Temporal conditional Wasserstein GANs for audio-visual affect-related ties

Temporal conditional Wasserstein GANs for audio-visual affect-related ties

Christos Athanasiadis, Enrique Hortal and Stelios Asteriadis
Department of Data Science and Knowledge Engineering
Maastricht University
Maastricht, the Netherlands
enrique.hortal@maastrichtuniversity.nl

***Abstract*—Emotion recognition through audio is a rather challenging task that entails proper feature extraction and classification. Meanwhile, state-of-the-art classification strategies are usually based on deep learning architectures. Training complex deep learning networks normally requires very large audio-visual corpora with available emotion annotations. However, such availability is not always guaranteed since harvesting and annotating such datasets is a time-consuming task. In this work, temporal conditional Wasserstein Generative Adversarial Networks (tc-wGANs) are introduced to generate robust audio data by leveraging information from a face modality. Having as input temporal facial features extracted using a dynamic deep learning architecture (based on 3dCNN, LSTM and Transformer networks) and, additionally, conditional information related to annotations, our system manages to generate realistic spectrograms that represent audio clips corresponding to specific emotional context. As proof of their validity, apart from three quality metrics (Fréchet Inception Distance, Inception Score and Structural Similarity index), we verified the generated samples applying an audio-based emotion recognition schema. When the generated samples are fused with the initial real ones, an improvement between 3.5 to 5.5% was achieved in audio emotion recognition performance for two state-of-the-art datasets.**

***Index Terms*—Domain Adaptation, Audio Emotion Recognition, Generative Adversarial Networks, Attention Mechanisms.**

I. Introduction

Recent progress in deep learning has fuelled advances in automated emotion recognition in the broader domain of affective computing. These advances came along with the prerequisite for the availability of large training datasets. On this ground, several big corpora have been introduced especially with respect to facial emotion recognition. However, for other modalities such as audio or brain signals, the generation and proper annotation of these types of huge datasets is not a forthright task [1]. One of the main constraints is associated with ambiguity when annotating affect-related instances.

On the other hand, there are many studies from the cognitive psychology perspective proposing connections between audio and facial cues. In particular, some studies [2] [3] have identified high-level brain structures that connect these cues and find the way the human brain commonly perceives these modalities. Similar research was also performed in an emotional context [4] [5].

Inspired by these works, we propose an AI model unlocking the relations between face and audio in emotion expressivity, with the end goal of generating data coming from one of the two cues (here, audio), when only the other one is available (here, face). On this basis, domain adaptation approaches [6] [7] are adopted in an effort to model the affective relationships that govern the two modalities. The proposed methodology aims at learning a projection which maps characteristics from a modality (face) to another (audio). Having learned this function, it can be used to generate new samples in the audio domain. These new audio samples can subsequently be fused with the actual recorded ones, to expand and enrich the initial datasets at hand. Hence, Audio Emotion Recognition (AER) could be applied as the evaluation schema of the proposed approach by comparing the classification performance when using the initial dataset and the dataset enriched with generated samples and, in this manner, improve performance of emotion classifiers.

In order to learn audio-visual mappings, Generative Adversarial Networks (GANs) [8] [9] can be employed. GANs is a state-of-the-art approach that generates data samples that approximate a target distribution. This architecture typically consists of two networks: the Generator G and the Discriminator D. Having as input a noise vector $z \sim P(z)$, the task of G is to generate samples from the target domain (in our case audio). On the other hand, the task of D is to discriminate between samples that are derived from the real dataset and instances that are generated by G. Both networks are trained in a min-max manner and the goal of the architecture is to converge to a point where G will generate samples that approximate as much as possible the authentic ones.

Starting from the aforementioned vanilla architecture [8], we focus on altering the initial schema to meet the demands of our task. The input to the generator G is the face modality, denoted as source cue X_S. Additionally, since we want to study time-related relations, temporal features extracted from this modality are given as input to network G as well. These are extracted using three state-of-the-art topologies, 3dCNN [10], LSTM [11], and the Transformer architecture [12], which is based on attention mechanisms. With respect to the audio domain, we make use of spectrogram representations, taking advantage of their strong encoding power of voice prosody.

This work builds on previous findings presented in [13], where static image-to-image knowledge transfer was applied. In the current work, temporal features are considered through

pre-trained deep topologies, while the added value of the recently introduces attention mechanisms, mainly Transformers, are leveraged. Moreover, a more dedicated approach for the loss function of generating topology (namely Wasserstein GANs [14]) [15]) is employed to improve the performance of the conducted knowledge transfer and alleviate time-related constraints. Wasserstein GANs can converge steadier, improve the quality of visual results and find a better equilibrium between the performance of the generator and the discriminator.

The structure of the remainder of this paper is as follows: Section II provides a review of the state-of-the-art while Section III describes the introduced methodology. In sections IV and V, the experimental protocol and the results are analyzed, respectively. Finally, Section VI contains the conclusions and possible future directions.

II. Related work

During the last years, generative architectures serving a number of purposes in data generation and classification have been widely used in different domains. Probably the most influential model is that referred to as Generative Adversarial Networks or GANs [8], briefly explained in the introduction of this paper. A significant number of works have been proposed making use of GANs for audiovisual analysis, with most prominent being those relating to synthetic face generation [16] or animation [18]. For example, the authors in [18] propose speech-driven facial animation, given an image and an audio clip using temporal GANs. Despite the high popularity of GANs in audiovisual problems, the potential of generative architectures in the emotion domain has not been fully studied yet, especially for generating audio or visual signals conditioned on emotion.

