IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Generative Adversarial Networks in Human Emotion Synthesis: A Review

NOUSHIN HAJAROLASVADI[1], MIGUEL ARJONA RAMÍREZ[2], (Senior Member, IEEE), WESLEY BECCARO[2], and HASAN DEMIREL.[1], (Senior Member, IEEE)

[1]Eastern Mediterranean University, Electrical and Electronic Engineering Department, Gazimagusa, 10 Via Mersin, Turkey
[2]University of São Paulo, Escola Politécnica, Department of Electronic Systems Engineering, São Paulo, Brazil

Corresponding author: Miguel Arjona Ramírez (e-mail: miguel@lps.usp.br).

ABSTRACT Deep generative models have become an emerging topic in various research areas like computer vision and signal processing. These models allow synthesizing realistic data samples that are of great value for both academic and industrial communities Affective computing, a topic of a broad interest in computer vision society, has been no exception and has benefited from this powerful approach. In fact, affective computing observed a rapid derivation of generative models during the last two decades. Applications of such models include but are not limited to emotion recognition and classification, unimodal emotion synthesis, and cross-modal emotion synthesis. As a result, we conducted a comprehensive survey of recent advances in human emotion synthesis by studying available databases, advantages, and disadvantages of the generative models along with the related training strategies considering two principal human communication modalities, namely audio and video. In this context, facial expression synthesis, speech emotion synthesis, and the audio-visual (cross-modal) emotion synthesis are reviewed extensively under different application scenarios. Gradually, we discuss open research problems to push the boundaries of this research area for future works. As conclusions, we indicate common problems that can be explored from the GAN topologies and applications in emotion synthesis.

INDEX TERMS Machine Learning, Generative Adversarial Networks, Learning Systems, Emotion Recognition, Speech Synthesis, Image Processing.

I. INTRODUCTION

DEEP learning techniques are known best for their promising success in uncovering the underlying probability distributions over various data types in the field of artificial intelligence. Some of these data types are videos, images, audio samples, biological signals, and natural language corpora. The success of the deep discriminative models owes primarily to the backpropagation algorithm and piece-wise linear units [1, 2]. In contrast, deep generative models [3] have been less successful in addressing deep learning applications due to difficulties that arise by intractable approximation in the probabilistic computation of methods like maximum likelihood estimation.

A specific type of neural network called Generative Adversarial Networks (GAN) models was introduced in 2014 by Goodfellow et al. [3]. These models are composed of a generative model pitting against an adversary model as a two-player minimax framework. The generative model captures data distribution. Then, given a sample, the adversary or the discriminator decides if the sample is drawn from the true data distribution (real) or from the model distribution (fake). The competition continues until the generated samples are indistinguishable from the genuine ones.

Many reviews studied the rapidly expanding topic of generative models and specifically GAN by investigating various points of view from algorithms, theory, applications [4, 5], to recent advances and developments [6, 7], comparative studies [8], GAN taxonomies [9], and its variants [10, 11, 12, 13]. Also, few review papers discussed the subject based on a specific application such as medical imaging [14], audio enhancement and synthesis [15], image synthesis [16], and text synthesis[17]. Howsoever, none of the existing surveys considered GAN in view of human emotion synthesis.

It is important to note that searching the scholarly literature "Generative Adversarial Network" on Web Of Science (WOS) and Scopus repositories, one can find that 2538 and 4405 documents published, respectively starting from 2014 up to present. The large number of researches published on this topic within only 6 years inspired us to conduct a comprehensive review considering one of the significant applications of GAN models called human emotion synthesis.

Humans communicate through various verbal and nonverbal channels to show their emotional state. All of the communication modalities are of high importance once interpreting the current emotional state of the user benefits from human emotion synthesis and data augmentation. Throughout this paper, we concentrate on the recent advances in the field of GAN and their possible acquisition of human emotion recognition which is known to be useful in other research areas like computer-aided diagnosis systems, security and identity verification, multimedia tagging systems, and human-computer and human-robot interactions. Humans communicate through various verbal and nonverbal channels to show their emotional state. All of the communication modalities are of high importance once interpreting the current emotional state of the user. We focus on the GAN-related works of speech emotion synthesis, face emotion synthesis, and audio-visual (cross-modal) emotion synthesis because face and speech are known as pioneer communication channels among humans [18, 19, 20, 21]. Researchers developed many GAN-based models to address problems such as data augmentation, improvement of emotion recognition rate, and enhancement of synthesized samples through unimodal [22, 23, 24, 25],[26, 27],[28, 29],[30],[31] and cross-modal analysis [32, 33, 34, 35].

This review deals with the GAN-based algorithms, theory, and applications in human emotion synthesis and recognition. The remainder of the paper is organized as follows: Section II provides a brief introduction to GANs and their variations. This is followed by a comprehensive review of related works on human emotion synthesis tasks using GANs in section III. This section covers unimodal and cross-modal GAN-based methods developed using audio/visual modalities. Section IV summarizes the review, identifies potential applications, and discusses challenges. Finally, section V concludes this survey.

II. BACKGROUND

In general, generative models can be categorized into explicit density models and implicit density models. While the former utilizes the true data distribution or its parameters to train the generative model, the latter generates sample instances without an explicit parameter assumption or direct estimation on real data distribution. Examples of explicit density modeling are maximum likelihood estimation and Markov Chain Method [36, 37]. GANs can be considered as implicit density modeling example [3].

A. GENERATIVE ADVERSARIAL NETWORKS (GAN)

Goodfellow et al. proposed Generative Adversarial Networks or vanilla GAN in 2014 [3]. The model works based on a two-player minimax game where one player seeks to maximize its cost function and the other seeks to minimize its own cost function. The game ends at a saddle point of the value function, where the first agent and the second agent reach a maximum and minimum of each cost function, respectively. This model draws samples directly from the desired distribution without explicitly modeling the underlying probability density function. The general framework of this model consists of two neural network components: a generative model G capturing the data distribution and a discriminative model D estimating the probability that a sample comes from the training samples or G.

Consider the input sample for G as $\boldsymbol{z}$, where $\boldsymbol{z}$ is a random noise vector sampled from a priori density $p_z(\boldsymbol{z})$, a real sample x_r that is taken from the data distribution P_r, and x_g output sample generated by G. Then, the idea is to get maximum similarity between the two samples. In fact, the generator G learns a nonlinear mapping function parametrized by θ_g and formulated as: $G(\boldsymbol{z};\theta_g)$. The discriminator D, gets both x_r and x_g to output a single scalar value $\mathcal{O}_1 = D(\boldsymbol{x};\theta_d)$ stating the probability of whether an input sample is a real or a generated sample [3]. It is important to highlight that $D(\boldsymbol{x};\theta_d)$ is the mapping function learned by D and parametrized by θ_d. The final probability density formed by generated samples is p_g and it is expected to approximate p_r after learning. Fig. 1(a) illustrates the general block diagram of the vanilla GAN model.

The traditional objective of generative modeling is related to maximizing likelihood, or minimizing the Kullback-Leibler (KL) divergence. The KL divergence allows quantifying the difference between two probability density functions for random variables ranging over $\mathbb{R}^D$. The KL divergence of p_r from p_g is defined for continuous distributions as (1):

$$\mathrm{KL}(p_r \parallel p_g) = \int_{\mathbb{R}^D} p_r(\boldsymbol{x}) \log\left(\frac{p_r(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right) d\boldsymbol{x}, \tag{1}$$

as long as $S_r \subseteq S_g$, where S_r and S_g are the support of the probability densities p_r and p_g, respectively, avoiding infinity when there are points such that $p_g(x) = 0$ and $p_r(x) > 0$ for $\mathrm{KL}(p_r \parallel p_g)$.

The KL divergence is asymmetric, i.e. $\mathrm{KL}(p_r \parallel p_g) \neq \mathrm{KL}(p_g \parallel p_r)$. A more convenient approach for GAN is the Jensen-Shannon (JS) divergence which may interpreted as a symmetrical version of KL divergence and it is defined as (2):

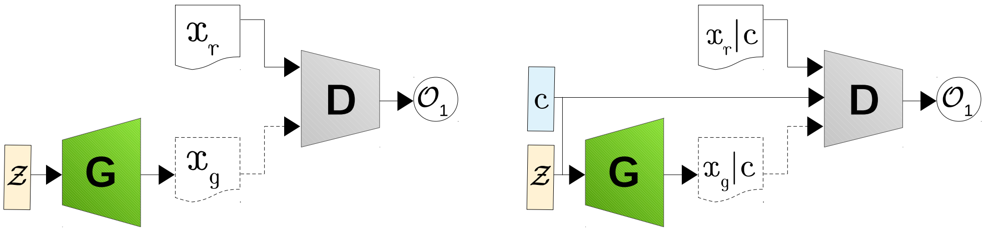
$$\mathrm{JS}(p_r, p_g) = \mathrm{KL}(p_r \parallel p_m) + \mathrm{KL}(p_g \parallel p_m), \tag{2}$$

where $p_m = (p_r + p_g)/2$.

