Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 1–20

Procedia Computer Science

Audio-visual domain adaptation using conditional semi-supervised Generative Adversarial Networks.

Christos Athanasiadis, Enrique Hortal and Stylianos Asteriadis

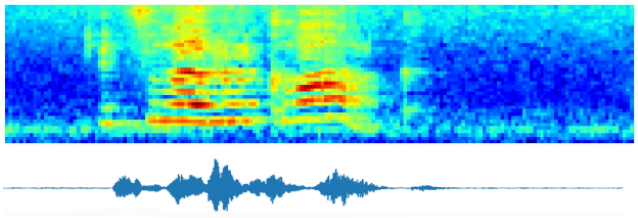
Abstract

Accessing large, manually annotated audio databases in an effort to create robust models for emotion recognition is a notably difficult task, handicapped by the annotation cost and label ambiguities. On the contrary, there are plenty of publicly available datasets for emotion recognition which are based on facial expressivity due to the prevailing role of computer vision in deep learning research, nowadays. Thereby, in the current work, we performed a study on cross-modal transfer knowledge between audio and facial modalities within the emotional context. More concretely, we investigated whether facial information from videos could be used to boost the awareness and the prediction tracking of emotions in audio signals. Our approach was based on a simple hypothesis: that the emotional state's content of a person's oral expression correlates with the corresponding facial expressions. Research in the domain of cognitive psychology was affirmative to our hypothesis and suggests that visual information related to emotions fused with the auditory signal is used from humans in a cross-modal integration schema to better understand emotions. In this regard, a method called dacssGAN (which stands for Domain Adaptation Conditional Semi-Supervised Generative Adversarial Networks) is introduced in this work, in an effort to bridge these two inherently different domains. Given as input the source domain (visual data) and some conditional information that is based on inductive conformal prediction, the proposed architecture generates data distributions that are as close as possible to the target domain (audio data). Through experimentation, it is shown that classification performance of an expanded dataset using real audio enhanced with generated samples produced using dacssGAN (50.29% and 48.65%) outperforms the one obtained merely using real audio samples (49.34% and 46.90%) for two publicly available audio-visual emotion datasets.

Keywords:

1. Introduction

Scientists in the domain of cognitive psychology have long studied the relationship between facial and vocal cues in humans [1] [2]. In particular, researchers suggested that infants, during the development of their auditory and visual perceptions, fuse facial cues together with audio information in an effort to better discriminate and recognize emotions. In the same spirit, authors in [3] performed a study of how deaf people perceive sounds of phonemes by eliciting their visual perceptual system with the purpose of performing lipreading (or speech-reading). In a similar manner, concerning the emotional cross-modal relationships, prosodic speech information (linguistics variation in speech like pitch tempo, loudness, etc.) and its correlation with facial features have been intensively studied in [4] [5] [6]. The cardinal outcome of these works was that speech prosodic information is associated with other social cues such as facial expression, body language or tone tempo. In particular, authors in [6] suggested that speech prosodic information could be extracted merely using facial cues. Hence, a worthwhile research question that is inspired by the theoretical conducted research is whether the connection between the audio-visual information could be examined from the emotional point of view. In this regard, we re-frame the question as: Does the emotional state content of a persons' voice correlate with their facial expression?



(a) Audio modality.

(b) Face modality.

Figure 1: An instance of the symbiotic audio-visual modalities from a CREMA-D dataset video clip.

In the light of the aforementioned cognitive research, we intend to transform the same questions posed from the cognitive psychologists into the domain of emotion recognition and Domain Adaptation (DA) from the computer science perspective and place the following question: Is it possible to transfer knowledge between facial expressions to mere audio? The importance of this question lies in the fact that, while emotion recognition through facial expressions has been studied extensively [7] [8] [9], emotion recognition through other modalities such as audio has produced fewer advancements concerning classification results [10]. An important reason behind this is the fact that there are not a lot of publicly available datasets for audio emotion recognition compared to the abundance of data that exists regarding facial expressivity. Therefore, generating training models for emotion recognition through audio modality can be a rather challenging task and requires the generation of new robust datasets. Meanwhile, the engineering of such big and complex corpora is not always a straightforward and feasible task.

In order to facilitate these limitations, Domain Adaptation (defined also as Transfer Learning) algorithms are fostered from researchers with the objective of developing classification methods for specific modalities by exploiting data from other similar ones coming from rich available datasets [11] performing the same in-hand task [11][12]. These techniques were inspired by the human behavior and the way that the learning process is materialized in human brain by "re-using" previous knowledge to handle new situations.

In the case under consideration, abundant emotional annotations for predicting emotional states from visual information are available. Therefore, traditional machine learning algorithms could be employed in order to develop efficient classifiers. However, sparse datasets available from the audio modality constrain this task. Training a classifier by using only the sparse available dataset of audio could be proven insufficient. Nonetheless, the datasets from both modalities are associated, since they are attributed to the same emotions and, therefore, to the same classification task. Yet, the distributions of face and audio modalities are not comparable, so the dense dataset from face cannot be directly employed into the emotion recognition classification task from audio. In the meantime, as aforesaid, it is already suggested from the cognitive psychologists that these two modalities are related to each other, thus, a mapping between these two domains could be attained. This mapping can be obtained using DA techniques and it can serve not only to improve the accuracy of the audio classification using facial information but also to expand or create new audio emotion-related datasets.

In the current work, we adopted a Generative Adversarial Networks (GANs) algorithm in order to study these cross-modal relationships between the symbiotic modalities of video (shown in Figure 1) and perform Domain Adaptation. In recent years, explosive popularity has emerged in the domain of GANs [13] which became one of the most promising developments in Deep Learning. The preliminary idea of GANs can be framed as follows: Given a vector of random noise z, the whole process endeavors to accomplish a good approximation of the data distribution in hand (in our case, represented as the target domain) by learning a mapping between the noise distribution and that domain. GANs usually consist of two different neural networks which compete with each other in a min-max manner. These networks are called Generator G and Discriminator D and they are depicted in Figure 2. An illustrative example of how GANs function is introduced in [13]. In this work, the target is to train a network G that, given a noise vector z, is able to generate new samples derived from the MNIST dataset domain (target domain) by trying to approximate the desired distribution. In the meantime, D tries to decide whether the generated samples are genuine or not.

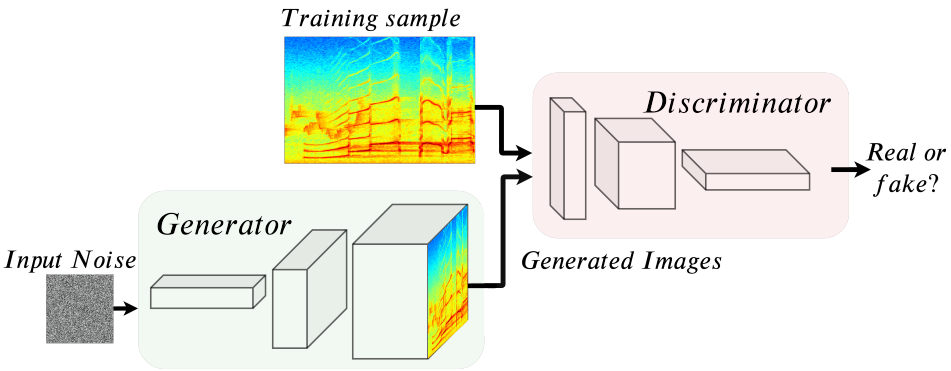


Figure 2: Initial version of the Generative Adversarial Networks.

Starting from the above-mentioned architecture (the so-called vanilla architecture), the scope of the proposed research is to modify it and adapt it to the needs of our goals. The desired objective is to develop a framework that will be able to not only generate data in the target domain but also to convert source samples into target domain ones. Thereby, several modifications were needed in the classical version of GANs for the sake of formalizing a system that will be able to perform as such. Recent advancements in GANs suggested several modifications that make them more suitable for the field of Domain Adaptation and audio-visual cross-modal mapping and provide fertile inspiration to the current work. A cardinal influence was the work done in [14] [15] and, particularly, the one in [16] where a conditional deep Generative Adversarial Network was proposed with the aim of performing image-to-image translation. In that modified version of GANs, a U-Net [17] architecture was proposed with a view to learning the domain shift between two different image datasets that share some characteristics. Contrary to the work done in [16], we propose a semi-supervised architecture, the so-called dacssGAN (Domain Adaptation Conditional Semi-Supervised Generative Adversarial Network) where the input of the generator contains, apart from the source modality data, conditional semi-supervised information extracted using a facial expression classifier (based on convolutional neural networks) and it is processed using conformal prediction (CP) [18] [19]. Conformal prediction is a framework for credible machine learning, constituting a methodology for obtaining error calibration in classification and regression tasks. This framework is based on hypothesis assumptions in an effort to provide rigorous error calibration. It allows obtaining confidence values for any class label given a test instance. In the current work, the implementation of CP is performed in order to provide robust conditional information as input to the proposed dacssGAN architecture.

On the whole, a synopsis of the current work's contributions is summarized as follows:

- The challenging task of heterogeneous semi-supervised domain adaptation between the symbiotic audio-visual modalities in the affective understanding context is explored.
- A novel label-agnostic architecture for GANs based on conditional information extracted using conformal predictions is introduced.
- Inductive conformal prediction [20] is evaluated with a view to remedying the high implementation cost of the traditional conformal prediction approach.
- A regulation mechanism over the generator that consists of an auxiliary classifier was opted in an effort to impose the emotion states over the generated samples.
- The evaluation of the domain adaptation procedure was performed by implementing a data augmentation schema (similar to [21] and [22]) where generated and real samples were fused together and emotion recognition was performed in this expanded dataset.
- An ablation study was performed in an attempt to investigate the capability of different architectures, loss functions and the performance of different conditional inputs to the presented GAN approach.