In terms of training stability of GANs, authors in [14] proposed a modified version (called Wasserstein Generative Adversarial Networks or wGANs) with the aim of defining a more stabilized version concerning optimization convergence by using Wasserstein distance as a loss function. An extension of that work is proposed in [15] where improvements were introduced for the wGANs. Whereas, the work Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities [19] introduced a novel loss based on Lipschitz regularization. In the proposed work, we make use of these advances in training optimization of GANs for the demanding and noisy problem of knowledge transfer among modalities in emotionally rich contexts.

A. Audio-visual cross-modal studies

With respect to cross-modal studies, authors in [20] introduced conditional GANs, implemented in an effort to generate audio-visual content in musical performances. For that purpose, they adopted two separate networks (image to audio and audio to image) in order to perform cross-modal generations in both ways.

The work done in [21] introduced a novel transfer learning scenario, which distils robust phonetic features from grounding models trained to predict correlations between image and audio clips. Semantics of speech are largely determined by its lexical content, represented by models that learn to preserve phonetic information while filtering uncorrelated information, such as speaker and channel.

Authors in [22] proposed audio-video synchronization between mouth and speech. To facilitate the task, a two-stream network was implemented by having one network dedicated to audio and one to video, coupled together by using contrastive loss judging whether or not the embeddings of the two streams belong to a synchronized video pair or not. Similarly, in [23], an audio-visual study was performed with the purpose of performing temporal synchronization. Likewise in [22], a two-stream network using constructive loss was implemented. In contrast with that work, the negative pairs were chosen to be within the same videos. Furthermore, authors employed 3D convolutional neural networks (3dCNN) with the purpose of learning spatio-temporal features that can model the correlation between the face and audio modalities.

B. Temporal encoding from video

Authors in [24] proposed a deep architecture to learn spatiotemporal features for gesture recognition based on 3dCNN and bidirectional convolutional long-short-term memory networks (ConvLSTM). Authors in [25] presented a new video representation, called temporal linear encoding (TLE) which was deployed within a CNN architecture as a new layer, capturing the appearance and motion throughout entire videos.

With respect to attention mechanisms, authors in [26] utilized a self-attention mechanism to learn the alignment between text and audio for emotion recognition in speech. Authors in [27] proposed recursive multi-attention based on Memory Networks. Their cross-modal approach showed that gated memory effectively achieves multimodal emotion recognition. In [28], authors deployed the Action Transformer model for localizing and recognizing actions in video clips. They re-purposed a Transformer-style architecture to aggregate features from the spatio-temporal context around the person whose actions are classified. They showed that by using high-resolution, person-specific, class-agnostic queries, the model spontaneously learns to track individual people and to pick up on semantic context from the actions of others.

III. Methodology

In order to validate our research contributions, two main research questions are posed. The first one concerns *the efficiency of Wasserstein loss in comparison with the conventional GANs in terms of knowledge transfer*. The second posed question is associated with studying the temporal tie between face and audio modalities within emotional expressivity contexts. More specifically, we want to *assess whether modeling the temporal dynamics that govern the audio-visual relationship can also help improve the performance of knowledge transfer in generating new emotion-enriched audio samples*.