In other words, a minimax game between G and D continues to obtain a normalized and symmetrical score in terms of the value function $V(G,D)$ with respect to a joint loss function for D and G as [3, 9]:

$$\min_G \max_D V(G,D) = \min_G \max_D \{\mathbb{E}_{\boldsymbol{x}_r \sim p_r(\boldsymbol{x})}[\log D(\boldsymbol{x}_r)] + \mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]\} \tag{3}$$

Here, the parameters of G are adjusted by minimizing $\log(1 - D(G(\boldsymbol{x}_g)))$. In a similar way, adjusting the parameters for D is performed by maximizing $\log D(\boldsymbol{x}_r)$. Minimizing $\log(1-D(G(\boldsymbol{x}_g)))$ is known [3] to be equivalent



(a) vanilla GAN [3]

(b) CGAN [38]

FIGURE 1: General Structure of vanilla GAN and CGAN models; $\mathcal{Z}$: input noise, **G**: **G**enerator, **D**: **D**iscriminator, $\boldsymbol{x}_r$: **r**eal sample, $\boldsymbol{x}_g$: **g**enerated sample, $\mathcal{O}_1$: **O**utput of binary classification to real/fake, c: **c**ondition vector.

to minimizing the JS divergence between P_r and P_g as expressed in Eq. (2). The value function $V(\theta_g, \theta_d)$ determines the payoff of the discriminator. Also, the generator takes the value $-V(\theta_g, \theta_d)$ as its own payoff. The generator and the discriminator, each attempts to maximize its own payoff [39] during the learning process.The general framework of this model is shown in Fig. 1(a).

B. CHALLENGES OF GAN MODELS

As pointed out before, the training objective of GAN models is often referred to as saddle point optimization problem [40] which is resolved by gradient-based methods. One challenge here is that D and G should be trained at a time so that they advance and converge together. Minimizing the generator's objective is proven to be equivalent to minimizing JS divergence if the discriminator D is trained to its optimal point before the next update of G. This means minimizing the JS divergence does not guarantee finding the equilibrium point between G and D through the training process. This normally leads to a better performance of D as opposed to G. Consequently, at some point classifying real and fake samples becomes such an easy task that gradients of D approach zero and it becomes ineffectual in the learning procedure of G. Mode collapse is another well-known problem in training GAN models where G produces a limited set of repetitive samples due to focusing on a few limited modes of the true data distribution, namely P_r, during learning and approximates distribution P_g. We discuss these problems in more detail in section IV-A.

C. VARIANTS BY ARCHITECTURES

The GAN model can be extended to a conditional GAN (CGAN) model if both the generator and discriminator are conditioned on some extra information $\boldsymbol{y}$ [38]. Fig. 1(b) shows the block diagram of the CGAN model. The condition vector $\boldsymbol{y} = c$ is fed into both the discriminator and the generator through an additional input layer. Here, the latent variable $\boldsymbol{z}$ with prior density $p_z(\boldsymbol{z})$ and condition vector $\boldsymbol{y}$ with some value $c \in \mathbb{R}^d$ are passed through one perceptron layer to learn the joint hidden representation. Conditioning on c changes the training criterion of Eq. (3) and leads to the following criterion (4):

$$\min_G \max_D \{\mathbb{E}_{\boldsymbol{x} \sim p_r(\boldsymbol{x})}[\log D(\boldsymbol{x}_r \mid \boldsymbol{y} = c)] \\ + \mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z} \mid \boldsymbol{y} = c)))]\}, \tag{4}$$

where c could be target class labels or auxiliary information from other modalities.

Another type of GAN model is the Laplacian Generative Adversarial Network (LAPGAN) [41] that is formed by combining CGAN models progressively within a Laplacian pyramid representation. LAPGAN includes a set of generative convolutional models, say $G_1, \ldots, G_K$. The synthesis procedure consists of two parts a sampling phase and a training phase. The sampling phase starts with generator G_1 that takes a noise sample z_1 and generates sample x_{g_1}. The generated sample is upscaled before passing to the generator of next level as a conditioning variable. G_2 takes both upscaled version of x_{g_1} and a noise sample z_2 to synthesize a difference sample called h_2 which is added to the upscaled version of x_{g_1}. This process of upsampling and addition repeats across successive levels to yield a final full resolution sample. Fig. 2 illustrates the general block diagram of the LAPGAN model.

SGAN is a second example formed by top-down stacked GAN models [42] to solve the low performance of GAN models in discriminative tasks with large variation in data distribution. Huang et al. [42] employ the hierarchical representation in a model trained discriminatively by stitching GAN models in a top-down framework and forcing the top-most level to take class labels and the bottom-most one to generate images. Alternatively, instead of stacking GANs on top of each other, Karras et al. [43] increased the depth of both the generator and the discriminator by adding new layers. All models are developed under conditional GAN [41, 42, 43].

Other models modify the input to the generator slightly. For instance, in SPADE [44] a segmentation mask is fed indirectly to the generator through an adaptive normalization layer instead of utilizing the standard input noise vector $\boldsymbol{z}$. Also, StyleGAN [45] injects $\boldsymbol{z}$, first to an intermediate latent space that helps to avoid entanglement of the input latent space to the probability density of the training data.

In 2015, Radford et al. [46] proposed Deep Convolutional Generative Adversarial Network (DCGAN) in which both the generator and the discriminator were formed by a class of architecturally constrained convolutional networks. In this model, fully convolutional downsampling/upsampling layers replaced the Fully Connected Layers (FC) of vanilla GAN along with other architectural restrictions like using batch-normalization layers and LeakyReLU activation function in all layers of the discriminator.

Another advancement in GAN models includes using the spectral normalization layer to adjust feature response criterion by normalizing the weights in the discriminator network [47]. Residual connections are another novel approach fetched into the GAN models by [48] and [47]. While models like CGANs incorporate the conditional information vector simply by concatenation, others remodeled the usage of a conditional vector by a projection approach leading to significant improvement in the quality of the generated samples [49].

The aforementioned GAN models expanded based on Convolutional Neural Networks (CNN). Further, along this line, a whole new research line of GAN models developed based on recent deep learning models called CapsuleNets (CapsNets) [54]. Consider $\mathbf{v_k}$ as the output vector of the final layer of a CapsNet that represents the presence of a visual entity by classifying to one of the K classes. Sabour et al. [54] provide an updated objective function that benefits from CapsNet margin loss (L_M) and it could be expressed as follows:

$$L_M = \sum_{k=1}^{K} T_k \max(0, m^+ - \|\mathbf{v_k}\|)^2 \\ +\lambda(1 - T_k)\max(0, \|\mathbf{v_k}\| - m^-) \quad (5)$$

where m^+, m^-, and λ are down-weighting factors set to 0.9, 0.1, and 0.5, respectively to stop initial learning from shrinking the lengths of the capsule outputs in the final layer. The length of each capsule in the final layer ($\|v_k\|$) can then be viewed as the probability of the image belonging to a particular class (k). Also, T_k denotes the target label.

CapsuleGAN [55] is a GAN model proposed by Jaiswal et al. based on CapsNet.The authors employ the CapsNet in the discriminator as opposed to conventional CNNs. The final layer of this discriminator consists of a single capsule representing the probability of being a real or fake sample. They used the margin loss introduced in Eq. (5) instead of the binary cross-entropy loss for training. The training criterion of the CapsuleGAN is then formulated as (6):

$$\min_G \max_D \{\mathbb{E}_{\boldsymbol{x}\sim p_r(\boldsymbol{x})}[-L_M(D(\boldsymbol{x}_r), \boldsymbol{T} = \mathbf{1})] \\ +\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[-L_M(D(G(\boldsymbol{z})), \boldsymbol{T} = \mathbf{0})]\} \quad (6)$$

Practically, the generator must be trained to minimize $L_M(D(G(\boldsymbol{z})), \boldsymbol{T} = \mathbf{1})$ rather than minimizing $-L_M(D(G(\boldsymbol{z})), \boldsymbol{T} = \mathbf{0})$. This helps to eliminate the down-weighting factor λ in L_M when training the generator, which does not contain any capsules.