Finally, the structure of the remainder of this paper is as follows: In Section 2, several definitions and the related work on Domain Adaptation (DA), GANs and CP are presented. Section 3 describes the introduced DA method that is based on the proposed dacssGAN while in Section 4 the experimental protocol, dataset and results are presented and analyzed. Finally, Section 5 contains the conclusion and the future work of this study.

2. Related work

In this section, some related works regarding the fields of Domain Adaptation (along with a definition of its terminology), GANs, audio-visual relationships and conformal prediction are presented.

2.1. Domain adaptation definitions

For readability purposes, this sub-section introduces a Domain Adaptation terminology dictionary which will help readers understand the concepts and methodologies used in the remainder of the paper. Primarily, the term domain D_{domain} is defined by using the following two notions. Firstly, the feature space $\mathcal{X}$, which contains all possible instances while $X = \{x_1, x_2, ..., x_n\} \in \mathcal{X}$ is a subset of the feature space and contains the in-hand available learning sample vectors. Secondly, the feature distribution probability of the learning samples features is defined as $P(X)$. Then, the classification task, denoted as T, is introduced. Two new terms should be defined in that case. Firstly, the label space $\mathcal{Y}$ and, secondly, the classification function $f(X)$, which can be used to map a new unknown input feature vector $x_{new} \in \mathcal{X}$ to the label space $\mathcal{Y}$. In particular, if the machine learning problem under study is the classification task of emotion recognition through facial expressions using six basic emotional states (happiness, sadness, fear, disgust, anger and neutral), then the feature space $\mathcal{X}$ is represented from all possible values (between 0-255) that the pixels of the images from the facial domain could take, while the labels $y_i \in \mathcal{Y}$ are represented from the aforementioned six basic emotions. It should be also stressed that for two different classification tasks, the feature domains and the feature distribution probability can be vastly different.

In addition, the terms source and target domain are introduced. Source domain is considered a space which contains data that will be used in order to perform the transfer of knowledge. For the current work, the facial expression modality is defined as the source domain. Formally, the source domain can be defined as $D_S = (X_S, Y_S, P(X_S))$ (with $X_S \subset \mathcal{X}_S$, the feature set which is subset of the feature space that represents the source data). On the other hand, the target domain is the sub-domain that needs to be enhanced through transferred knowledge stemming from the source domain. This domain is defined as: $D_T = (X_T, Y_T, P(X_T))$ (with $X_T \subset \mathcal{X}_T$, the feature set that is subset of the feature space that represents the target data). In the same spirit, a definition for source and target classification tasks could be defined as follows: Insofar as the first is concerned, the source task T_S is the classification task that can be trained using the data from the source domain D_S while for the latter, the classification task T_T is the one that can be applied using the target domain data D_T. The classification task for each domain consists in calculating the predictive classification function for each case: $f_S(X_S)$ and $f_T(X_T)$. This is done by incorporating the feature vectors from the training set and learn the relation between the feature vectors and the corresponding labels. By definition, the scope of Domain Adaptation is to extract the knowledge from the source task and to apply this knowledge on the target task. This transfer of knowledge is implemented with the purpose of improving the performance of the classification task in the target domain by incorporating knowledge from the source domain, thus improving the performance of the predictive classification function $f_T(X)$.

Having defined basic terms associated with Domain Adaptation, several emerging scenarios arise regarding the nature of the available data and the way that DA could be utilized. For the source and target domains $D_S = (X_S, Y_S, P(X_S)$ and $D_T = (X_T, Y_T, P(X_T))$, the emerging cases for the DA are correlated to the following conditions: $\mathcal{X}_S \neq \mathcal{X}_T$, $T_S \neq T_T$ and $P(X_S) \neq P(X_S)$. In the case that the source and target domains are not the same ($\mathcal{X}_S \neq \mathcal{X}_T$), the approach is defined as *Heterogeneous Domain Adaptation* while if $\mathcal{X}_S = \mathcal{X}_T$ and $P(X_S) \neq P(X_S)$ the approach is defined as *Homogeneous Domain Adaptation*.

Another division between the DA approaches is associated with the availability in label information. There are three main scenarios: supervised, semi-supervised and unsupervised Domain Adaptation. As far as the first is concerned, both source and target domains contain dense datasets with fully available label information. In the second case, there exist some small amount of label information or there is no label information regarding the target domain but there is an auxiliary way to calculate them (likewise with the usage of a classifier). Finally, in the last case there

is a lack of information concerning labels in the target domain and any auxiliary classifier. In the framework of the current work, the applied the Domain Adaptation schema could be characterized as a heterogeneous semi-supervised approach.

2.2. Domain Adaptation related work

The work done in [11][12] introduced state of the art approaches for Domain Adaptation while the works in [23] [24] address state of the art approaches for visual Domain Adaptation. Most of the recent approaches could be roughly categorized into two groups depending on whether they are deep-learning based or not.

In the former case, deep-learning based approaches, several works have been presented. For example, authors in the work titled Domain-Adversarial Training of Neural Networks (DANN) [25] facilitate Domain Adaptation from the learning representation perspective. They attempt to jointly learn representations for both source and target domain samples by introducing a neural network that is having as loss function a domain divergence loss ($\mathcal{H}$-divergence) that calculates the distance between the two domains. This loss was coupled together with classification loss that is using the supervised information exclusively from the source domain. Inspired from this work, authors in [26] proposed VRADA model (which employs variational recurrent adversarial networks) for the purpose of capturing and transferring temporal latent dependencies across domains via domain-invariant representations (for real-world health-care time-series data). The work done in [27] facilitates the implicit discourse classification problem in a principled adversarial manner. A deep learning architecture was established which was composed from a network (i-CNN) that extracts embeddings related to implicit input and a network (a-CNN) that extracts embeddings for the same implicit input enhanced with explicit connectors. Furthermore, an adversarial network (discriminator) judges whether the inputs comes from the i-CNN or a-CNN. Finally, on top of the previous networks, a final CNN network performs the final discourse classification. Authors in [28] propose a covariant multimodal attention based multimodal domain adaptation neural network (MDANN). In that work authors tried to investigate whether it is possible to perform domain adaptation that can transfer knowledge from one multimodal dataset to another one. That was done by trying to learn a common feature representation for multiple modalities and mitigate inter-domain divergence by applying jointly adversarial loss among the different modalities.

Regarding the latter case, non-deep learning approaches, authors in [29] tackle the homogeneous domain adaptation task in an unsupervised manner by trying to spot correspondences between samples in the source and target domains. The correspondences were obtained by treating the source and target samples as graphs and using a convex criterion to match them. The criteria used were first-order and second-order similarities between the graphs as well as a class-based regularization. Experiments performed in several image classification datasets as well as in toy datasets. Similarly, authors in [30] authors accommodated the same task by considering also higher-order similarities. In the work presented in [31] titled Optimal Transport for Domain adaptation authors accommodated the unsupervised and semi-supervised Domain Adaptation problem as in [29] graph matching problem by trying to bring close source and target modality by using Monge-Kantorovich (alternatively called as Wasserstein) distance coupled together with several regularizers and by using generalized conditional gradient (GCG). Finally, in [32] a hybrid version of Deep neural networks and graph matching approaches was developed. Neural networks were employed to extract domain invariant representations, that used graph matching loss as the domain discrepancy metric.

2.3. Generative Adversarial Networks

In this sub-section, state of the art techniques for Generative Adversarial Networks which influence our work are presented. In conditional GANs introduced in [15], networks G and D were conditioned to some variables c that represent the label information of the class. The model was not only managed to generate data that represent those labels but also improved the quality of the generated data. Similarly, in [14], a modified version of the initial GANs which makes use of Deep Convolutional Neural Networks for the G and D network was proposed. Authors in [33] presented an approach to learn how to translate an image from a source domain X_S to a target domain X_T without having any available paired information among these two domains. The main objective of the approach was to learn a mapping $G : X_S \rightarrow X_T$ such that the distribution of images from $G(X_S)$ domain is equivalent with the distribution X_T using an adversarial loss. Since that mapping is highly under-constrained, authors banded it together with an inverse mapping $F : X_T \rightarrow X_S$ and introduced a cycle consistency loss to force $F(G(X_S)) \approx X_S$ (and vice versa). Qualitative and quantitative results were delivered on several tasks where paired training data were not available. In DiscoGAN [34],

authors tried also to uncover cross-domain relations given unpaired data (that were representing two different image domains). Authors proposed a method based on GANs that learns to discover relations between different domains. Using these uncovered relations, the proposed network effectively transferred style from one domain to another while maintaining important image features such as orientation and face identity. Cycle-Consistent Adversarial Domain Adaptation or CyCADA [35] proposed a novel discriminatively-trained technique. It suggests that GANs, combined with cycle-consistency constraints, are surprisingly effective at mapping images between domains, even without the use of aligned image pairs. CyCADA transforms image domains at both pixel and feature levels and enforces cycle-consistency constraints while leveraging a task loss. To validate this approach, the authors applied their model in a variety of visual recognition and prediction tasks. In [36], authors tried to establish a bridge between the definitions of GANs and Variational Autoencoders (VAE) by reformulating the definition of GANs, which translates the generation of samples as performing posterior inference. In that way, several state of the art approaches from VAE can be transferred to GANs and vice versa.