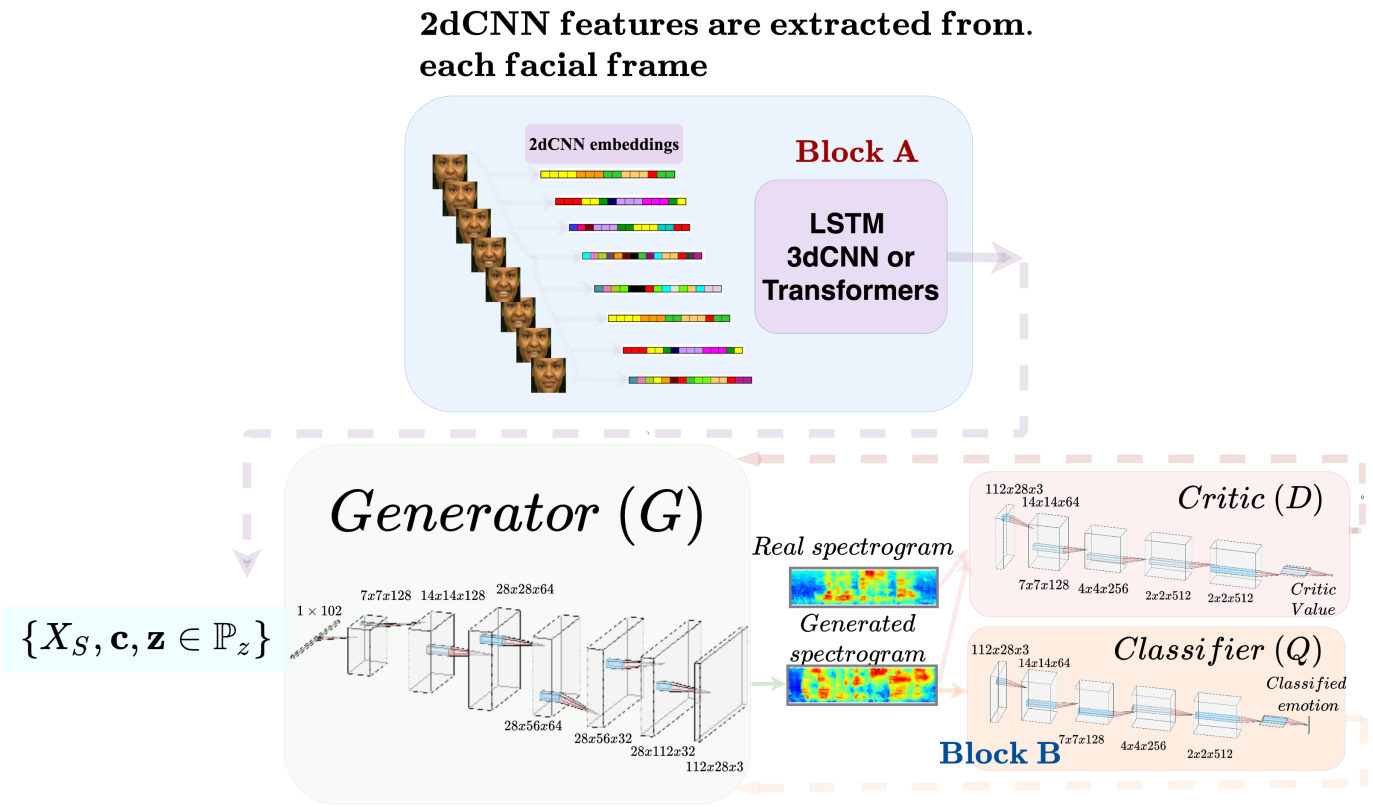


Fig. 1: The tc-wGANs approach. Block A contains the information regarding the temporal feature extraction process (LSTM, 3dCNN and Transformers versions). In Block B, the wGANs architecture with the size of each tensor for G, D and Q networks (inputs, outputs) can be seen.

The overall architecture of the approach can be seen in Figure 1. The activation function used for each of the layers in our approach is Leaky REctified LinearUnit (LeakyRELU). As an exception, in the output layer, to normalize the output to be between [-1,1], the hyperbolic tangent function was applied. Furthermore, batch normalization and dropout operations are performed in every layer for the three networks (except for the output layer). The rest of the information (regarding the tensors' size in each layer) can be also found in Figure 1.

Our objective is to implement a domain shift and calculate a transformation between source (D_S, representing face modality) and target domain (D_T, audio modality), denoted as $X_S \rightarrow X_T$. Thereby, instead of having as input to the generator merely a noise vector $z \sim P_z \in \mathbb{R}^{32}$, we couple it together with samples that are distributed from the source domain $X_S \subseteq D_S \in \mathbb{R}^{64}$ (which are represented by temporal extracted information as 64-dimensional vectors). Additionally, since the goal is to generate data that approximate the target domain $X_T \subseteq D_T \in \mathbb{R}^{28\times112\times3}$ conditioned upon emotional information, the conditional information, denoted as $\mathbf{c} \in \mathbb{R}^d$, is added to the equation (where $d = 6$ is the dimensionality of the conditional vector corresponding to the target emotion classes). Finally, emotion-wise loss from a classifier network Q is added. Considering the above formulation, we propose the following overall loss:

$$\min_G \max_D F(D,G) = E_{y\sim X_T}[\log D(y)] + E_{q\sim G}[\log Q(q|c)] + E_{z\sim P_z, x\sim X_S}[\log(1 - D(G(x,c,z)))] \quad (1)$$

where $x \in X_S$, $y \in X_T$ and q stands for a generated sample. Eventually, as explained in [13], in order to calibrate the generator to extract high-quality samples, we add, in Equation 1, an L1-norm loss which is back-propagated to the generator.

A. Wasserstein Generative Adversarial Networks

It has been shown that GANs can optimize network D much easier than G [19]. Minimizing the GANs objective function with an optimal D is equivalent to minimizing the Jensen–Shannon-divergence loss [14]. An optimal D back-propagates the proper gradient for G to be tuned. But if the weights of G are not sufficiently trained yet, the gradient for G diminishes and it does not tune properly. With this in mind, Wasserstein GANs [14] have been introduced in an attempt to mitigate the diminishing gradient problem. This approach implements the Earth-Mover distance as a loss function, which in fact calculates the minimum cost of moving and transforming a pile of mass in order to match the shape of one probability distribution with the shape of another one.

In the case of continuous probability domains, Wasserstein distance can be framed as:

$$W(P_r, P_q) = \inf_{\gamma \sim \Pi(P_r, P_g)} \mathbb{E}_{x,y}[||x - y||] \quad (2)$$

In the formula above, $\Pi(P_r, P_g)$ is the set of all possible joint probability distributions between P_r and P_g (which are representing the real and the generated distributions); and x and y are samples from these distributions. $\inf$ in Equation 2 illustrates that we are searching for the joint distribution that minimizes the amount of mass movement. One joint distribution $\gamma \in \Pi(P_r, P_g)$ describes one possible strategy for mass moving. In particular, $\gamma(x, y)$ states the percentage of mass that should be transported from a sample point x to y so as to make x distribution to approximate the distribution of y. However, it is infeasible to investigate all possible joint distributions in $\Pi(P_r, P_g)$ to compute $\inf \gamma \sim \Pi(P_r, P_g)$. Hence, a common approach is to use the Kantorovich-Rubinstein duality [14], framed as:

$$W(P_r, P_q) = \min_G \max_{D \in \mathcal{D}} \mathbb{E}_{x \sim P_r}[D(x)] - \mathbb{E}_{y \sim P_g}[D(y)] \quad (3)$$