D. VARIANTS BY DISCRIMINATORS

Stabilizing the training and avoiding mode collapse problem could be achieved by employing different loss functions for D. An entropy-based loss is proposed by Springenberg [50] called Categorical GAN (CatGAN) in which the objective of discriminator changed from real-fake classification to entropy-based class predictions. WGAN [56] and an improved version of it called WGAN-GP [48] are two GAN models with a loss function based on Wasserstein distance used in the discriminator. The Earth-Mover (EM) distance or Wasserstein-1 is expressed as follows:

$$W(P_r, P_g) = \inf_{\gamma\epsilon\prod(P_r,P_g)} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma}[\|\boldsymbol{x} - \boldsymbol{y}\|], \quad (7)$$

where $\prod(P_r, P_g)$ is the set of all joint distributions $\gamma(\boldsymbol{x}, \boldsymbol{y})$ whose marginals are respectively, P_r and P_g. Here, $\gamma(\boldsymbol{x}, \boldsymbol{y})$ describes how much "mass" needs to be transported from $\boldsymbol{x}$ to $\boldsymbol{y}$ in order to transform the distribution P_r into P_g. The EM distance is then the "cost" of the optimal transport plan.

Other alternative models that benefit from a different loss metric are GAN based on Category Information (CIGAN) [57], hinge loss[47], least-square GAN [58], and f-divergence

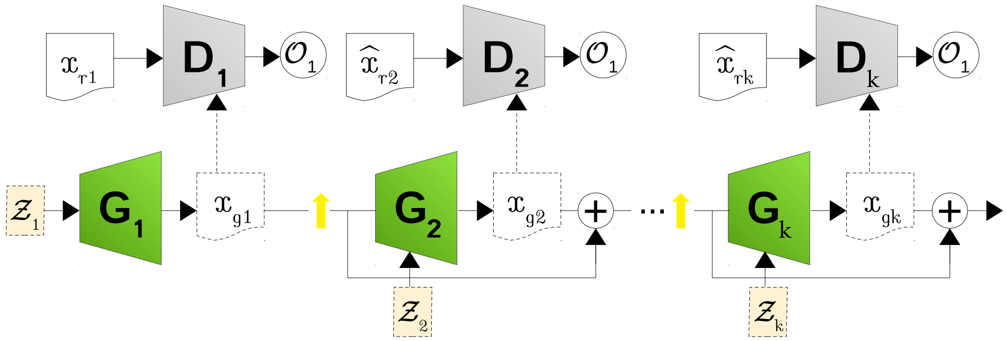
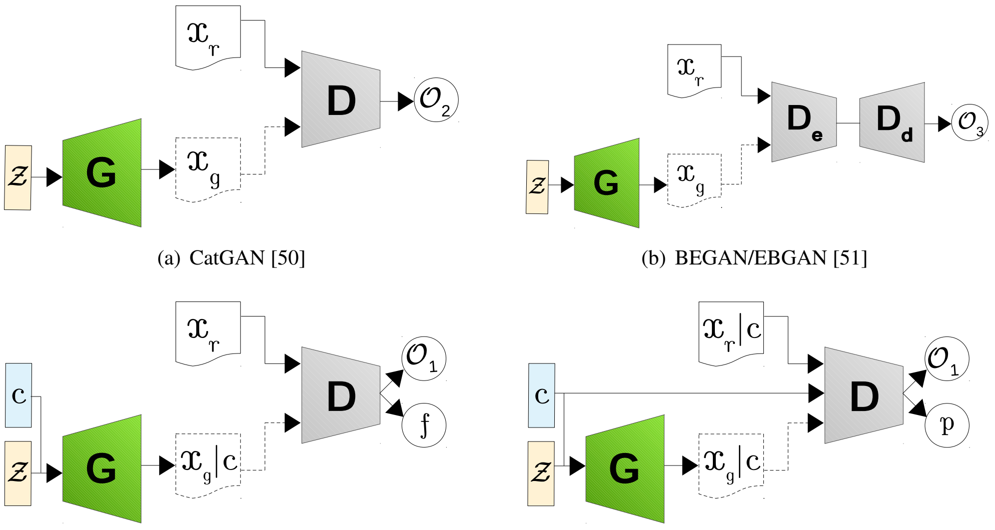


FIGURE 2: Block diagram of LAPGAN model [41]; $\boldsymbol{G_k}$: k-th **G**enerator, $\boldsymbol{D_k}$: k-th **D**iscriminator, $\boldsymbol{\mathcal{X}_{r1}}$: **r**eal sample, $\boldsymbol{\mathcal{X}_{rk}}$: k-th **r**eal residual, $\boldsymbol{\mathcal{X}_{gk}}$: **g**enerated sample, $\boldsymbol{\mathcal{O}_1}$: **O**utput of binary classification to real/fake.



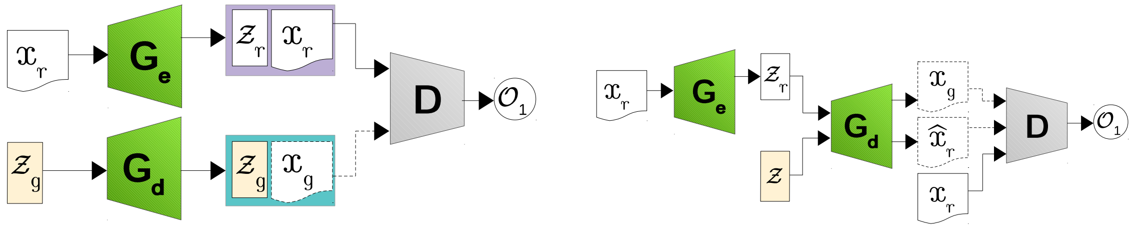
(a) CatGAN [50]

(b) BEGAN/EBGAN [51]

(c) InfoGAN [52]

(d) ACGAN [53]

FIGURE 3: General Structure of GAN models varying by discriminator; $\mathcal{Z}$: input noise, **G**: **G**enerator, **D**: **D**iscriminator, $\boldsymbol{x_r}$: **r**eal sample, $\boldsymbol{x_g}$: **g**enerated sample, $\mathcal{O}_1$: **O**utput of binary classification to real/fake, $\mathcal{O}_2$: **O**utput of category classification, $\mathcal{O}_3$: **O**utput of reconstruction loss for binary classification to real/fake, c: condition vector, $\boldsymbol{D_e}$: incorporated **D**iscriminator-**e**ncoder, $\boldsymbol{D_d}$: incorporated **D**iscriminator-**d**ecoder, $\boldsymbol{f}$: a semantic feature vector extracted from $\boldsymbol{x_g}$, $\boldsymbol{p}$: confidence level.



(a) BiGAN/ALI [61]

(b) VAEGAN [62]

FIGURE 4: General Structure of GAN models varying by generator; $\mathcal{Z}$: input noise, **G**:**G**enerator, **D**:**D**iscriminator, $\boldsymbol{x_r}$: **r**eal sample, $\boldsymbol{x_g}$: **g**enerated sample, $\mathcal{O}_1$: **O**utput of binary classification to real/fake, $\boldsymbol{G_e}$: incorporated **G**enerator-**e**ncoder, $\boldsymbol{G_d}$: incorporated **G**enerator-**d**ecoder, $\boldsymbol{\mathcal{Z}_r}$: latent vector that encodes the input for $\boldsymbol{G_e}$, $\boldsymbol{\mathcal{Z}_g}$: latent vector that encodes the input for $\boldsymbol{G_d}$.

GAN [59]. Research developments include replacing the encoder structure of the discriminator with an autoencoder structure. In fact, a new loss objective is defined for the discriminator which corresponds to the autoencoder loss distribution instead of data distribution. Examples of such GAN frameworks are Energy-based GAN (EBGAN) [60] and Boundary Equilibrium GAN (BEGAN) [51]. Fig. 3 illustrates the block diagram of GAN models developed by modification in the discriminator.

Another interesting GAN model proposed by Chen et al. [52] is the Information Maximizing Generative Adversarial Net (InfoGAN), which simply modifies the discriminator to output both the fake/real classification result and the semantic features of x_g illustrated as f in Fig. 3(c). The discriminator performs real/fake prediction by maximizing the mutual information between the x_g and conditional vector c. Other models like CIGAN [57] and ACGAN [53] focused on improving the quality of the generated samples by employing the class labels during synthesis and then impelling D to provide entropy loss information as well as class probabilities. Fig. 3(d) shows the structure of ACGAN.