2.4. Audio-Visual relationships

The state of the art audio-visual studies that mainly influence our study are presented in this sub-section. An interesting study of cross-modal relationships of audio and visual cues was introduced in [37] where conditional GANs were applied with the purpose of generating one modality while another modality was given as an input. In order to do so, authors introduced two separate networks (image-to-sound and sound-to-image) in order to perform cross-modal generations in both ways. Inspired by this work, authors in [38], built a model called Cross-Modal Cycle Generative Adversarial Model in an effort to perform cross-modal mappings between image and audio. Authors in [39] introduced a system that performs audio-video synchronization between mouth and speech in a video. To facilitate the task, a two-stream network was implemented by having one network dedicated for audio and one for video and coupled together by using the constructive loss that is judging whether or not the embeddings from the two streams belong to a synchronized video pair or not. Similarly in [42], a audio-visual study was performed with the purpose of performing temporal synchronization. Likewise in [39] a two stream network using constructive loss function was implemented. In contrast with that work, the negative pairs were chosen to be within the same videos and furthermore, authors employed 3D CNN with the purpose of learning spatio-temporal features that can model the correlation between face and audio modalities. In [40], an audiovisual approach for the emotion recognition in the wild challenge was introduced. This approach contains two pre-processing steps. Firstly, a voice activity detector based on Recurrent Neural Network (RNN) that returns only the speech segments from the videos, then secondly, a lip activity detection returns only the segments where the speech is in-line with the person seen in the video. Afterwards, low level descriptors and Mel Frequency based features were used to extract useful information from audio, while local binary patterns (LBP) were employed for extracting features from video. Finally, mono-model and multi-modal emotion classification were performed by using Support Vector Machines. The same authors recently published an open source toolbox that could be employed for studying these audio-visual relationships [41].

2.5. Conformal prediction

In this sub-section, state of the art techniques for conformal prediction and the works that mainly influence our study are presented. In [18], the initial formulation of the conformal prediction is presented. Conformal prediction was proposed as a more rigorous and efficient way to extract confidence of the prediction of a classifier. Authors introduced two important definitions for their approach. Firstly, the nonconformity measure, a value that is calculated based on the classifier prediction, was defined. Secondly, given this value and a calibration set, the second term that was established is the p-value defined as the improved prediction confidence of the whole conformal prediction approach. In [43], an extension of the initial formulation of the algorithm is introduced. This technique was noted as inductive conformal prediction. In this case, the data used to validate the approach is split into three sets, the training bucket which is used to train the classifier, the calibration set used for calculating the nonconformity values and finally, the test set which is the data used to calculate the conformal prediction values. Furthermore, in contrast with the initial conformal prediction, the training of the classifier occurs just once at the beginning. Cross conformal prediction [44] is a modification of the previous approach for more efficient results. In this case, a split between training and calibration set is performed in a cross-validation manner in K steps and then, in each case, the calculation of the p-values of the test set is done. Finally, the final p-values of the test set are the average of the K different values. More details about the implementation of the CP for the current work can be found in Section 3.1.

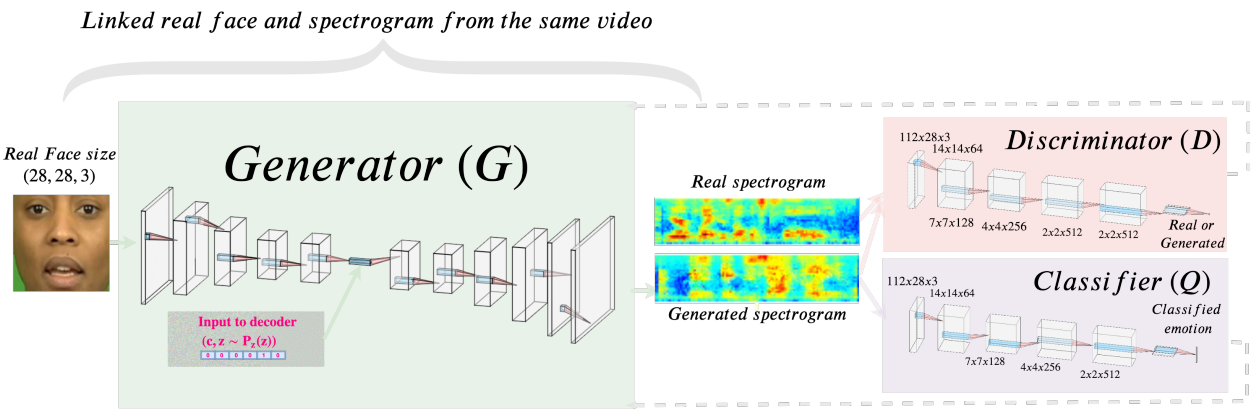


Figure 3: Complete architecture of the dacssGAN approach.

3. DacssGAN approach

In this section, the proposed dacssGAN approach is discussed in detail. The overall architecture of the approach can been seen in Figure 3. While in Figures 4a and 4b two different architectures were implemented for the network G and were examined during the ablation study in sub-section 4.4 are depicted. GANs, as mentioned before, consist of two networks, a generator G and a discriminator D. Given a noise vector z as an input to the network G and a dataset of samples that comes from the target domain distribution $X_T = \{x_1, x_2, ..., x_n\} \subseteq D_T$, network G is calibrated to generate unseen samples that resemble that distribution, while D is trained to examine whether the generated samples are genuine or not. The whole training procedure is occurring in an adversarial fashion implemented as a min-max algorithm. The initial formalization of that game could be framed with the following equation:

$$\min_G \max_D V_1(D, G) = \mathbb{E}_{y \sim X_T(x)}[\log D(x)] + \mathbb{E}_{z \sim P_z}[\log(1 - D(G(z)))] \tag{1}$$

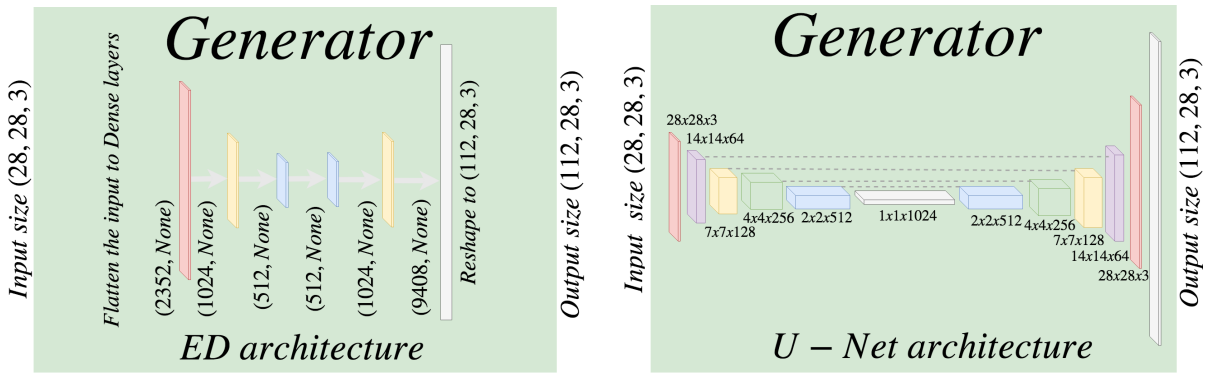
where $y \in X_T$ (samples coming from target domain D_T) corresponds to the data that we want to approximate while $z \in P_z$ corresponds to the distribution domain that the noise vector z is sampled from. However, the objective in the current work is to implement a domain shift and calculate a transformation between source (D_s) and target domain (D_T). Thereby, instead of having as input to the G network the noise vector $z \in P_z$, samples that are distributed from the source domain $X_S = \{x_1, x_2, ..., x_n\} \subseteq D_S$ should be utilized. However, if we proceed by neglecting completely the noise vector z, that may result in the development of a network that only produces deterministic outputs. Thus, noise vector z should be fused together with the source domain samples in G [45]. Consequently, Equation 1 becomes:

$$\min_G \max_D V_2(D, G) = \mathbb{E}_{y \sim X_T}[\log D(y)] + \mathbb{E}_{z \sim P_z, x \sim X_S}[\log(1 - D(G(x, z)))] \tag{2}$$

where x are samples derived from the source domain ($x \in X_S$) and y are samples belonging to the target domain ($y \in X_T$). Additionally, since the goal is to generated data that approximate the target domain $X_T \subseteq D_T$ conditioned to emotional information, Equation 2 could be easily re-framed for the conditional scenario as:

$$\min_G \max_D V_3(D, G) = \mathbb{E}_{y \sim X_T}[\log D(y)] + \mathbb{E}_{z \sim P_z, x \sim X_S}[\log(1 - D(G(x|c, z)))] \tag{3}$$

where the input in G network is conditioned to variable information c. In the current framework, we examined the possibility of having three alternative sources of conditional information as input to the network G namely: class label information, prediction of a classifier (trained using datasets that derived from source domain X_S), and conformal predictions [19]. The latter is a technique that has been used for error calibration in classification settings. It is analyzed in more details in sub-section 3.1. Furthermore, similar to [17], in our study we investigated the possibility of combining the initial GANs objective with a more classical loss, such as L1 distance [46]. The discriminators task remains the same, however, the generator is deputed to not only fool the discriminator but also be near the ground truth output in an L1 manner (that, in our case, is calculated in a pixel-wise manner). It was found [17] that L1 norm



(a) The architecture of the network G (ED). (b) The architecture of the network G (U-NET).