where $\mathcal{D}$ is a set of 1-Lipschitz functions. In this case, instead of judging whether a sample comes from the real or the generated distribution, the task of network D is to minimize the distance in Equation 3 between the distribution of real and generated datasets, and therefore try to learn a proper function $\mathcal{D}$. In order to ensure the 1-Lipschitz condition, one possible way is to implement weight clipping (for D) and to constrain them within a specific range. This formula can become part of GANs approach. However, this approach has proven to be problematic [15] (weight clipping biases critic towards much simpler functions, led to exploding and vanishing gradients and unstable training). Hence, to circumvent tractability issues and to enforce the 1-Lipschitz condition without using weight clipping, we utilize the gradient penalty described in [15], which is an alternative to weight clipping for enforcing the 1-Lipschitz constraint and has been proven to be more stable. Eventually, as in Equation 1, we need to incorporate $E_{q \sim G}[\log Q(q|c)]$ as well.

B. Data pre-processing protocol

The implemented temporal conditional Wasserstein Generative Adversarial Networks architecture (called tc-wGANs for brevity) is trained by making use of CREMA-D [29] and RAVDESS [30] datasets, applying a pre-processing similar to the one described in [13]. Moreover, for the audio modality, an augmentation strategy is utilized [31] in an attempt to increase the number of samples in the dataset. This approach modifies the spectrogram by warping it, in the time direction, by using masking blocks. In this way, for each spectrogram, we are able to construct another 10 different spectrograms. Regarding face modality, firstly, the middle frames of each video are kept (about 50-60 frames were enough for our purposes), removing superfluous frames with low emotional content at the beginning and end of the clips. The selected subsequence is further processed by splitting it in overlapping time-windows of 10 frames. To that end, the window slides by 5 frames each time, which results in an overlap of 5 frames between the consecutive time-windows. Therefore, for each sequence sample, we obtain 10 frame sequences and 10 spectrograms, obtaining 100 different correspondences between the frame sequences and the spectrograms. In this manner, a significantly increased body of training data is obtained, to be given as inputs to the generator and the discriminator networks.

Subsequently, the temporal feature extraction process from face is performed (Figure 1, Block A). Firstly, for each frame, 2dCNN features are extracted, using a CNN of the same topology as the Q network shown in Figure 1 (Block B). The only alteration is the modified input layer to accommodate faces instead of spectrograms. In order to obtain temporal features, we decided to investigate three different strategies, namely 3d Convolutional Networks (3dCNN), Long short-term memory (LSTM), and our Transformer-LSTM mechanism. In the following sections, these strategies are presented.

1) 3dCNN: The input to a 3dCNN is 3D data and more specifically, a 3D tensor $f \in \mathbb{R}^{10 \times 28 \times 28 \times 3}$ (a set of 10 28×28 frames derived from an RGB video, in our case, which entails facial expressions), while the output is a common representation denoted as $\mathbf{h} \in \mathbb{R}^{64}$ (the extracted temporal features). The 3dCNN model extracts features from both spatial and temporal dimensions by performing 3D convolutions, thereby, capturing the motion information encoded in multiple adjacent frames. In this architecture, from the initial tensor, and after applying intermediate 3D convolutions and max pooling operations, the extracted feature map is obtained.

The 3dCNN network was first trained using a classification scheme based on Facial Emotion Recognition (FER). Then, the last fully connected layer ($\mathbf{h} \in \mathbb{R}^{64}$) is kept and is deployed for extracting temporal features.

2) LSTM: As a reference approach, an LSTM architecture [11] was deployed with the same goal to encode temporal information from the input facial frames. The employed strategy was similar to 3dCNN. The input again is a 3D Tensor $f \in \mathbb{R}^{10 \times 28 \times 28 \times 3}$. Thereby, to extract the common representation $\mathbf{c}$, an LSTM architecture using 10 cells was introduced. This network was trained again using an FER scheme. Since we are interested in the encoding, the last output hidden layer was kept, enabling the extraction of temporal features. The output of the network was the hidden layer $\mathbf{h}_t \in \mathbb{R}^{64}$ from the last cell ($t = 10$).

3) Transformer-LSTM: Transformer mechanisms were originally introduced in the domain of machine translation [12] and typically consists of an encoder part (encoding the input language), as well as a decoder (translating to the target language). In this work, however, the focus is on extracting audiovisual representations and, thus, only the encoder part is considered. The extracted representations $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\}$ from the Transformer encoder are combined using a LSTM network into a common representation $\mathbf{c}$, as shown in Figure 2. The encoder architecture is composed of the four following

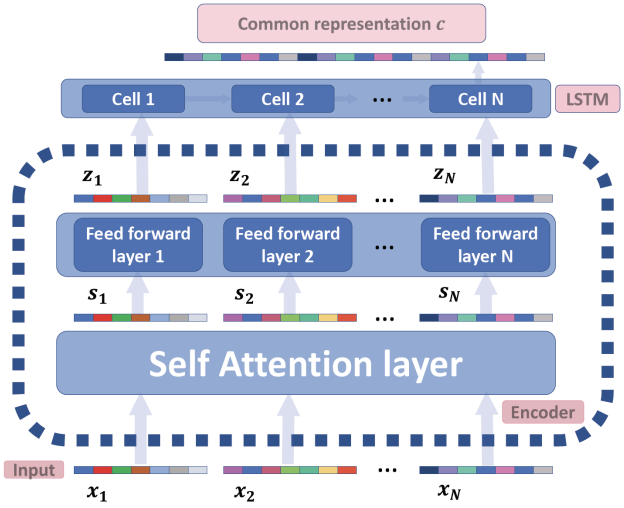


Fig. 2: The building modules of the implemented Transformers architecture when using N input frames.

modules: multi-head self-attention, feed-forward layer and positional encoder, followed by the LSTM layer.