E. VARIANTS BY GENERATORS

The objective of generators is to transform noise input vector $\boldsymbol{z}$ to a sample $x_g = G(\boldsymbol{z})$. In the standard vanilla GAN, this objective is achieved by successively improving the state of the generated sample. The procedure stops when the desired quality is captured. Variational AutoEncoder GAN network (VAEGAN) [65] is arguably the most popular GAN model proposed by varying on the generator architecture. The VAEGAN computes the reconstruction loss in a pixel-wise approach. The decoder network of VAE outputs patterns resembling the true samples (see Fig. 4(b)).

One challenge in designing GAN models is controlling the attributes of the generated data known as a mode of data. Using supplemental information leads to sample generation with control over the modification of the selected properties. The generator output then becomes $x_g = G(\boldsymbol{z}, c)$. GANs lack the capability of interpreting the underlying latent space that encodes the input sample. ALI [66] and BiGAN [61] are proposed to resolve this problem by embedding an encoder network in the generator as shown in Fig. 4(a). Here, the discriminator performs real/fake prediction by distinguishing between the tuples (z_g, x_g) and (z_r, x_r). This can categorize the model as a discriminator variant as well.

Other researchers developed the generators to solve specific tasks. Isola et al. [64] designed pix2pix as an image-to-image translation network to study relations between two visual domains and Milletari et al. [68] proposed VNet with Dice loss for image segmentation. The disadvantage of such networks was the aligned training with paired samples. In 2017, Zhu et al. [63] and Kim et al. [69] found a solution to perform unpaired image-to-image translation using cycle consistency loss and cross-domain relations, respectively. Here, the idea was to join two generators together to perform translation between sets of unpaired samples. Fig. 5(a) and Fig. 5(b) show block diagrams of the CycleGAN and pix2pix, respectively.

CycleGAN[63] and UNIT [67] are successful examples derived from VAEGAN model. Fig. 6 illustrates the layout for UNIT framework. It is important to highlight that considering the generators, the conditional input may vary from class labels [38] and text descriptions [70], [71] to object location and encoded audio features or cross-modal correlations.

III. APPLICATIONS IN HUMAN EMOTION SYNTHESIS

In this section, we discuss applications of GAN models in human emotion synthesis. We categorize related works into unimodal and cross-modal researches based on audio and video modalities to help the reader discover applications of interest without difficulty. Also, we explain each method in terms of the proposed algorithm and its advantages and disadvantages. Generally, applications of GAN for human emotion synthesis focus on two issues. The first one is data augmentation that helps to obviate the need for the tedious job of collecting and labeling large scale databases and the second one is improving the performance on emotion recognition.

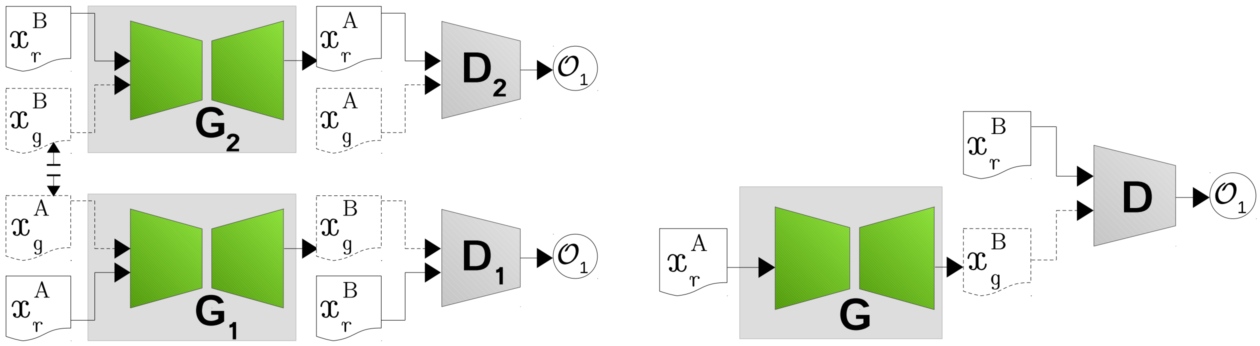
A. FACIAL EXPRESSION SYNTHESIS

Facial expression synthesis using conventional methods confronts several important problems. First, most methods require paired training data, and second, the generated faces are of low resolution. Moreover, the diversity of the generated faces is limited. The works reviewed in this section are taken from the computer-vision-related researches that focus on facial expression synthesis.

1) Image Synthesis

One of the foremost works on facial expression synthesis was the study by Susskind et al. [72] that could embed constraints like "raised eyebrows" on generated samples. The authors build their framework upon a Deep Belief Network (DBN) that starts with two hidden layers of 500 units. The output of the second hidden layer is concatenated with identity and a vector of the Facial Action Coding System (FACS) [73] to learn a joint model of them through a Restricted Boltzmann Machine (RBM) with 1000 logistic hidden units. The trained DBN model is then used to generate faces with different identities and facial Action Units (AU).

Later, with the advent of GAN models, DyadGAN [88] is designed specifically for face generation and it can generate facial images of an interviewer conditioned on the facial expressions of their dyadic conversation partner. ExprGAN [22] is another model designed to solve the problems mentioned above. ExprGAN has the ability to control both the target class and the intensity of the generated expression from weak to strong without a need for training data with intensity values. This is achieved by using an expression controller



(a) CycleGAN (unaligned training pairs) [63]

(b) pix2pix (aligned training pairs) [64]

FIGURE 5: General Structure of GAN models varying by generator; $\mathcal{Z}$: input noise, A: domain A, B: domain B, **G**: **G**enerator, **D**: **D**iscriminator, $\boldsymbol{x}_r^A$ and $\boldsymbol{x}_r^B$: **r**eal sample taken from domain A and B, respectively, $\boldsymbol{x}_g^A$ and $\boldsymbol{x}_g^B$: **g**enerated sample from domain A or B, respectively, $\mathcal{O}_1$: **O**utput of binary classification to real/fake.

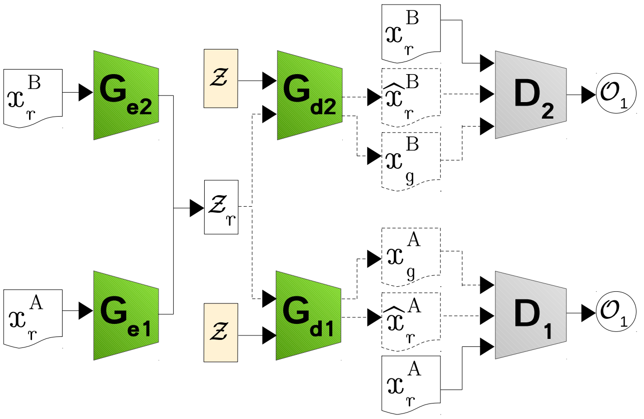


FIGURE 6: General Structure of UNIT GAN model [67]; $\mathcal{Z}$: input noise, A: domain A, B: domain B, $\boldsymbol{x}_r^A$ and $\boldsymbol{x}_r^B$: **r**eal sample taken from domain A and B, respectively, $\boldsymbol{x}_g^A$ and $\boldsymbol{x}_g^B$: **g**enerated sample from domain A or B, respectively, $\hat{\boldsymbol{x}}_r^A$ and $\hat{\boldsymbol{x}}_r^B$: **g**enerated fake sample from domain A or B, respectively, $\mathcal{O}_1$: **O**utput of binary classification to real/fake, $\boldsymbol{G_e}$: incorporated **G**enerator-**e**ncoder, $\boldsymbol{G_d}$: incorporated **G**enerator-**d**ecoder, $\boldsymbol{\mathcal{Z}_r}$: latent vector that encodes the input for $\boldsymbol{G_e}$.

module that encodes complex information like expression intensity to a real-valued vector and by introducing an identity preserving loss function.

Other proposed methods before ExprGAN had the ability to synthesize facial expressions either by manipulating facial components in the input image [111, 112, 113] or by using the target expression as a piece of auxiliary information [72, 114, 115]. In 2017, Shu et al. [117] and Zhou and Shi [116] proposed two GAN-based models. The models proposed by Shu et al. learns a disentangled representation of inherent facial attributes by manipulating facial appearance, while the model described by Zhou and Shi is able to synthesize facial appearance of unseen subjects using AUs and a conditional adversarial autoencoder.