Figure 4: Different tested architectures for the network G.

encourages less blurring in the results than other metrics like L2 norm. The L1 loss function that was implemented for our framework is formulated as:

$$L_1(G) = E_{y\sim X_Y, x\sim X_S, z\sim p_z(z)}[||y - G(x, z)||_1] \tag{4}$$

The complete optimization schema that derives after combining Equations 3 and 4 is formulated as follows:

$$\min_G \max_D (V_3(D, G(x|c, z)) + L_1(G(x|c, z))) \tag{5}$$

Hitherto, the system made use of some conditional information about the label, however, during the experimental phase it was found that by just adding this conditional variable c is not efficient enough to produce genuine samples from the target domain (that also represent the desired emotional states). Thereby, we decided to add an extra Network $Q = f_T(x_{generated} \in G(x|c, z))$ that is producing an error based on the correct or wrong classification of the emotional states. This model Q is presented in Figure 3 as "Classifier". By using this network, we managed to solve a problem emerged in that framework, the conditional information inputted in the network G was not itself sufficient of reproducing well the class information in the generated samples. The proposed network Q is a CNN network with an architecture similar to that used in the network D. However, they differ in the last layer that in the classifier case outputs the predicted emotion state for the input audio samples instead of the binary decision provided by the network D. The input of this network is the output samples of G and the cross-entropy error was passed to the Generator optimization in tandem with Formula 5. Intuitively, we want to calibrate and influence the output of the network G by considering its capability to reproduce samples that are not only governed from the target domain X_T distribution but also represent as good as possible the class information that the samples of the X_T are associated with. In the work presented in [47] and [48], authors already introduced an extra classifier in the whole training process of GANs. In our approach, the introduced error of the classifier, the cross-entropy of the generated samples from the network G, could be denoted as:

$$L_C = \mathbb{E}[\log(P(x_k|D(G(x|c, z))))] \tag{6}$$

where x_k denotes the probability of a sample to belong to the specific class k. Eventually, the complete loss function is the summary of Equation 6 and 5:

$$\min_G \max_D (V_2(D, G(x|c, z)) + L_1(G(x|c, z)) + \mathbb{E}[\log(P(x_k|G(x|c, z))]) \tag{7}$$

3.1. Semi-supervised GANs using conformal prediction

The conditional information c that has been applied as a supplementary input of the network G (together with the vector noise z) in the classical version of conditional GANs was mainly associated with the label information

($c = y_i \in \mathcal{Y}$) of the target domain samples ($x \in X_T$). However, these labels are not always available, and being able to construct a network G regardless of the label information is one of the main objectives in the current paper. Having said that, the very first approach that was fostered to displace the conditional label information was to make use of a trained classifier $f_S(X_S)$ that related to source classification task T_S. This classifier will output the prediction confidence that depicts the probability for each sample of being associated with each class ($y_i \in \mathcal{Y}$ that represent in our case every different emotion, denoted in Equation 7 as c). In this respect, by neglecting the class information in the input of the network G, we could state that our generator is operating in a semi-supervised and class-agnostic manner. Apart from this though, as we already mentioned, we investigated the efficiency of conformal prediction [19] as a way to output a better class-confidence and use it as the conditional information that is fed into the network G instead of simply use the classifier prediction output. The CP framework is a probabilistic approach focused on post-processing of classification results for more reliable predictions. The CP framework combines a methodology of algorithmic randomness and hypothesis testing to provide error calibration in online settings. For the sake of being more concrete, an illustrative example of the way CP functions is described: Given a dataset $X_D = \{(x_1, y^p), (x_2, y^p), ..., (x_n, y^p)\}$ (where $p \in \{1, 2, ..., w\}$ with w the number of classes), a classifier $f(X)$ and a new test data point x_{n+1}, the hypothesis that x_{n+1} is assigned to a specific class label $c = y^p \in \mathcal{Y}$ is adopted. Having defined the test hypothesis, a re-train process for the adopted classifier $f(X)$, with $D \cup \{x_{n+1}, y^p\}$ is performed. Subsequently, a nonconformity function for all the data points $X_D = \{(x_1, x_2, ..., x_n)\}$ is re-computed assuming this hypothesis true. That nonconformity function is correlated with the chosen classifier that was selected. In [49], several nonconformity measures that correspond to several classifiers are presented. For the case of CNN the nonconformity measure could be framed as (with the following two cases):

$$a_{n+1}^{y^p} = \begin{cases} 1 - o_{n+1}^{y^p} \\ -o_{n+1}^{y^p} + \max_{i=1,..,w, i\neq p} o_{n+1}^{y^i} \end{cases} \tag{8}$$

where $o_{n+1}^{y^p}$ denotes the output of the last layer of the CNN for the specific class and in particular, corresponds to the softmax output function of the CNN architecture. The $\max_{i=1,...,M, i\neq p} o_{n+1}^{y^i}$ corresponds to the higher value among all conformity hypothesis excluded the case that $i \neq p$. In the current approach, we used the second part of the Equation 8 as the established nonconformity measure. The next step of the approach was to define the p-value measurement, given in Equation 9:

$$p(a_{n+1}^{y^p}) = \frac{count\{i \in \{1, ..., n+1\} : a_i^{y^P} \geq a_{n+1}^{y^P}\}}{n+1} \tag{9}$$

where $a_{n+1}^{y^P}$ denotes the nonconformity measure of x_{n+1} when it is assumed that it belongs to class label $c = y^p$. This test hypothesis is performed with all available classes (and the corresponding p-value for each hypothesis is calculated). It is obvious that the p-value is highest when all nonconformity measures of training data belonging to class $c = y^p$ are lower than that of the new test point x_{n+1}, which points out that x_{n+1} is most conformal to the class $c = y^p$. This process is repeated by performing the null hypothesis for all the class labels, and the highest p-value is used to decide the actual class label to be assigned to x_{n+1}. Considering p_j as the highest p-value and p_k as the second highest p-value, p_j is called the credibility of the decision while 1 - p_k represents the confidence of the classifier's decision.

Algorithm 1 Pseudo-code for the conformal prediction process [49].

1: Given a training set $D = (x_i, y^p), ..., (x_n, y^p), x_i \in X$, number of classes $y^p \in Y = y^1, y^2, ..., y^w$ and a classifier $f(X)$:
2: Get a new unlabeled sample x_{n+1}.
3: **for** all class labels y^j, where j = 1, ..., w **do**
4: Assign label y^j to x_{n+1}.
5: Re-train the classifier $f(X)$, with $D \cup \{x_{n+1}, y^{(j)}\}$.
6: Compute nonconformity measure value, $a_i^{y^j}$ with $i = 1...., n+1$ to compute the p-value
7: **end for**
8: Output the conformal prediction based on the p-value prediction regions.

On the whole, the methodology is summarized in Algorithm 1. However, as it was already mentioned before, in the current approach, we made use of the inductive conformal predictions that is introduced in [20], where a set of size $l = n - r$ was denoted as training set and a set of size r that is denoted as the calibration set of the conformal prediction. For a new x_{n+1} now the calculation of p-value is occurring without the re-training of the classifier $f(X)$ but just with directly comparing its nonconformity value of that sample with the nonconformity values of the calibration set.

In the light of the above, the calculated p-values were given as input to the proposed dacssGAN architecture instead of directly using the outcome of the classifier $f_S(X_S)$, as illustrated in Figure 3. The rationale behind using inductive CP was the high complexity (computational) cost of the initial conformal prediction algorithm in combination with a GAN architecture. Additionally, experimental tests resulted in similar behaviours when using inductive CP in our dacssGAN approach in comparison with the plain version of the CP algorithm. The turning of the calibration set size r, was based on techniques that could be traced in the literature and are described in [20]. Those techniques are referred in that work as query function. In an effort to efficiently reduce the computational cost, we decided to employ as a query function the random split of the sets and make a comparison of the classification performance with the performance of the initial conformal prediction algorithm.

3.2. Mode collapse problem and remedies

One of the most profound drawbacks of GANs algorithm is the so-called "Mode collapse". Mode collapse is a phenomenon where the network G generates a limited diversity of samples, or even the same samples, regardless of the input. Authors in [50] presented mode collapse and provided a precise explanation about the reasons why the phenomenon occurs. Since training is a stochastic process, due to the randomness introduced with vector z, during the early stages in training, the generated samples will deviate depending on $z \in P(z)$ and the samples drawn from $x \in \mathcal{X}$. In other words, the gradients back-propagated to the network G will deviate between training steps relying on input information.

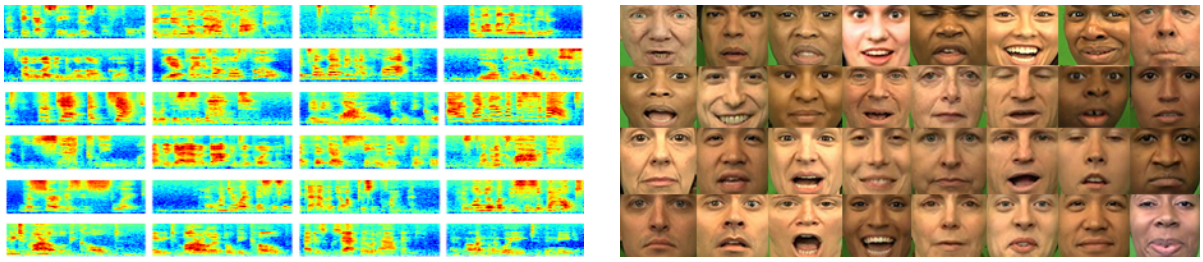
However, in practice, there exists a single fixed point for the weights that network G considers as the optimum ones for the generation process regardless of the input information we fed into it and there is nothing in the objective function that explicitly forces the network G to generate different samples given a different input. For its part, Network D eventually is not imposing any more variety in the generated samples or forcing the partially collapsed G to a different direction.