Multi-head self-attention: First module of the encoder topology. Its input is represented by vectors $\mathbf{x}_i \in \mathbb{R}^{d_1}$ where $d_1 = 128$, the dimensionality of faces after applying the 2dCNN. In this module, these input vectors interact with each other to discover which ones need more attention. The outputs are aggregates of these interactions and attention scores. In this way, multiple representations for each input are constructed: $\mathbf{s}_i \in \mathbb{R}^{d_2}$, where $d_2 = 64$, the dimensionality of these representations. Moreover, instead of relying on only computing attention once, a multi-head mechanism goes through the attention operation a number of times in parallel. In its essence, this mechanism produces several attention outputs for a single input. Hence, if we decide, for example, to have four different outputs in our approach, four different z matrices for each input frame $\mathbf{x}_i$ need to be calculated. More details about how these representations are constructed can be found in [12].

Feed-forward layer: Subsequently, the feed-forward layer module combines multiple representations for each input extracted from the previous module. For each input frame, this module results in one representation $\mathbf{z}_i \in \mathbb{R}^d$. The dimensionality d of this output is a parameter that needs to be tuned during training.

Positional encoding: This operator is a way to account for the order of the input sequences. The positional encoding step allows the model to recognize which part of the sequence an input belongs to. That is a pre-processing step that is applied to the input embeddings of the transformer network. It is worth noting that this operation is not part of the learning procedure, meaning that is just an operation and not a trainable component. More details for this operation can be found in [12].

Transformer-LSTM: In order to combine all the extracted embeddings from the feed-forward layer, we implement an LSTM layer which outputs a common shared representation for all time windows (with the dimensionality of $\mathbf{c} \in \mathbb{R}^{64}$). Ultimately, for each input video, the common representation layer (of the LSTM network) outputs one temporal embedding.

C. Datasets

The first dataset used to tune the architecture is CREMA-D [29], an audio-visual emotion expression database, publicly available[1]. It encompasses 7442 videos from 91 actors (43 females and 48 males) aged from 20 to 74 and stemming from a diversity of races and ethnicities (African, American, Asian, Caucasian and Hispanic). Actors were requested to pose 12 sentences associated with six different emotions (Anger, Disgust, Fear, Happiness, Sadness and Neutral) with four levels of intensity (Low, Medium, High and Unspecified). The dataset's annotation is based on videos shown to the participants. A batch of spectrograms for this dataset can be seen in Figure 3a.

The second dataset is RAVDESS [30]. This corpus is a large-scale multimodal emotion expression dataset (derived from speech and song segments) made public in 2018[2]. The database is gender-balanced consisting of 24 actors posing the following expressions: Neutral, Calm, Happy, Sad, Angry, Fear, Surprise, and Disgust. Each expression is produced at two intensity levels. All cases are available in face-and-voice, face-only, and voice-only formats. In our experiments, we made use of the speech segments, ignoring singing segments. The 7356 recordings were rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals. Finally. only the clips with emotional expressions overlapping the ones in CREMA-D were kept (thus, avoiding Surprise and Calm). A batch of spectrograms for this dataset can be seen in Figure 3b.

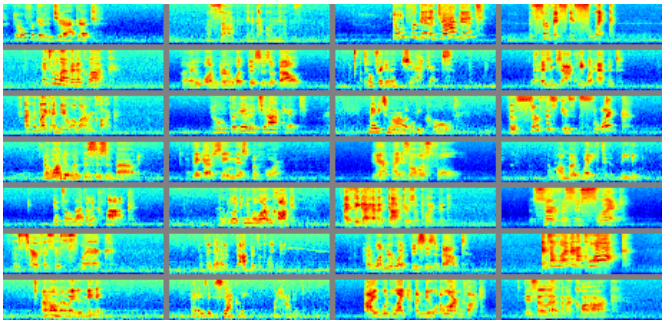
It should be noted that both CREMA-D and RAVDESS, during the whole experimental phase, were balanced with the purpose of containing approximately the same amount of videos for each class. Both datasets were split into three different sets, namely training (60%), validation (30%) and test sets (10%).

IV. Experiments

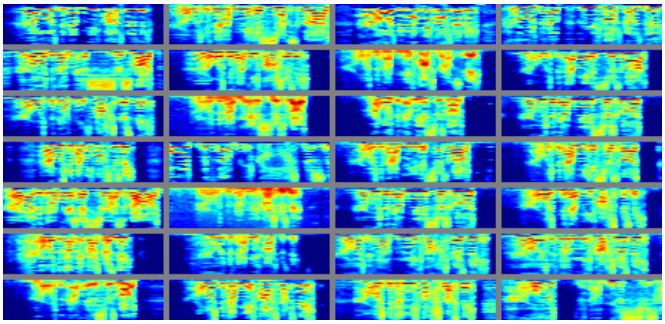
A. Experimental protocol

Baseline method: Firstly, for evaluation purposes, we developed a CNN network for performing Audio-based Emotion Recognition (AER). This architecture is displayed in Figure 1 as classifier Q. This baseline evaluation is intended not only to assess the quality of the generated samples but also to compare our architecture when applied as a data augmentation approach with a basic data augmentation method. For the evaluation of the model, real and generated samples are fused together in a common dataset. Therefore, with the aim of performing a fair evaluation, it is important to use the same amount of samples in both cases. To that end, the methodology introduced in [31]