In addition to the GAN models, we summarize the databases, measurements and metrics used by various researchers. In this context, databases and loss functions are listed in Tables 1 and 3. Furthermore, various performance measurements and purposes are given in Tables 4 and 5. Then, in Table 2 we compare the reviewed publication based on various metrics, databases, loss functions and purposes used by researchers. Following those models and through the many variations of facial expression synthesis proposed by researchers, the GAN-based model proposed by Song et al. [98] was one of the interesting and premier ones, called G2GAN. G2GAN generates photo-realistic and identity-preserving images. Furthermore, it provides fine-grained control over the target expression and facial attributes of the generated images like widening the smile of the subject or narrowing the eyes. The idea here is to feed the face geometry into the generator as a condition vector which guides the expression synthesis procedure. The model benefits from a pair of GANs that while one removes the expression, the other synthesizes it. This leverages on the ability of unpaired training.

	Database	#Sbj.	#Smp.	#Cls.
$\mathrm{D}v_{1}$	Oulu-CASIA [74]	80	2,880	$6^{\star}$
$\mathrm{D}v_{2}$	MULTI-PIE [75]	337	755,370	$6^{\dagger}$
$\mathrm{D}v_{3}$	CK+ [76]	123	593	$6^{\diamond}$
$\mathrm{D}v_{4}$	BU-3DFE [77]	100	2,500	$6^{\bullet}$
$\mathrm{D}v_{5}$	RaFD [78]	67	1,068	$6^{\diamond}$
$\mathrm{D}v_{6}$	CelebA [79]	10,177	202,599	45^{+}
$\mathrm{D}v_{7}$	EmotioNet [80]	N/A	1,000,000	$23^{\ddagger}$
$\mathrm{D}v_{8}$	AffectNet [81]	N/A	450,000	$6^{\bullet}$
$\mathrm{D}v_{9}$	FER2013 [82]	N/A	35,887	$6^{\bullet}$
$\mathrm{D}v_{10}$	SFEW [83]	N/A	1,766	$6^{\bullet}$
$\mathrm{D}v_{11}$	JAFFE [84]	10	213	$6^{\bullet}$
$\mathrm{D}v_{12}$	LFW [85]	5,749	13,233	-
$\mathrm{D}v_{13}$	F^{2}ED [86]	119	219,719	$54^{\triangleleft}$
$\mathrm{D}v_{14}$	MUG [87]	52	204,242	$6^{\triangleright}$
$\mathrm{D}v_{15}$	Dyadic [88]	31	-	$8^{\times}$
$\mathrm{D}v_{16}$	IRIS Dataset [89]	29	4,228	$3^{\oplus}$
$\mathrm{D}v_{17}$	MMI [90]	31	236	$6^{\bullet}$
$\mathrm{D}v_{18}$	Driver [91]	26	N/A	$6^{\bullet}$
$\mathrm{D}v_{19}$	KDEF [92]	70	N/A	$6^{\bullet}$
$\mathrm{D}v_{20}$	UVDB [93]	5,793	77,302	-
$\mathrm{D}v_{21}$	3dMD [94]	12,000	N/A	-
$\mathrm{D}v_{22}$	4DFAB [95]	180	1,800,000	$6^{\bullet}$

D stands for Database, A: Audio, V:Visual, A-V:Audio-Visual
$6^{\star}$: 6 basic expressions including angry, disgust, fear, happiness, sadness, and surprise
$6^{\dagger}$: 6 classes including smile, surprised, squint, disgust, scream, and neutral
$6^{\diamond}$: $6^{\star}$ + neutral and contempt
$6^{\bullet}$: $6^{\star}$ + neutral
$23^{\ddagger}$: 6 basic expressions + 17 compound emotions
45^{+}: 5 landmark locations, 40 binary attributes
$54^{\triangleleft}$: 54 emotion types (categories are not mentioned clearly in the source paper)
$6^{\star}$ + neutral, also landmark point annotation is provided)
$6^{\times}$: Joy, Anger, Surprise, Fear, Contempt, Disgust, Sadness and Neutral
$3^{\oplus}$: surprised, laughing, angry

TABLE 1: List of databases used for facial emotion synthesis in the reviewed publications.

StarGAN [23] is the first approach with a scalable solution for multi-domain image-to-image translation using a unified GAN model (i.e. only a single generator and discriminator). In this model, a domain is defined as a set of images sharing the same attribute. The attributes are the facial features like hair color, gender, and age which can be modified based on the desired value. For example, one can set hair color to be blond or brown and set the gender to be male or female.

Likewise, Attribute editing GAN (AttGAN) [106] provides a GAN framework that can edit any attribute among a set of attributes for face images by employing adversarial loss, reconstruction loss, and attribute classification constraints. Also, DIAT [118], CycleGAN [63] and IcGAN [119] could be compared as baseline models.

In 2018, the G2GAN [98] was extended by Qiao et al. [62]. The authors derived a model based on VAEGANs to synthesize facial expressions given a single image and several landmarks through some transferring stages. Different from ExprGAN their model does not require the target class label of the generated image. Also, unlike G2GAN, it does not

Author	Based on	Model	Loss	Data	M	RS	RM
Original GANs							
Li et al.	GAN	-	Lv_1, Lv_3	Dv_6	Mv_5 Mv_6	0.84 20.20	Pv_2, Pv_4
Wang et al.	GAN	CompGAN CycleGAN pix2pix VAEGAN AttGAN	Lv_1, Lv_3, Lv_6, Lv_7, Lv_{10}, Lv_{14}	Dv_{14}	Mv_2	74.92 65.38 71.31 70.88 71.92	Pv_2, Pv_4, Pv_{13}
Deng et al.	GAN	UVGAN	Lv_1, Lv_3, Lv_{20}	Dv_{20}	Mv_5 Mv_6	0.89 25.06	Pv_2, Pv_4, Pv_{13}
Huang and Khan	CGAN	DyadGAN	Lv_1	Dv_5	Mv_7	-	Pv_3, Pv_5, Pv_{15}, Pv_{19}
Ding et al.	CGAN	ExprGAN	Lv_1, Lv_2, Lv_3, Lv_4, Lv_5	Dv_1	Mv_2	84.72	Pv_1, Pv_2, Pv_3, Pv_5
Zhang et al.	CGAN	FaceID	Lv_1, Lv_3, Lv_{18}	Dv_{12}	Mv_4	97.01	Pv_2, Pv_4, Pv_{13}
Song et al.	CGAN	G2GAN	Lv_1, Lv_2, Lv_3, Lv_6, Lv_{12}	Dv_1	Mv_2 Mv_4 Mv_5 Mv_6	58.94 100 0.94 29.50	Pv_1, Pv_2, Pv_5, Pv_{10}
				Dv_2	Mv_5 Mv_6	0.85 24.81	
				Dv_3	Mv_5 Mv_6	0.82 27.30	
Bozorgtabar et al.	CGAN	ExprADA* CycleGAN	Lv_1, Lv_3, Lv_7, Lv_{15}	Dv_4	Mv_2	73.20 71.60	Pv_1, Pv_5, Pv_7, Pv_{13}, Pv_{14}
		ExprADA* CycleGAN		Dv_{17}		70.7 63.50	
		ExprADA* CycleGAN		Dv_{18}		86.90 71.20	
		ExprADA* CycleGAN		Dv_{19}		86.90 82.40	
GAN Variants by Generator							
Choi et al.	CycleGAN	StarGAN	Lv_1, Lv_6, Lv_7	Dv_5	Mv_{10} Mv_3	52.20 2.12	Pv_4, Pv_7, Pv_{14}
Lu et al.	CycleGAN	AttGGAN	Lv_1, Lv_3, Lv_6	Dv_6	Mv_5	0.92	Pv_8, Pv_{10}, Pv_{14}
Peng and Yin	CycleGAN	ApprGAN	Lv_1, Lv_3, Lv_{10}	Dv_3 Dv_{14}	Mv_{15}	0.95 0.97	Pv_8

Continue on the next page

TABLE 2: Comparison of facial expression image synthesis models, description of loss functions (L), metrics (M), databases (D) and purposes (P) used in the reviewed publications are given in Tables 1, 3, 4, and 5.

