approx. Reasoning 46 (3) (2007) 542–567.

[21] Y. Saatci, C. Town, Cascaded classification of gender and facial expression using active appearance models, in: IEEE FGR, Southampton, UK, 2006, pp. 393–400.

[22] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: ICML, Sydney, NSW, Australia, 2017, pp. 2642–2651.

[23] K. Zechner, A. Waibel, Minimizing word error rate in textual summaries of spoken language, in: ANLP, Seattle, Washington, USA, 2000, pp. 186–193.

[24] K. Koumpis, S. Renals, Transcription and summarization of voicemail speech, in: Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000, 2000, pp. 688–691.

[25] C. Hori, S. Furui, R. Malkin, H. Yu, A. Waibel, Automatic summarization of english broadcast news speech, in: Proc Human Language Technology Workshop, 2002.

[26] S. Furui, T. Kikuchi, Y. Shinnaka, C. Hori, Speech-to-text and speech-to-speech summarization of spontaneous speech, IEEE Trans. Speech and Audio Processing 12 (4) (2004) 401–408.

[27] A. Stolcke, B. Y. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, M. K. Sönmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, Q. Zhu, Recent innovations in speech-to-text transcription at SRI-ICSI-UW, IEEE Trans. Audio, Speech & Language Processing 14 (5) (2006) 1729–1744.

[28] The ibm watson speech to text service, https://speech-to-text-demo.ng.bluemix.net/.

[29] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: NAACL-HLT, New Orleans, Louisiana, USA, 2018, pp. 2227–2237.

[30] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: ACL, Melbourne, Australia, 2018, pp. 328–339.

[31] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805.

[32] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in: NeurIPS, Montreal, Quebec, Canada, 2014, pp. 2672–2680.

[33] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, V. Palade, Deep learning for emotion recognition in faces, in: ICANN, Barcelona, Spain, pp. 38–46.

[34] Cifar-10 dataset, https://www.cs.toronto.edu/~kriz/cifar.html.

[35] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: A large-scale hierarchical image database, in: IEEE CVPR, Miami, Florida, USA, 2009, pp. 248–255.

[36] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: ACM SIGKDD, San Francisco, CA, USA, 2016, pp. 785–794.

[37] F. Pedregosa, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (10) (2013) 2825–2830.

[38] J. Huang, M. Westphal, S. F. Chen, O. Siohan, D. Povey, V. Libal, A. Soneiro, H. Schulz, T. Ross, G. Potamianos, The IBM rich transcription spring 2006 speech-to-text system for lecture meetings, in: MLMI, Bethesda, MD, USA, 2006, pp. 432–443.

[39] V. Siivola, T. Hirsimaki, S. Virpioja, On growing and pruning kneser-ney smoothed -gram models, IEEE Transactions on Audio Speech & Language Processing 15 (5) (2007) 1617–1624.

[40] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: IEEE ICCV, Santiago, Chile, 2015, pp. 19–27.

[41] A. Ghaddar, P. Langlais, Wikicoref: An english coreference-annotated corpus of wikipedia articles, in: LREC, Portorož, Slovenia, 2016.

[42] Kaggle dataset, https://www.kaggle.com/c/sa-emotions/leaderboard.

[43] C. F. Benitez-Quiroz, R. Srinivasan, A. M. Martínez, Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, in: IEEE CVPR, Las Vegas, NV, USA, 2016, pp. 5562–5570.

[44] D. Kingma, J. Ba, Adam: A method for stochastic optimization, Computer Science.

[45] Fer2013 dataset, https://www.kaggle.com/c.

[46] Openface api, https://github.com/TadasBaltrusaitis/OpenFace.

[47] Ibm watson python sdk, https://github.com/watson-developer-cloud/python-sdk.

[48] W. Cao, Study of an algorithm for face pose adjustment based on eye location.

[49] W. Cao, S. Wang, An algorithm for face pose adjustment based on gray-scale static image.