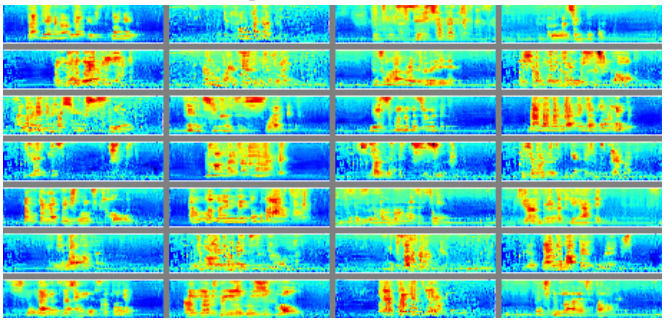
[1]https://github.com/CheyneyComputerScience/CREMA-D
[2]https://zenodo.org/record/1188976



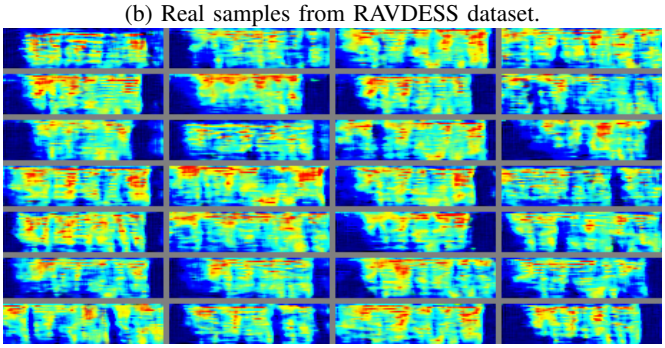
(a) Real samples from CREMA-D dataset.



(b) Real samples from RAVDESS dataset.



(c) Generated spectrograms from CREMA-D.



(d) Generated samples from RAVDESS.

Fig. 3: Real (top) and generated (bottom) spectrograms derived from CREMA-D (left) and RAVDESS (right) datasets when using 3dCNN for feature extraction.

TABLE I: Classification performance, FID, IS and SSIM for all the methods analysed (the best results in bold).

Case	CREMA-D				RAVDESS			
	Classification	FID	IS	SSIM	Classification	FID	IS	SSIM
Baseline	49.34%				44.73%			
dacssGANs [13]	52.52%	59.44	2.16	0.77	47.11%	49.77	2.21	0.90
wGANs-im2im	52.88%	51.55	2.50	0.90	49.81%	41.34	2.27	**0.95**
tc-wGANs:3dCNN	**55.87**%	**38.55**	**2.65**	0.91	**51.81**%	41.34	2.27	**0.95**
tc-wGANs:LSTM	51.07%	39.03	2.60	0.87	49.84%	41.12	**2.32**	**0.96**
tc-wGANs:trs	54.72%	39.15	**2.66**	**0.93**	51.76%	**41.10**	2.30	**0.95**

and briefly described in Section III-B is applied to duplicate the number of training samples in the baseline evaluation.

Conventional GAN approach: This approach, described in [13], is denoted as dacssGANs which stands for Domain Adaptation Conditional Semi-Supervised GANs. This model can be considered as the baseline for the GANs approaches and it is intended to be compared against Wasserstein loss-based frameworks.

Wasserstein loss based approaches: In the same vein, we apply a static version of the whole architecture which utilized an image-to-image approach. This model is based on the dacssGANs but implementing the Wasserstein distance as its loss function. This modification was implemented due to the remarkable difference in training stability using both approaches as stated in the literature [14] [15]. This architecture is denoted in this paper as wGANs-im2im.

Temporal GANs approaches: Finally, the study of temporal information as input to the Wasserstaein network is performed. To prove its efficiency, three different temporal strategies, by making use of the 3dCNN, LSTM and Transformer-LSTM topologies, were evaluated. They are denoted as tc-wGANs:3dCNN, tc-wGANs:LSTM and tc-wGANs:tfs respectively.

B. Evaluation metrics

For a direct comparison, we make use of the same evaluation metrics proposed in [13]. Specifically, first, Audio-based Emotion Recognition (AER) classification is evaluated. In this approach, real and generated samples are fused in a common dataset and an AER classifier was trained using this enriched dataset. The evaluation is performed using a test-set with only real spectrograms and the classification performance is kept as the final evaluation score.

Furthermore, three quality metrics assessing the quality of the generated images are introduced in an attempt to endorse the results from the classification evaluation approach. These metrics are the Fréchet Inception Distance (FID), the Inception Score (IS) and the Structural Similarity index (SSIM). The reader is referred to [32] for further details about these metrics. It is important to mentioned that, the comparison between two images is based on the following three measurements:

luminance, contrast and structure. Therefore, for a batch of images (e.g. when using SSIM), the result were averaged.

Finally, the total amount of time needed for the training, using the same number of epochs and the same hardware, was assessed.

V. Results and Discussion

In Table I, the results of the conducted experiments are shown (for the first four evaluation metrics). The first row contains results for both datasets for the baseline evaluation. This can be used as a reference evaluation for the performance of knowledge transfer of different approaches and to quantify the quality of the samples. It is worth noting that this baseline has been defined to perform a fair evaluation of our models, using a comparable dataset size by applying a simpler data augmentation technique (refer to Section III-B for further details). The results obtained when using the dacssGANs approach can be seen in the second row. As it is displayed in Table I, this approach managed to outbid the baseline evaluation regarding classification performance.