Author	Based on	Model	Loss	Data	M	RS	RM
Lee et al.	CycleGAN	CollaGAN	Lv_1, Lv_6, Lv_7, Lv_{17}	Dv_2	Mv_7	-	Pv_7, Pv_{10}, Pv_{16}, Pv_{17}, Pv_{18}
Caramihale et al.	CycleGAN	-	Lv_1, Lv_6	Dv_3 Dv_9 Dv_{10} Dv_{11} Dv_{12}	Mv_2	98.30 75.20 60.80 94.80 75.70	Pv_1, Pv_{13}
Zhu et al.	CycleGAN	-	Lv_1, Lv_6	Dv_9 † ‡	Mv_2	94.71 39.07 95.80	Pv_1
Lai and Lai	VAEGAN	-	Lv_1, Lv_2, Lv_7, Lv_8, Lv_{13}	Dv_2 Dv_4	Mv_2	87.08 73.13	Pv_1, Pv_{12}, Pv_{13}
Lindt et al.	VAEGAN	-	Lv_1, Lv_3, Lv_4, Lv_{10}	Dv_8	Mv_3	6.07	Pv_3, Pv_4, Pv_9
He et al.	VAEGAN IcGAN	AttGAN* CycleGAN VAEGAN StarGAN	Lv_1, Lv_7, Lv_{10}	Dv_6	Mv_{11}	88.20 67.80 52.00 86.20	Pv_3, Pv_4, Pv_7, Pv_{14}
		AttGAN* CycleGAN StarGAN		Dv_{12}		88.00 63.40 73.40	
Shen et al.	Pix2Pix	TVGAN	Lv_1, Lv_3	Dv_{16}	Mv_4	50.90	Pv_1, Pv_2, Pv_{20}
Vielzeuf et al.	StarGAN	-	Lv_1, Lv_7, Lv_{10}	Dv_8	Mv_3	3.4	Pv_2, Pv_9, Pv_{11}
GAN Variants by Discriminator							
Wang et al.	ACGAN	UNet GAN	Lv_1, Lv_7, Lv_{16}	Dv_1 Dv_5	Mv_2	43.33 82.59	Pv_2, Pv_4
Cheng et al.	BEGAN ChebNet CoMA	MeshGAN* CoMA	Lv_1	Dv_{21}	Mv_{16}	1.43 1.60	Pv_2, Pv_4
		MeshGAN* CoMA		Dv_{22}	Mv_3	0.85 1.89	
GAN Variants by Architecture							
Shen et al.	StackGAN	FaceFeat	Lv_1, Lv_3, Lv_8, Lv_{20}	Dv_{12}	Mv_4	97.62	Pv_2, Pv_4

- M: Metric, RS: Results, RM:Remarks
- *: shows the proposed method by authors, other mentioned methods are implemented by the authors for the sake of comparison
- the result reported for expression classification accuracy (Mv_2) belongs to the synthesized image datasets
-†: $Dv_9 + Dv_{10}$ is used as the database
-‡: $Dv_9 + Dv_{11}$ is used as the database
- All papers provide visual representation of the synthesized images (Mv_7)

TABLE 2: Comparison of facial expression image synthesis models, description of loss functions (L), metrics (M), databases (D) and purposes (P) used in the reviewed publications are given in Tables 1, 3, 4, and 5.

require the neutral expression of a specific subject as an intermediate level in the facial expression transfer procedure. While G2GAN and its extension focus on geomterical features to guide the expression synthesis procedure, Pumarola et al. [120] use facial AU as a one-hot vector to perform an unsupervised expression synthesis while smooth transition and unpaired samples are guaranteed.

Another VAEGAN-based model is the work of Lai and

Lai [104] where a novel optimization loss called symmetric loss is introduced. Symmetric loss helps to preserve the symmetrical property of the face while translating from various head poses to frontal-view of the face. Similar to Lai and Lai is the FaceID-GAN [107] where a classifier of face identity is added to the two-players of vanilla GANs and symmetry information. This classifier is the third player that competes with the generator and it distinguishes the identities of the real and synthesized faces.

Lai and Lai [104] also used GAN to perform emotion-preserving representations. In the proposed approach, the generator can transform the non-frontal facial images into frontal ones while the identity and the emotion expression are preserved. Moreover, a recent publication [108] relies on a two-step GAN framework. The first component maps images to a 3D vector space. This vector is issued from a neural network and it represents the corresponding emotion of the image. Then, a second component that is a standard image-to-image translator uses the 3D points obtained in the first step to generate different expressions. The proposed model provides fine-grained control over the synthesized discrete expressions through the continuous vector space representing the arousal, valence, and dominance space.

It should be noted that a series of GAN models focus on 3D object/face generation. Examples of these models are Convolutional Mesh Autoencoder (CoMA) [121], MeshGAN[94], UVGAN [93], and MeshVAE [122]. Despite the successful performance of GANs in image synthesis, they still fall short when dealing with 3D objects and particularly human face synthesis.

2) Video Synthesis

In addition to GAN-based models that synthesize single images, there are models with the ability to generate an image sequence or a video/animation. Video GAN (VGAN) [130] and Temporal GAN (TGAN) [131] were the first two models in this research line. Although these models could learn a semantic representation of unlabeled videos, they produced a fixed-length video clip. As a result, MoCoGAN is proposed by Tulyakov et al. [24] to solve the problem. MoCoGAN is composed of 4 sub-networks. These sub-networks are a recurrent neural network, an image generator, an image discriminator, and a video discriminator. The image generator generates a video clip by sequentially mapping a sequence of vectors to a sequence of images.

While MoCoGAN uses content and motion, Depth Conditional Video generation (DCVGAN) proposed by Nakahira and Kawamoto [125] utilizes both the optical information and the 3D geometric information to generate accurate videos using scene dynamics. DCVGAN solved the unnatural appearance of moving objects and assimilation of objects into the background in MoCoGAN. Other methods like Warp-

	Name	Remarks
$\mathrm{L}v_{1}$	$\mathcal{L}_{\text{adv}}$	adversarial loss presented by the discriminator (see section II-D)
$\mathrm{L}v_{2}$	$\mathcal{L}_{\text{pixel}}$	pixel-wise image reconstruction loss
$\mathrm{L}v_{3}$	$\mathcal{L}_{\text{id}}$	identity preserving loss
$\mathrm{L}v_{4}$	$\mathcal{L}_{\text{regl}}$	loss of a regularizer
$\mathrm{L}v_{5}$	$\mathcal{L}_{\text{tv}}$	total variation regularizer loss
$\mathrm{L}v_{6}$	$\mathcal{L}_{\text{cyc}}$	cycle-consistency loss
$\mathrm{L}v_{7}$	$\mathcal{L}_{\text{cls}}$	classification loss (expression)
$\mathrm{L}v_{8}$	$\mathcal{L}_{\text{feat}}$	feature matching loss
$\mathrm{L}v_{9}$	$\mathcal{L}_{\text{contr}}$	contrastive loss
$\mathrm{L}v_{10}$	$\mathcal{L}_{\text{rec}}$	image reconstruction loss
$\mathrm{L}v_{11}$	$\mathcal{L}_{\text{att}}$	attention loss
$\mathrm{L}v_{12}$	$\mathcal{L}_{\text{cnd}}$	conditional loss
$\mathrm{L}v_{13}$	$\mathcal{L}_{\text{symm}}$	Symmetry loss to preserve symmetrical property of the face
$\mathrm{L}v_{14}$	$\mathcal{L}_{\text{pose}}$	cross-entropy loss used to ensure correct pose of the face
$\mathrm{L}v_{15}$	$\mathcal{L}_{\text{bi}}$	bidirectional loss to avoid mode collapse
$\mathrm{L}v_{16}$	$\mathcal{L}_{\text{triplet}}$	triplet loss to minimize the similarity between $\boldsymbol{x}_r$ and $\boldsymbol{x}_g$
$\mathrm{L}v_{17}$	$\mathcal{L}_{\text{SSIM}}$	Structural Similarity Index Loss that measures the image quality
$\mathrm{L}v_{18}$	$\mathcal{L}_{\text{shape}}$	learn the shape feature by minimizing the weighted distance
$\mathrm{L}v_{19}$	$\mathcal{L}_{\text{recurr}}$	loss for a recurrent temporal predictor to predict future samples
$\mathrm{L}v_{20}$	$\mathcal{L}_{\text{3DMM}}$	3D Morphable Model loss to ensure correct pose and expression
$\mathrm{L}v_{21}$	$\mathcal{L}_{\text{motion}}$	motion loss consisting of a VAE loss, a video reconstruction loss, and KLD between the prior and posterior motion latent distribution
$\mathrm{L}v_{22}$	$\mathcal{L}_{\text{cnt}}$	contents loss consisting of a reconstruction loss for the current frame and a KLD between the prior and posterior content distribution
$\mathrm{L}v_{23}$	$\mathcal{L}_{\text{reg}}$	a regression loss between input target and its regression estimates on generated image

TABLE 3: List of loss functions used for facial emotion synthesis in the reviewed publications.