Possible remedies that are proposed in the literature (mainly in [51] [52] [53]) and which were proven extremely useful during the experimental procedure for the current work were: input normalization, batch normalization, and the use of *LeakyRelu* [54] as activation function in all the networks of the proposed architecture. Further remedies that were implemented in the current approach to mitigate the mode collapse phenomenon were the application of soft and noisy labels (in the case that the conditional variable c was represented from the real label information, see "Supervised conditional GAN" in Section 4.3) [53] when they appear as inputs to the network G, as well as adding some noise to the input data (in all the networks) [53]. Finally, the last remedy that was utilized is the implementation of Adagrad [53] as the optimizer for all networks of the dacssGAN architecture.

3.3. Calculation of spectrograms

Motivated by several works dealing with audio classification tasks [55] [56] [57], it was decided to extract and make use of spectrogram representation (instead of making use of the raw audio signals for the target domain X_T) for the feature extraction process. A spectrogram is a visual depiction of the spectrum of frequencies from a signal (audio signals in our case) and its fluctuations over time. In that manner, the whole approach of knowledge transfer could be transformed into an image-to-image translation task, that will also make easier the implementation of our GAN architecture. Additionally, it also facilitates the qualitative inspection of the generated spectrograms produced during experimental results. The whole process of extracting spectrograms given raw audio is detailed in [58].

Finally, the spectrograms extracted from audio signals were narrowed using a fixed size (with three alternative configurations) of 28×28, 56×28 or 112×28. That fixed size was established by always starting from the middle part of the spectrogram and symmetrically keeping the surrounding region (that roughly corresponds to auditory information of 0.2, 0.5 and 1 second respectively) in order to avoid initial of final silence appearing in some file. In Figure 5a, samples of spectrograms extracted from the CREMA-D dataset are visualized.



(a) Samples of extracted spectrograms.

(b) Samples of aligned and cropped faces.

Figure 5: Real samples derived from the CREMA-D.

3.4. Face cropping and alignment

As described in Section 4.1, in this work we made use of CREMA-D dataset. Using the data included in this dataset, the following strategy to extract the facial features was employed. Firstly, faces were tracked and facial landmarks were obtained using the Ensemble of Regression Trees (ERT) method described in [59]. ERT provides efficient and precise landmark positions in demanding settings, such as varying illumination and poses, with solid and high performing tracking in real-time. Using the face tracking, each face was aligned by registering it in reference to pre-defined facial landmarks, namely, eyes centers, nose, mouth, and chin, to a canonical frame through a similarity transformation [60]. Facial images were cropped and re-sized to a fixed resolution that was chosen to be 28×28. In Figure 5b, instances of the CREMA-D database after the whole cropping and alignment processes are shown.

4. Experimental phase

4.1. Datasets

In this sub-section, the datasets that were utilized for evaluating the framework by training the networks G, D and Q as well as imposing the class-related information of emotional context that governs audio-visual data are presented. In this regard, the whole architecture was tuned by making use of the CREMA-D [61] and RAVDESS [62] datasets in two independent evaluation procedures. 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 data samples for each class. Both datasets were split into four different sets.

- Insofar as the first is concerned (this set was denoted as S_1), was mainly utilized in order to train the classifier from the source domain $f_S(x_S)$ to perform facial expression recognition, with the purpose of using it for the semi-supervised GANs.
- Secondly, the set S_2 was used with a view to calibrating the classifier by calculating the p-values of conformal prediction (Equation 9).
- Thirdly, the samples in a set denoted as S_3 were used for retrieving their p-values and using them for training the dacssGAN architecture (networks G, D and Q).
- Finally, the rest of the subjects, grouped as the set S_4 were used for the testing of the whole approach.

4.1.1. CREMA-D

CREMA-D is a large-scale multimodal emotion expression dataset made public recently on GitHub[1]. It encompasses 7442 videos from 91 actors (43 females and 48 males). Their age varies between 20 and 74 and they stem from a diversity of races and ethnicities (African American, Asian, Caucasian, Hispanic). Actors were requested to pose 12 sentences that are associated with six different emotions (anger, disgust, fear, happy, neutral, and sad) with

[1] github.com/CheyneyComputerScience/CREMA-D

four different levels of intensity (Low, Medium, High and Unspecified). The dataset's annotations are based on the presented videos that were shown up to the participants. The actors that participate in the CREMA-D dataset were requested to rate the emotions and their levels based on the combined audiovisual presentation, video only, and audio only. Each participant rated 90 unique clips, 30 audio, 30 visual and 30 audio-visual. 95% of the videos have, at least, 8 ratings. The dataset is a result of an effort to generate standard emotional stimuli for neuro-imaging studies which require a wide range of intensity and separation for visual and auditory modalities presentation.

4.1.2. RAVDESS

RAVDESS is an audio-visual affect-related (derived from speech and song segments) dataset recently made public[2]. The database is gender balanced consisting of 24 actors and it includes posed emotions regarding speech (including calm, happy, sad, angry, fearful, surprise, and disgust) and songs (with calm, happy, sad, angry, and fearful labels). In our experiments, we made use only the speech segments. Each expression is produced at two intensity levels. All cases are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. An additional set of 72 individuals provided test-retest data.

4.2. Metrics for evaluating generated samples

Three different metrics have been applied in an attempt to evaluate the quality of the generated samples. Firstly, the classification performance of the data augmentation schema was assessed as the first metric. In that case, we fused real samples from the original dataset with the ones generated from our architecture and we extracted the classification performance of the expanded set. This was done with the aim of testing whether the generated spectrograms encompass efficiently the emotion recognition performance and whether can improve the classification performance of the initial real dataset. In this sense, we can perceive data augmentation as an affordable alternative to easily expand audio-related datasets.

Secondly, in an attempt to evaluate the quality of the generated samples, the Inception Score (IS) [63] was utilized as an evaluation measurement. The approach made use of an Inception network pre-trained on performing emotion recognition in real spectrogram datasets. This pre-trained model was applied to the generated samples in an effort to compare the conditional label distribution with the marginal label distribution. The score is measured based on two criteria: 1) whether the generated images have diversity, and 2) whether the generated images have good quality.

This idea could be framed with the following equation:

$$IS(x) = \exp(E_x[KL(p(y|x)p(y))]) \tag{10}$$

where x is a generated sample and $p(y|x)$ represent the distribution of classes for this sample. We want $p(y|x)$ to be highly predictable so to have low entropy. Furthermore, $p(y)$ is the overall distribution of classes across the sampled data and should have a high entropy which means the absence of dominating classes and a well-balanced training set. Altogether, the higher the IS is the better the quality of the generated samples.

Finally, another qualitative metric, the Fréchet Inception Distance (FID) [64] was employed. This approach compares the statistics of generated samples to real ones, instead of only evaluating generated ones. This approach is based on the same Inception model (previously used for the IS) and it is applied in the generated and real images to calculate the prediction using the Inception network. In more detail, FID could be framed as:

$$FID(\mathcal{X}_R, \mathcal{X}_G) = \|\mu_R - \mu_G\| + Tr(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{(\frac{1}{2})}) \tag{11}$$

where $\mathcal{X}_R$ and $\mathcal{X}_G$ are distributions of real and generated images (after the utilization of the Inception network) respectively and $\mu_{R,G}$ and $\Sigma_{R,G}$ correspond to mean and covariance of the real and generated datasets respectively. Lower FID values mean better image quality and diversity.

[2]https://zenodo.org/record/1188976#.XHPO8-hKi70

4.3. Experimental protocol

In the experimental phase, the proposed architecture was evaluated in several steps. The objective of the current work was to evaluate the capacity and the amount of knowledge transferred between the source and target domains as well as, to inspect the quality of the generated samples. For that reason we employed the metrics described in the previous sub-section for each step of the experimental protocol.

Primarily, the classification performance of audio ($f_T(X_T)$) and visual domains ($f_S(X_S)$) was established as the baseline in the evaluated datasets accordingly without utilizing any domain adaptation strategy. The classifier employed to establish this baseline (for both modalities, audio and face) is a network similar to the one used as part of the dacssGAN and represented in Figure 3 as "Classifier". The only difference between the audio and face cases was one extra convolutional layer that was added in an effort to tackle the different sizes of the two domains. For the training of this baseline for face and audio, the S_1 set was utilized while for the validation of the classifier, we made use of the S_4 set (see Section 4.1 for further details). As a preliminary step, in the case of audio, three different experiments were performed, evaluating three spectrograms sizes: $28 \times 28, 56 \times 28$ and 112×28 (in pixels). In this analysis, it was found that the best results were obtained in the case of spectrograms with a resolution of 112×28. From this point on, the results for the rest of the experiments in this paper are referring to that case.

In this phase, the classification performance using the CNN classifier was established as $49,34\%$ and $64,50\%$ for the audio and face modalities respectively for the CREMA-D dataset while for the RAVDESS dataset these values were established as $46,28\%$ and $59,79\%$ respectively. Those aforesaid results will be noted henceforth as the **baseline** scores for the whole evaluation schema. The next step of our evaluation schema was to train the whole GAN architecture (shown in Figure 3) by making use of the available training S_3 set. To that end, three different approaches were considered:

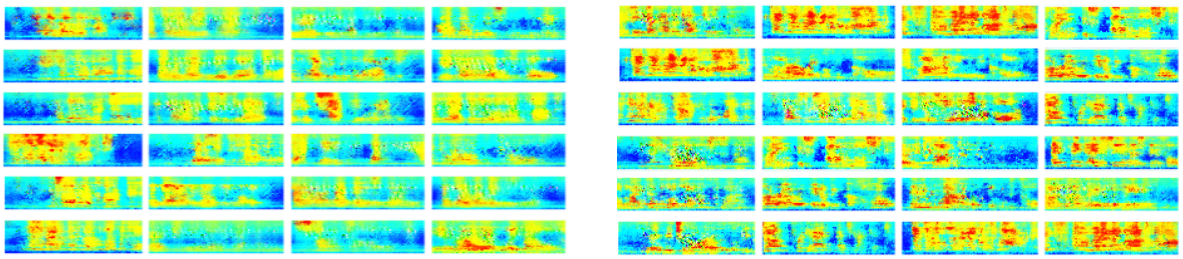
- **Supervised conditional GAN**: In the first evaluated approach we have as input to the network G together with the samples from the target domain and a noise vector z the conditional information c that is represented from the label information that comes together with our datasets (groundtruth).
- **Semi-supervised conditional GAN**: In this case, we explored the possibility of replacing the label information that was given as input to the generation G with the output of a classifier $f_S(X_S)$ trained using S_1 to perform emotion recognition on the target domain. The output of that classifier was a six-featured vector that contains, in each feature, the probability of the input sample to be derived from a specific emotional label.
- **Semi-supervised conditional CP GAN**: Finally, in that case, we explored the possibility of replacing that classification conditionality by the calibrated version that could be provided using inductive conformal prediction. As it was already mentioned, for calculating the inductive CP we made use of the set denoted as S_2 as a validation bucket that helped us to calculate the p-values for the denoted set S_3 that played the role of the conditional information.