The second and third rows concern the experiments related to the first posed question, related to the efficiency of Wasserstein loss. All quality metrics improve with the application of the Wasserstein loss, while, classification scores improve significantly (47.11% vs 49.8%) in the case of RAVDESS. Classification remains comparable for the CREMA-D data. Regarding wGANs-im2im, it is important to notice that this approach also improved training time significantly. The total amount of training time was reduced by a factor of 4 (see Table II). While the training of dacssGANs when using 80 subjects from CREMA-D for training 100 epochs took approximately $110h$, when using Wasserstein loss, it was approximately $50h$. Therefore, we can conclude that, by using the Wasserstein loss, we can improve the time efficiency and the quality of the generated samples while the performance in knowledge transfer is also increased.

The last three rows of Table I (denoted as tc-wGANs:3dCNN, tc-wGANs:LSTM and tc-wGANs:trs) are referring to the second core research question, whether temporal information improves the performance in comparison with dacssGANs and wGANs-im2im. Regarding tc-wGANs:LSTM, AER performance diminished with regards to dacssGANs. However, in terms of quality metrics, we observed that this technique managed to perform well and generates high-quality samples. Our assumption is that while this method managed to approximate efficiently the distribution of real spectrograms, it does not provide any improvement through capturing temporal audio-visual dynamics and at incorporating nuanced emotion patterns that can boost emotion recognition through generating new audio samples. The results for Transformer-LSTM (tc-wGANs:trs), however, are better than tc-wGANs:LSTM with an increase of about 3.6% and 2% for CREMA-D and RAVDESS, respectively.

However, from Table I, it is obvious that the improvements in most metrics were significantly better when using tc-wGANs:3dCNN. As shown in Table I, in both CREMA-D and RAVDESS datasets, there is approximately 3.3% and 6.7% improvement in terms of recognition rates in comparison to the previously published results with the dacssGANs approach.

TABLE II: The time performance for the experiments conducted regarding the first posed research question. In all cases, we measure the total amount of time after 100 epochs. All the experiments were performed in the same hardware (Titan XP GPU).

Method	Training time (in hours)
dacssGANs [13]	112h
wGANs-im2im	58h
tc-wGANs:3dCNN	47h
tc-wGANs:LSTM	44h
tc-wGANs:trs	48h

Therefore, our tc-wGANs approach when using 3dCNN (tc-wGANs:3dCNN), not only managed to outbid the dacssGANs and the wGANs-im2im approaches for the generation of spectrograms, and reduce significantly the training time, but it also managed to generate samples of high quality for both datasets as seen in the figures 3c and 3d. The quality of the generated results is validated also by the IS, FID and SSIM quality scores. In comparison with the visual results and the introduced metrics, it is obvious that our approach managed to outbid dacssGANs and wGANs-im2im.

Additionally, to verify that the improvement obtained in the paper is consistent with other classification models, some experiments using two additional classifiers were conducted. Firstly, we decided to implement a more complicated network based on the one reported in the paper. While the initial one (classifier Q shown in Figure 1) consisted of 6 layers (leading to 1.4M parameters), the new version includes 16 layers (with 7.5M parameters). The second model applied is the VGG16. This model, with the necessary modifications to inject our data, with 16 layers, has 55M parameters. It is necessary to remark that, although these results should be considered preliminary due to the need for further tuning, which is essential to these networks, these outcomes obtained are promising. In the case of the modification of the classifier Q, both the baseline and also the performance of the model were improved when using the augmented dataset in CREMA-D. The performance for only real data was 54.9% while, when using the augmented dataset (real+generated samples) it reached 58.1%. For VGG16, we obtained 47.7% when using only real data and 50.5% for the augmented dataset again when using CREMA-D.

Finally, regarding the duration of training, we can observe a significant discrepancy between the conventional GANs and the Wasserstein loss cases. However, no significant differences are observed for the tc-wGANs strategies evaluated. As shown in Table II, the training for the tc-wGANs:LSTM was the fastest, lasting approximately 44 hours while wGANs-3dCNN and wGANs-trs performed similarly with 47h and 48h respectively. All the results shown in this table are obtained when training the models during 100 epochs running in an NVIDIA

Titan XP.

VI. Conclusions

This work present a study on audio-visual domain adaptation within emotion expressivity contexts by introducing an approach called temporal conditional Wasserstein GANs (tc-wGANs). The focus of this work is placed on the following two research questions: Firstly, whether Wasserstein loss can help in improving the performance of knowledge transfer between face and audio when using a GANs architecture. Secondly, whether the temporal governed relations between face and audio can be elicited (using 3dCNN, LSTM and a Transformer architecture) in an attempt to optimize the results. From the experimental results, both research questions were validated and it was shown that both time efficiency and the quality of the generated samples improve when using the Wasserstein loss. It has been proved that the efficiency of knowledge transfer improves by 3% and 2% in each dataset for the task of automated emotion recognition by using temporal information extracted when using a 3dCNN architecture, in comparison to a corresponding model not considering temporal data. Additionally, the overall increase in performance, taking into account both temporal information, as well as Wasserstein losses, compared to previous approaches with GANs is even higher, 3.3% and 6.7% for each dataset. Future work will consider the opposite challenging problem, that is, how to generate facial expressivity from emotion-enriched audio. In other words, research will focus on how audio spectrograms associated with a certain emotion can drive the synthesis of facial expressions making use of generative, deep architectures.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the NVIDIA GeForce GTX Titan XP GPU used throughout the experimental phase.