	Measurement	Remarks
Mv_1		Ground truth, costly, not scaleable
Mv_2	expression classification (accuracy)	down stream task
Mv_3	expression classification (error)	down stream task
Mv_4	identity classification (accuracy)	down stream task
Mv_5	Structural Similarity Index Measure	measures image quality degradation
Mv_6	Peak Signal to Noise Ratio (PSNR)	measures quality of representation
Mv_7	visual representation	down stream task
Mv_8	real/fake classification (accuracy)	down stream task
Mv_9	real/fake classification (error)	down stream task
Mv_{10}	attribute classification (accuracy)	down stream task
Mv_{11}	attribute classification (error)	down stream task
Mv_{12}	Average Content Distance (ACD)	content consistency of a generated video
Mv_{13}	Motion Control Score (MCS)	capability in motion generation control
Mv_{14}	Inception Score (IS)	measures quality and diversity of $p_g(\boldsymbol{x})$
Mv_{15}	texture similarity score	measuring texture similarity
Mv_{16}	identity classification (error)	down stream task

TABLE 4: List of evaluative metrics used for facial emotion synthesis in the reviewed publications.

guided GAN [123] generate real-time facial animations using a single photo. The method instantly fuses facial details like wrinkles and creases to achieve a high fidelity facial expression.

Recently, Yang et al. (2018) proposed a pose-guided method to synthesize human videos. This successful method relies on two concepts: first, a Pose Sequence Generative Adversarial Network (PSGAN) is proposed to learn various motion patterns by conditioning on action labels. Second, a Semantic Consistent Generative Adversarial Network (SCGAN) is employed to generate image sequences (video) given the pose sequence generated by the PSGAN. The effect of noisy or abnormal poses between the generated and ground-truth poses is reduced by the semantic consistency. We show this method as PS/SCGAN in Table 6. It is worth to mention that two of the recent and successful methods in video generation are MetaPix [132] and MoCycleGAN [133] that used motion and temporal information for realistic video synthesis. However, these methods are not tested for facial expression generation. Tables 1 and 3 include databases and loss functions used for video synthesis. Also, lists of metrics and applications can be found in Tables 4 and 5. Later, in Table 6 the video synthesis models are compared using those metrics and applications.

More recently, Yu et al. [129] proposed a video-based synthesis version of the StarGAN called StarGAN-EgVA. The StarGAN-EgVA model generates continuous facial emotions based on arousal and valence emotional representation. The authors subtly utilized the discrete emotions to guide the training on arousal and valence intensities. Hence, the

	Purpose or characteristic		Purpose or characteristic
Pv_1	is tested for data augmentation	Pv_{12}	preserves the emotion of the subject
Pv_2	preserves the identity of the subject	Pv_{13}	employs arbitrary head poses and applies face frontalization
Pv_3	controls the expression intensity	Pv_{14}	modifies facial attributes based on a desired value
Pv_4	generates images with high quality (photo-realistic)	Pv_{15}	generates image sequences (video)
Pv_5	employs geometrical features as conditional vector	Pv_{16}	employs multiple inputs
Pv_6	employs facial action units as conditional vector	Pv_{17}	modifies image attributes like illumination and background
Pv_7	employs a unified framework for multi-domain tasks	Pv_{18}	is tested for data imputation
Pv_8	employs combination of appearance and geometric features	Pv_{19}	generates video animation using a single image
Pv_9	provides a smooth transition of facial expression	Pv_{20}	generates visible faces from thermal face image
Pv_{10}	supports training with unpaired samples	Pv_{21}	employs temporal/motion information for video generation
Pv_{11}	employs arousal-valence and dominance-like inputs	Pv_{22}	supports unsupervised learning

TABLE 5: List of purposes and characteristics of GAN models used for facial emotion synthesis by the reviewed publications.

Author	Based on	Method	Loss	Data	M	RS	RM
Original GANs							
Kim et al.	GAN	DVP/VDub	Lv_1	-	Mv_8	51.25	Pv_4, Pv_5, Pv_{15}
Huang and Khan	CGAN	DyadGAN	Lv_1	Dv_5	Mv_7	-	$Pv_3, Pv_5, Pv_{15}, Pv_{19}$
Geng et al.	CGAN	wg-GAN/ Elor et al.	Lv_1	Dv_{14}	Mv_8	62.00	$Pv_8, Pv_9, Pv_{14}, Pv_{19}$
Nakahira and Kawamoto	CGAN	DCVGAN	Lv_1	Dv_{14}	Mv_{14}	6.68	Pv_{15}
Sun et al.	CGAN	2SVAN*† */MoCoGAN	$Lv_1, Lv_7, Lv_{21}, Lv_{22}$	Dv_{14}	Mv_{14} Mv_8	5.48 88.00	$Pv_8, Pv_9, Pv_{14}, Pv_{19}$
Tulyakov et al.	CGAN	MoCoGAN* VGAN TGAN	Lv_1, Lv_7	Dv_{14}	Mv_{12}	0.201 0.322 0.305	Pv_4, Pv_7, Pv_{14}
		MoCoGAN*			Mv_{13}	0.581	
		*/VGAN */TGAN			Mv_8	84.20 54.70	
Yang et al.	GAN CGAN	PS/SCGAN* VGAN MoCoGAN	Lv_1, Lv_{10}, Lv_{12}	Dv_3	Mv_{14}	1.92 1.68 1.83	Pv_1, Pv_{15}
		*/VGAN */MoCoGAN			Mv_8	93.00 86.00	
GAN Variants by Generator							
Pumarola et al.	CycleGAN	GANimation	$Lv_1, Lv_3, Lv_5, Lv_{11}, Lv_{12}$	Dv_7	Mv_7	-	Pv_6, Pv_9, Pv_{15}
Bansal et al.	CycleGAN	RecycleGAN	Lv_1	-	Mv_8	76.00	$Pv_{10}, Pv_{15}, Pv_{21}, Pv_{22}$
Qiao et al.	VAEGAN	-	Lv_1, Lv_9, Lv_{10}	Dv_2	Mv_5 Mv_6	0.69 26.73	$Pv_4, Pv_8, Pv_9, Pv_{19}$
				Dv_3	Mv_5 Mv_6	0.77 27.67	
Yu et al.	StarGAN	StarGAN-EgVA GANimation	Lv_1, Lv_7, Lv_{23}	Dv_8	Mv_3	0.26 0.27	Pv_7, Pv_{11}

- M: Metric, RS: Results, RM:Remarks
- *: shows the proposed method by authors, other mentioned methods are implemented by the authors for the sake of comparison
- the result reported for expression classification accuracy (Mv_2) belongs to the synthesized image datasets
- All papers provide visual representation of the synthesized images (Mv_7)
- *†:TwoStreamVAN