4.4. Ablation study

Over the above-mentioned evaluation approaches, an ablation study was performed by evaluating the performance of different architectures for the generator network G, different loss functions, different input for the network G, the different sparsity levels concerning the data availability in the target domain and different algorithms for conformal predictions.

Insofar as the first is concerned, a deep convolutional U-Net and a structure that resembles an encoder and decoder approach (EncDec) using merely Dense layers were evaluated as the possible architectures for the network G. Figures 4a and 4b depict these two architectures for the network G respectively. Table 1 shows that U-Net was proven incapable to outperform the simplified (EncDec) architecture. Additionally, L1 norm was proven to be crucial for the outcome of the dacssGAN (as displayed in columns 2 and 4). When this part from the optimization function was omitted, it was observed that the results were deteriorating not only classification-wise but also and most notably regarding the forfeiting of visual fidelity of the generated images. A possible explanation is that pixel to pixel distance as a loss helped the calibration of network G and force the generated distributions to be closer to the real ones.

Moreover, we performed some further experiments wherein we neglected the source domain $x \in X_S$ from the input to the network G. These experiments were conducted, in an effort to evaluate the importance of the source



(a) Supervised conditional GAN. (b) Semi-supervised conditional GAN.

Figure 6: Generated spectrograms of the GAN approach calibrated using the CREMA-D dataset.

domain $x \in X_S$ as input to the network G. In this phase, we also evaluated the three experimental cases (supervised, semi-supervised conditional GANs and semi-supervised conditional CP GAN). In particular, for these experiments we made use of a different EncDec and U-Net architectures, where we dropped the encoding part in both cases together with the input $x \in X_S$. In all the cases, the results were inferior in comparison with all the aforesaid DA cases and these architectures were therefore rejected.

Additionally, we decided to monitor the performance of the whole approach by having different sparsity availability in the target domain. In that case we decided to make use of the 50% and 20% of the initial datasets and extract the results for the Semi-supervised conditional CP GAN case. In the course of this experiment we wanted to determine how crucial is the availability of data for the proper training of the whole GAN approach. That study is explained in more details in the sub-section 4.6.

Finally, it is important to note here that for the case of conformal prediction we validated both the initial algorithm that is explained in Algorithm 1 and the inductive conformal prediction. It was found that the performance of both techniques was similar, however, the computation complexity of the initial version of conformal prediction in comparison to inductive conformal prediction was remarkable. Therefore, it was decided to stick with the results of inductive conformal prediction and introduce it as our semi-supervised technique. Furthermore, since the classification performance of the inductive conformal prediction had similar results with the results using random split for the training and calibration sets, it was decided that it is not necessary to employ a more sophisticated way to split the sets.

4.5. Experimental results and discussion

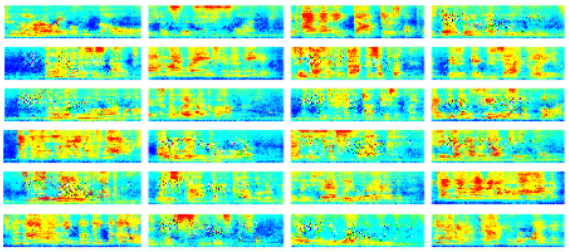
In this section, the results obtained during the evaluation of all the cases of the experimental protocol which presented in Section 4 are described. In Table 2 the performance regarding all three metrics of all the aforementioned cases from experimental protocol is presented. Firstly, in the initial GAN case, the so-called supervised conditional GAN was evaluated. Figure 6a represents some generated spectrograms derived from this approach. In that case, the approach reached the best performance (52.52% for CREMA-D and 47.11% for RAVDESS). The same behaviour was implied for the qualitative results based on the IS and FID metrics (see Table 2).

However, in the current work our main effort was focused on the much more interesting semi-supervised case where the goal is to generate annotated audio samples coming from rich but not necessarily annotated video samples. As a subsequent step, the semi-supervised conditional GAN was evaluated. Figure 6b represents the generated spectrograms that derived from this approach. In that case, the obtained results were 49.92% for CREMA-D and 46.23% for RAVDESS.

Finally, the evaluation of the semi-supervised conditional CP GAN was conducted. The extracted results were 50.29% for CREMA-D and 46.55% for RAVDESS. The rationale behind that improvement in the results (in contrast to the previous case) was mainly that, after the application of conformal prediction, the calculated p-values contain better-distributed confidences in the rest of the labels in comparison to confidences derived using merely the classifier. In Figure 7a, the results in that scenario (where conformal prediction is used as part of the input to the networks G) are visualized. From Figures 6a, 6b and 7a that represent the generated spectrograms for all three cases, we can deduce that the dacssGAN approach in all three steps managed to approximate well the target domain in each case while the visual results can be considered faithful representations of the target distribution domain.

Table 1: The classification performances of the target domain task T_T for the performed ablation study.

Baseline	49,34%		
Case	EncDec	U-NET	EncDec_without L1
Supervised GANs	52.52%	50.24%	38.11%
Semi-Supervised GANs-classifier	49.92%	50.12%	31.67%
Semi-Supervised GANs-CP	50.29%	50.09%	31.69%



(a) Generated spectrograms for the semi-supervised CP GAN.

(b) Generated faces for the supervised GAN case.

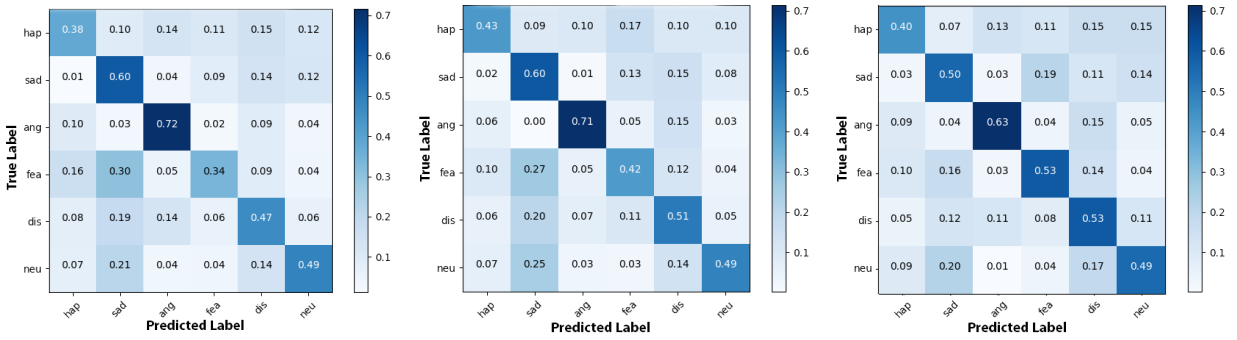
Figure 7: Generated samples of the GAN approach that calibrated using CREMA-D dataset.

Supplementary to the previous conducted experiments, for validation purposes, it was decided to utilize the whole dacssGAN architecture in a vice versa manner, in order to generate faces given as an input noise vector ($z \in Z$) fused with the samples from source domain ($x \in X_S$, that in that case are represented from audio domain). Primarily, we want to emphasize that our approach, although intended to expand an audio-related dataset by using facial information, it is not limited to this objective and can be easily modified to address slightly different approaches. This experiment, was not the core objective of the current work. Therefore, only elemental experiments (with CREMA-D) and some basic hyper-parameter tuning for the supervised version of GANs were conducted. We believe that there is a significant room for further improvement, classification-wise, as well as regarding the visual results and could be very interesting research for future work. However, still the visual results were worth presenting due to the capacity of the approach not only to approximate sufficiently the face distribution, but further, to generate faces with immense emotion context. In Figure 7b the generated images that are produced from the network G in that case are portrayed.

Table 2: Classification, FID and IS performance for all the experimental protocol scenarios.

Case	CREMA-D			RAVDESS		
	Classification	FID	IS	Classification	FID	IS
Baseline	49.34%			44.73%		
Supervised GANs	**52.52%**	**59.44**	**2.16**	**47.11%**	**49.77**	**2.21**
Semi-Supervised GANs-classifier	49.92%	60.13	2.01	46.23%	50.33	2.05
Semi-Supervised GANs-CP	50.29%	60.10	2.00	46.55%	49.95	2.01

Further stimulating observations can be found in the confusion matrices (CM) extracted when using CREMA-D dataset for the baseline case (displayed in Figure 8a), supervised GAN approach (Figure 8b) and semi-supervised CP (Figure 8c). Firstly, all figures show that the considered emotions are well discriminated, since the diagonal elements of the matrix have (in all cases) the highest classification performances. Also, in all three cases it is evident that the strongest captured emotion is anger. It is noteworthy that this behaviour is consistent with the study performed in [61] concerning the human accuracy in audio emotion recognition, where it was stated that the best performed emotion label was anger for the CREMA-D dataset. In [65] authors observed a similar finding when using other state of the art audio-based datasets. In the case of the data augmentation schema of the supervised case, in Figure 8b we can



(a) Baseline. (b) Supervised approach. (c) Semi-supervised CP approach.