References

[1] S.Albanie, A.Nagrani, A.Vedaldi and A.Zisserman, Emotion Recognition in Speech using Cross-Modal Transfer in the Wild, ACM Multimedia 2018.

[2] T.Grossman, The development of emotion perception in face and voice during infancy, Resorative Neurology and Neuroscience 28, Pages: 219–236, 2010.

[3] D.W.Massaro and M.M.Cohen, Perceiving Talking Faces, Current Directions in Psychological Science, Volume 4, Number 4, 1995.

[4] A.Schirmer, Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing, Social Cognitive and Affective Neuroscience, Volume 13, Number 1, Pages 1–13, 2018.

[5] S.Siddharth, T.P.Jung and T.J.Sejnowski, Impact of Affective Multimedia Content on the Electroencephalogram and Facial Expressions, Nature scientific reports, Volume 9, Number 16295, Pages 257–286, 2019.

[6] S.Jialin Pan and Q.Yang, A Survey on Transfer Learning, IEEE Transactions on knowledge and data engineering, Volume 22, Number 10, Pages 1345 – 1359, 2009.

[7] K.Weiss, T.Khoshgoftaar and D.D.Wang, A survey of transfer learning, Journal of Big Data, Volume 2, Number 9, Pages 1345 – 1359, 2016.

[8] I.J.Goodfellow, J.P.Abadie, M.Mirza, B.Xu, D.W.Farley, S.Ozair, A.Courville and Y. Bengio, Generative Adversarial Networks, NIPS, 2014.

[9] I.Goodfellow, Tutorial: Generative Adversarial Networks, NIPS, 2016.

[10] S.Ji, W.Xu, M.Yang and K.Yu, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 35, Number 1, 2012.

[11] S.Hochreiter and J.Schmidhuber, Long short term memory, Neural Computation, 9(8), Pages: 1735-1780, 1997.

[12] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N.Gomez, L.Kaiser and I. Polosukhin, Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS), 2017.

[13] C.Athanasiadis, E.Hortal and S.Asteriadis, Audio–visual domain adaptation using conditional semi-supervised Generative Adversarial Networks, Neurocomputing, 2019.

[14] M.Arjovsky, S.Chintala and L.Bottou, Wasserstein GAN, 2017.

[15] I.Gulrajani, F.Ahmed, M.Arjovsky, V.Dumoulin and A.Courville, Improved Training of Wasserstein GANs, Advances in Neural Information Processing Systems 30 (NIPS), 2017.

[16] K. Olszewski, Z. Li, C. Yang, Y. Zhou, R. Yu, Z. Huang, S. Xiang, S. Saito, P. Kohli, and H. Li. Realistic dynamic facial textures from a single image using gans. In Proceedings of the IEEE International Conference on Computer Vision, pages 5429–5438, 2017.

[17] A.Radford, L.Metz and S.Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, International Conference on Learning Representations (ICLR), 2016.

[18] K.Vougioukas, S.Petridis and M.Pantic, Realistic Speech-Driven Facial Animation with GANs, International Journal of Computer Vision, Pages: 1–19, 2019.

[19] G.J.Qi, Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities, International Journal of Computer Vision (2019).

[20] L.Chen, S.Srivastava, Z.Duan and C.Xu, Deep Cross-Modal Audio-Visual Generation, Thematic Workshops '17 Proceedings of the on Thematic Workshops of ACM Multimedia, 2017.

[21] W.N.Hsu, D.Harwath and J.Glass, Transfer Learning from Audio-Visual Grounding to Speech Recognition, Interspeech, 2019.

[22] J. S. Chung and A. Zisserman, Out of time: automated lip sync in the wild, Workshop on Multi-view Lip-reading, ACCV, 2016.

[23] B. Korbar, D. Tran and L. Torresani, Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization,Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS), 2018.

[24] L.Zhang, G.Zhu, P.Shen, J.Song, S.A.Shah and M.Bennamoun, Learning Spatiotemporal Features using 3DCNN and Convolutional LSTM for Gesture Recognition, International Conference on Computer Vision (ICCV), 2017.

[25] A.Diba, V.Sharma, L.V.Gool, Deep Temporal Linear Encoding Networks, Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[26] H.Xu, H.Zhang, K.Han, Y.Wang, Y.Peng and X.Li, Learning alignment for multimodal emotion recognition from speech, Interspeech, 2019.

[27] R.Beard et al., Multi-modal sequence fusion via recursive attention for emotion recognition, in Proceedings of the 22nd Conference on Computational Natural Language Learning, 2018.

[28] R.Girdhar, J.Carreira, C.Doersch and A.Zisserman, Video Action Transformer Network, Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[29] H.Cao, D.G.Cooper, M.K.Keutmann, R.C.Gur, A.Nenkova and R.Verma, CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset, IEEE Transactions on Affective Computing, Volume: 5, Number: 4, Pages: 377–390, 2014.

[30] S.R.Livingstone and F.A.Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, Volume 13, Number 5, Pages: 1—35, 2018.

[31] D.S.Park, W.Chan, Y.Zhang, C.C.Chiu, B.Zoph, E.D.Cubuk and Q.V.Le, SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, Interspeech 2019.

[32] Z.Wang, A.C.Bovik, H.R.Sheikh and E.P.Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, EEE Transactions on image processing, Volume 13, Number 4, 2004.