Figure 8: Confusion matrix for the emotion recognition classifier through audio spectrograms using different approaches for the six distinct emotion of CREMA-D dataset.

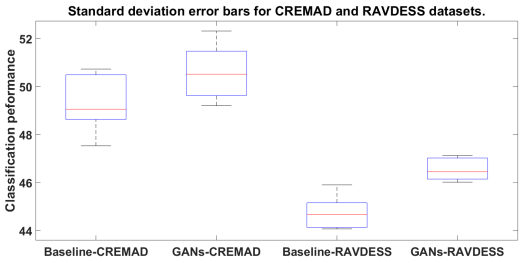


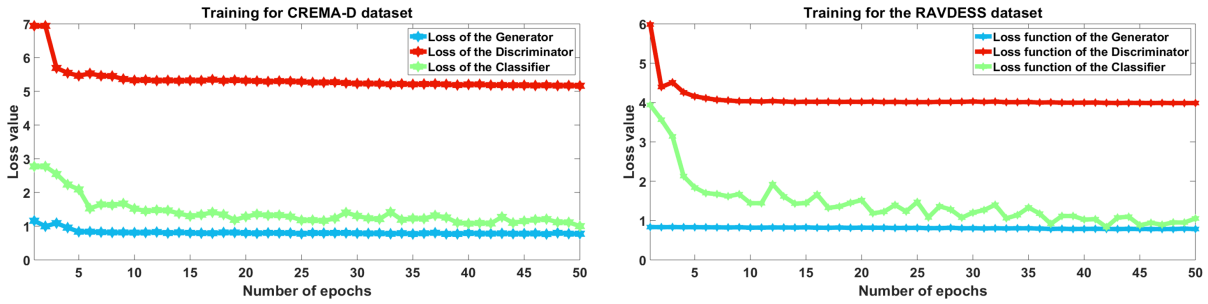
Figure 9: Std error bars for baseline and the Semi-supervised CP.

observe that while the anger emotion performance drops, the efficiency of the rest of emotions roughly increases and thus, a more smooth allocation of the emotion recognition is achieved.

In an attempt to better understand the performance of the proposed semi-supervised CP GAN in comparison with the baseline a statistical analysis was performed. From this analysis, the statistical results (mean and standard deviation) for the classification performance were extracted from different folds (each fold contain different subject of dataset). The results obtained for both datasets are illustrated in Figure 9. From this figure we can observe that, in the case of RAVDESS, the standard deviation is narrow and there exists a significant different in the mean value. In the case of CREMA-D, the deviation from the mean is higher which could be attributed to the bigger variety of the subjects included in the dataset.

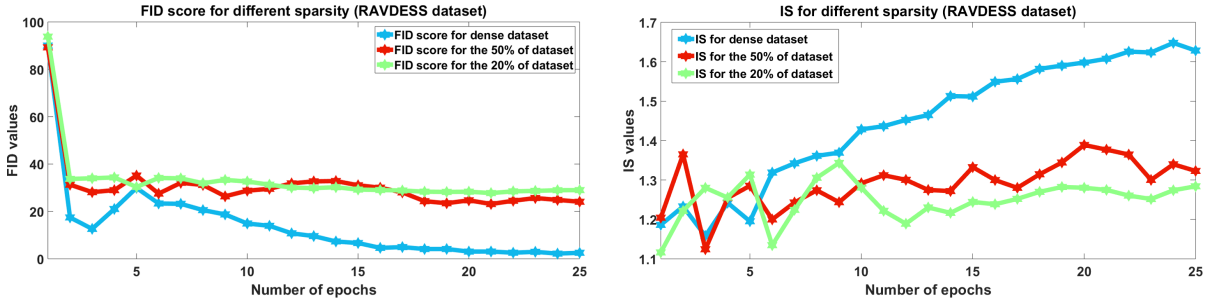
4.6. Training process evaluation

In order to examine the performance of the training process of our dacssGANs algorithm, we visualize the loss function of the G, D and Q networks, as well as, IS and FID score during the training procedure. Figures 10a and 10b render the loss function of the three networks for the CREMA-D and RAVDESS datasets respectively while Figures 11a and 11b display the IS and FID scores for RAVDESS dataset and also the evaluation performance for the sparsity test. From these figures, we can deduce that during training even by using 50% of the dataset leads to noticeably poorer results in the qualitative performance. On the whole, an evident observation during the training procedure was that the approach was steadily converging and the quality of the visual results were improving with the increasing number of epochs. Finally, regarding the complexity of our approach, all the conducted experiments were performed by using a NVIDIA Titan X graphic card. For a single experiment, the total time duration was approximately 26 and 48 hours (for RAVDESS and CREMA-D respectively).



(a) CREMA-D dataset. (b) RAVDESS dataset.

Figure 10: Loss function during the training procedure of dacssGANs CP.



(a) Féchet Inception Distance. (b) Inception Score.

Figure 11: FID and IS values during the training procedure of dacssGANs CP.

4.7. Limitations

In the context of the current work, we applied a trained classifier for extracting predictions from the source domain samples (also for calculating the CP values). That classifier was trained by using data samples with a distribution similar to the real source domain samples used for training the network G. During this process, the data samples are different subjects of the same dataset which were captured under similar background and illumination conditions. However, the availability of similar distributions is not always ensured. To tackle this challenge, several approaches can be considered. Firstly, we can make use of another dataset with similar distribution to train our algorithm to be applied as a face classifier. Another possible solution is to generate spectrograms only by using faces and noise, by neglecting the conditional information in the input. Both strategies require several new experiments for tuning the whole network hyper-parameters. This constraint is further affiliated with time limitation posed for the training of the whole network (that was described in the previous sub-section). Hence, our study consists of the most crucial experiments.

Another limitation is the employment of posed and in-lab environment datasets that contained acted emotions and not into the wild behaviours. We performed so in an effort to establish the reliability of the whole framework by studying the cross-modal relationships of the two modalities. As a further work we would like to study these relationships under uncontrolled environments.

5. Conclusion

In the current work, we investigated the research question of whether it is possible to transfer knowledge between facial expressions to the audio information from the same sequences for the purpose of expanding audio datasets for emotion recognition. For that purpose, we introduced a novel approach to study the cross-modal relationships between audio-visual modalities, called domain adaptation conditional semi-supervised Generative Adversarial Networks or dacssGAN. The core objective of the approach was to implement a network G which will generate samples that are

distributed from the target domain (audio domain in the presented case) and represent specific emotion states. The input to network G is a random noise vector $z \in Z$, fused together with samples from the source domain ($x \sim X_S$) and with conditional information (c). That conditional information was calculated using a semi-unsupervised technique called conformal Prediction. We propose the use of these confidence values, instead of labels, as a softer and more reliable manner to introduce knowledge into the generator. Furthermore, we investigated the efficiency of a network Q that works as a classifier in the target domain and calibrates the generated samples from network G. The efficiency of our approach was established during the experimental phase by employing a data augmentation schema where it was proven that our posed hypothesis is valid and facial expression could be efficiently used and fused together with the audio information in order to generate samples for audio domain that can help improving classification results in the target domain.

Acknowledgements

This work was supported by the Horizon 2020 funded project MaTHiSiS (Managing Affective-learning THrough Intelligent atoms and Smart InteractionS) nr. 687772 (http://www.mathisis-project.eu/). Furthermore, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Nvidia GeForce GTX TITAN X GPU used throughout the experimental phase.

References

[1] T.Grossman, The development of emotion perception in face and voice during infancy, Resorative Neurology and Neuroscience 28, Pages: 219–236, 2010.
[2] D.W.Massaro and M.M.Cohen, Perceiving Talking Faces, Current Directions in Psychological Science, Volume 4, Number 4, 1995.
[3] C.Bayard, C.Colin and J.Leybaert, How is the McGurk effect modulated by Cued Speech in deaf and hearing adults?, Frontiers in Psychology, Volume 5, Number 416, 2014.
[4] E.Cvejic, J.Kim and C.Davis, Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion, Journal of Speech Communication, Volume 52, Number 6, Pages: 555-564, 2010.
[5] M.Swerts and E.Krahmer, Facial expression and prosodic prominence: Effects of modality and facial area, Journal of Phonetics, Volume 36, Number 2, Pages: 219–238, 2008.
[6] M.D.Pell, Prosodyface Interactions in Emotional Processing as Revealed by the Facial Affect Decision Task, Journal of Nonverbal Behavior, Volume 29, Number 4, Pages: 193-215, 2005.
[7] C.A.Corneanu, M.Oliu, J.F.Cohn and S.Escalera, A Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition:History, Trends, and Affect related Applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 38, Number 8, Pages: 1548–1568, 2016
[8] J.Kumari, R.Rajesh and K.M.Pooja, Facial Expression Recognition: A Survey, Procedia Computer Science Volume 58, Pages 486-491, 2015.
[9] P.Liu, S.Han, Z.Meng and Y.Tong, Facial Expression Recognition via a Boosted Deep Belief Network, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[10] S.Albanie, A.Nagrani, A.Vedaldi and A.Zisserman, Emotion Recognition in Speech using Cross-Modal Transfer in the Wild, ACM Multimedia 2018.
[11] K.Weiss, T.M.Khoshgoftaar and D.D. Wang, A survey of transfer learning, Journal of Big Data, Volume 4, Number 1, 2017.
[12] S.J. Pan and Q. Yang, A Survey on Transfer Learning, IEEE Transactions on Knowledge and Data Engineering, Volume 22, Number 10, Pages: 1345–1359, 2009.
[13] I.J.Goodfellow, J.P.-Abadie, M.Mirza, B.Xu, D.W.Farley, S. Ozair, A.Courville and Y. Bengio, Generative Adversarial Networks, 27th conference on Advances in Neural Information Processing Systems (NIPS), 2014.
[14] A.Radford, L.Metz and S.Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, International Conference on Learning Representations (ICLR), 2016.
[15] M.Mirza and S.Osindero, Conditional Generative Adversarial Nets, Computing Research Repository (CoRR), 2014.
[16] O.Ronneberger, P.Fischer and T.Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.
[17] P.Isola, J.Yan, Z.Tinghui, Z.Alexei and A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[18] G.Shafer and V.Vovk, A Tutorial on Conformal Prediction, The Journal of Machine Learning Research archive Volume 9, Pages 371-421, 2008.
[19] V.Vovk, Conformal prediction, Algorithmic Learning in a Random World, Springer US, Pages: 17–51, 2005.
[20] S.Matiz, K.E.Barner, Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification, Volume 90, Pages: 172–182, 2019.
[21] X.Liu, Y.Zou, L.Kong, Z.Diao, J.Yan, J.Wang, S.Li, P.Jia and J. You, Data Augmentation via Latent Space Interpolation for Image Classification, International Conference on Pattern Recognition (ICPR), 2018.
[22] S.Sahu, R.Gupta and.E.Wilson, On Enhancing Speech Emotion Recognition using Generative Adversarial Networks, Interspeech, 2018.

[23] G.Csurka, Domain Adaptation for Visual Applications: A Comprehensive Survey, Domain Adaptation in Computer Vision Applications, Pages: 1–35, 2017.
[24] M.Wang, W.Deng, Deep Visual Domain Adaptation: A Survey, Neurocomputing 2018.
[25] Y.Ganin, E.Ustinova, H.Ajakan, P.Germain, H.Larochelle, F.Laviolette, M.Marchand and V.Lempitsky, Domain-Adversarial Training of Neural Networks, Journal of Machine Learning Research, Volume 17, Pages: 1–35, 2016.
[26] S.Purushotham, W.Carvalho, T.Nilanon and Y.Liu, Variational Recurrent Adversarial Deep Domain Adaptation, International Conference on Learning Representations (ICLR), 2017.
[27] L. Qin, Z. Zhang, H. Zhao, Z. Hu and E.P. Xing, Adversarial Connective-exploiting Networks for Implicit Discource Relation Classification, Conference: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017.
[28] F.Qi, X.Yang and C.Xu, A Unified Framework for Multimodal Domain Adaptation, Proceedings of the 26th ACM international conference on Multimedia, 2018
[29] D. Das. C.S.G. Lee, Sample-to-sample correspondence for unsupervised domain adaptation, Journal for Engineering Applications of Artificial Intelligence, Volume 73, Pages: 80–91, 2018.
[30] D. Das. C.S.G. Lee, Unsupervised domain adaptation using regularized hyper-graph matching, International Conference on Image Processing, 2018.
[31] N. Courty, R. Flamary, D. Tuia and Alain Rakotomamonjy, Optimal Transport for Domain Adaptation, IEEE Transactions of Pattern Analysis and Machine Intelligence, 2016.
[32] D.Das and S.G.Lee, Graph Matching and Pseudo-Label Guided Deep Unsupervised Domain Adaptation, Graph Matching and Pseudo-Label Guided Deep Unsupervised Domain Adaptation, 2018.
[33] J.Zhu, T.Park, P.Isola and A.A.Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, IEEE International Conference on Computer Vision (ICCV), 2017.
[34] T.Kim, M.Cha, H.Kim, J.K.Lee and J.Kim, Learning to Discover Cross-Domain Relations with Generative Adversarial Networks, International Conference on Machine Learning (ICML), 2017.
[35] J.Hoffman, E.Tzeng, T.Park, J.Y.Zhu, P.Isola, K.Saenko, A.Efros and T.Darrell, CyCADA: Cycle-Consistent Adversarial Domain Adaptation, International Conference on Learning Representations (ICLR), 2018.
[36] Z.Hu, Z.Yang, R.Salakhutdinov, E.P.Xing, On Unifying Deep Generative Models, International Conference on Learning Representations (ICLR), 2018.
[37] 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.
[38] W.Hao, Z.Zhang and H.Guan, CMCGAN: A Uniform Framework for Cross-Modal Visual-Audio Mutual Generation, Association for the Advancement of Artificial Intelligence (AAAI), 2018.
[39] J. S. Chung and A. Zisserman, Out of time: automated lip sync in the wild, Workshop on Multi-view Lip-reading, ACCV, 2016.
[40] F.Ringeval, S.Amiriparian, F.Eyben, K.Scherer and B.Schuller, Emotion Recognition in the Wild: Incorporating Voice and Lip Activity in Multimodal Decision-Level Fusion, Proceedings of the 16th International Conference on Multimodal Interaction ICMI, 2014.
[41] P.Buitelaar, I.D.Wood, S.Negi, M.Arcan, J.P.McCrae, A.Abele, C.Robin, V.Andryushechkin, H.Ziad, H.Sagha, M.Schmitt, B.W.Schuller, S.Fernando, C.A.Iglesias, C.Navarro, A.Giefer, N.Heise, V.Masucci, F.A.Danza, C.Caterino, P.Smrz, M.Hradis, F.Povolny, M.Klimes, P.Matejka and G.Tummarello, 2018, MixedEmotions: An Open-Source Toolbox for Multimodal Emotion Analysis, IEEE TRANSACTIONS ON MULTIMEDIA, Volume 20, Pages: 2454–2465, 2018.
[42] 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.
[43] H.Papadopoulos, Inductive Conformal Prediction: Theory and Application to Neural Networks, Tools in Artificial Intelligence, 2008.
[44] J.Sun, L.Carlsson, E.Ahlberg, U.Norinder, O.Engkvist and H.Chen, Applying Mondrian Cross-Conformal Prediction To Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets, in Journal of Chemical information and modelling, Volume 57, Number 7, Pages: 1591-1598, 2017.
[45] X.Wang and A.Gupta. Generative image modeling using style and structure adversarial networks, European COnfeence on Computer Vision (ECCV), 2016.
[46] D.Pathak, P.Krahenbuhl, J.Donahue, T.Darrell, and A.A. Efros. Context encoders: Feature learning by inpainting, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[47] A.Odena, C.Olah and J.Shlens, Conditional Image Synthesis with Auxiliary Classifier GANs, Proceedings of the 34th International Conference on Machine Learning, (PMLR), 2017.
[48] M.Lee and J.Seok, Controllable Generative Adversarial Network, Published in ArXiv, 2017.
[49] V.N.Balasubramanian, A.Baker, M.Yanez, S. Chakraborty, and Sethuraman Panchanathan, PyCP: An Open-Source Conformal Predictions Toolkit.
[50] L.Metz, B.Poole, D.Pfau and J.S.Dickstein, Unrolled Generative Adversarial Networks, International Conference on Learning Representations (ICLR), 2017.
[51] I.Goodfellow, NIPS 2016 Tutorial: Generative Adversarial Networks, Conference on Neural Information Processing Systems (NIPS), 2016.
[52] T.Salimans, I.Goodfellow, W.Zaremba, V.Cheung, A.Radford and X.Chen, Improved Techniques for Training GANs, Conference on Neural Information Processing Systems (NIPS), 2016.
[53] I.Tolstikhin, S.Gelly, O.Bousquet, C.Johann, S.Gabriel and B.Schlkopf, AdaGAN: Boosting Generative Models, Advances in Neural Information Processing Systems (NIPS), 2017.
[54] K.He, X.Zhang, S.Ren and J.Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, IEEE Computer Vision and Pattern Recognition (CVPR), 2015.
[55] C.Donahue, J.McAuley and M.Puckette, Adversarial Audio Synthesis, International Conference on Learning Representations (ICLR), 2019.
[56] C.Donahue, J.McAuley and M.Puckette, Synthesizing Audio with GANs, International Conference on Learning Representations (ICLR), 2018.

[57] P.Yenigalla, A.Kumar, S.Tripathi, C.Singh, S.Kar and J.Vepa, Speech Emotion Recognition Using Spectrogram & Phoneme Embedding, in Interspeech, 2018
[58] H.M.Fayek, M.Lech and L.Cavedon, Towards Real-time Speech Emotion Recognition using Deep Neural Networks, 9th International Conference on Signal Processing and Communication Systems (ICSPCS), 2015.
[59] V.Kazemi and J.Sullivan. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[60] X.Xiong and F.D.Torre. Supervised descent method and its applications to face alignment. In Computer Vision and Pattern Recognition (CVPR), 2013.
[61] 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.
[62] 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.
[63] T.Salimans, I.Goodfellow, W.Zaremba, V.Cheung, A.Radford and X.Chen, Improved Techniques for Training GANs, In Advances in Neural Information Processing Systems (NIPS), 2016.
[64] M.Heusel, H.Ramsauer, T.Unterthiner, B.Nessler and S.Hochreiter, Gans trained by a two time-scale update rule converge to a local Nash equilibrium, In Advances in Neural Information Processing Systems (NIPS), (2017).
[65] E.Ghaleb, M.Popa, E.Hortal and S.Asteriadis, Multimodal Fusion Based on Information Gain for Emotion Recognition in the Wild, IEEE Intelligent Systems Conference 2017.