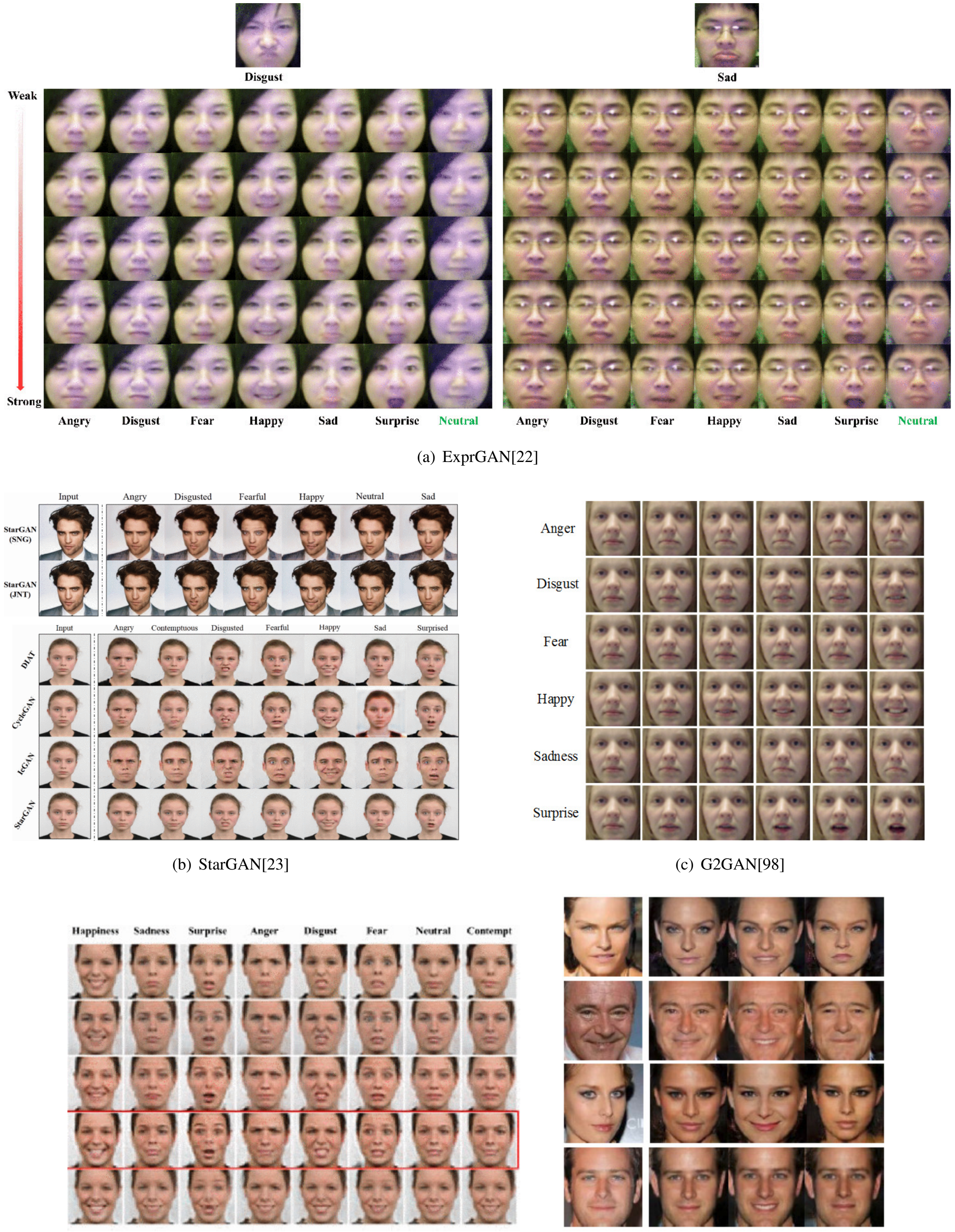
TABLE 6: Comparison of facial expression video generation models, description of loss functions (L), metrics (M), databases (D) and purposes (P) used in the reviewed publications are given in Tables 1, 3, 4, and 5.

model learns both the domain-specific features and embedded changes introduced by different arousal and valence intensities. StarGAN-EgVA modifies the original StarGAN architecture by adding an arousal-valence channel to the generator and also having an output channel of arousal-valence from the discriminator. Furthermore, regression loss is added to the loss functions of StarGAN.

One of the main goals in synthesizing is augmenting the number of available samples. Zhu et al. [103] used GAN models to improve the imbalanced class distribution by data augmentation through GAN models. The discriminator of the model is a CNN and the generator is based on CycleGAN. They report up to 10% increase in the classification accuracy (Mv_2) based on GAN-based data augmentation techniques. In fact, generative-based face data augmentation approaches like GANs, VAEs, PixelCNN, and Glow have became a trending approach in facial expression data augmentation. A fine source of such approaches can be found in [134].

The objective function or the optimization loss problem categorizes into two groups: synthesis loss and classification loss. Although the definitions provided by the authors are not always clear, we tried to list all different losses used by authors and we propose a symbolic name for each to provide harmony in the literature. The losses are used in a general point of view. That is, marking different papers by classification loss (L7) in Table 2, does not mean necessarily that the exact same loss function is used. In other words, it shows that the classification loss is contributed in some way. A comprehensive list of these functions is given in Table 3.

We compare synthesized images of the aforementioned methods qualitatively in Fig. 7 and Fig. 8 (See next page). Additionally, we compare some of the video synthesis models in Fig. 9. Images are taken from the corresponding papers. As the images show, most of the generated samples suffer from blurring problem.



(a) ExprGAN[22]

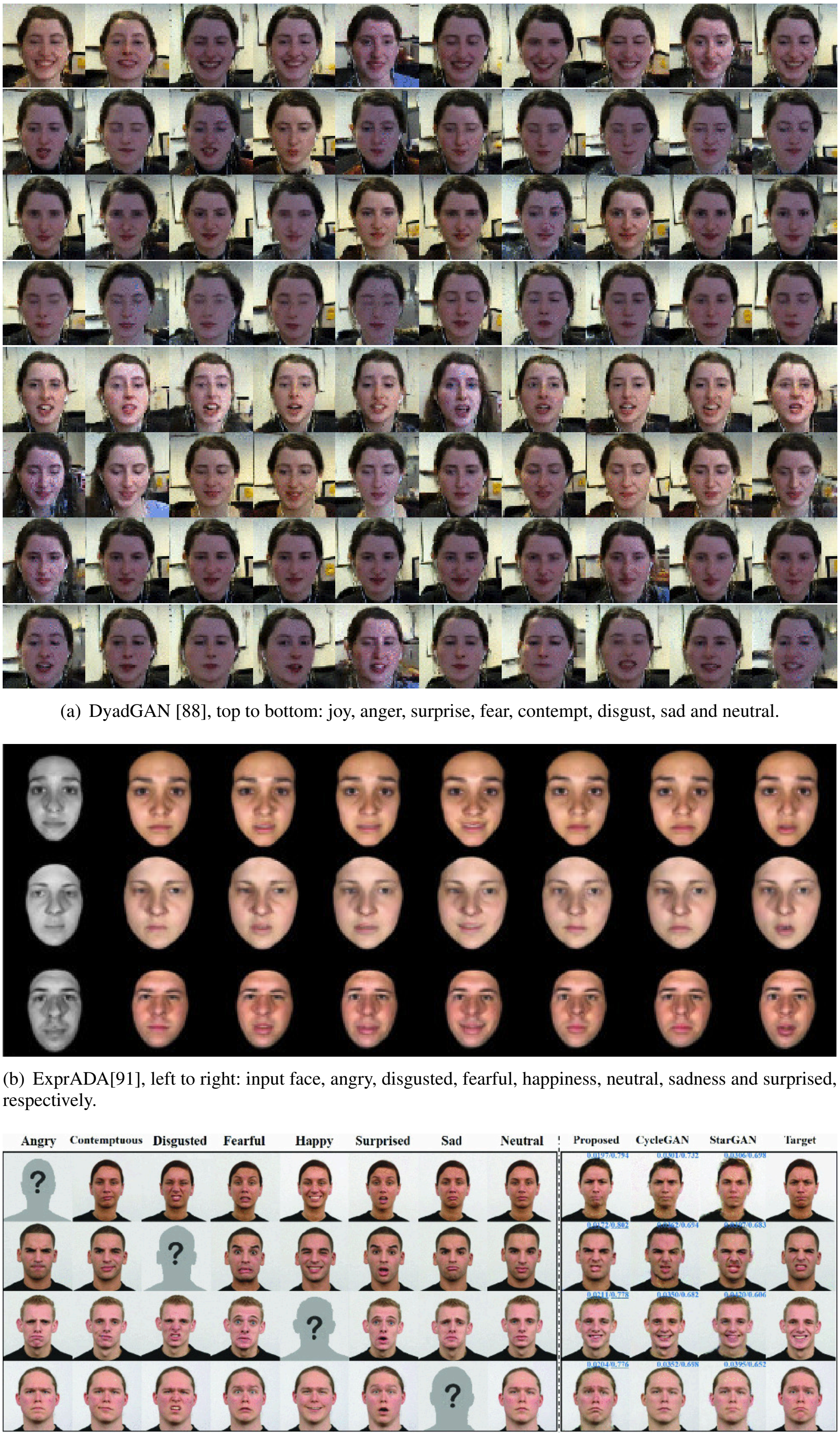
(b) StarGAN[23]

(c) G2GAN[98]

(d) Zhou and Shi [116]

(e) FaceFeat[110]

FIGURE 7: Visual comparison of the GAN models for image generation of facial expression, images are in courtesy of the reviewed papers.

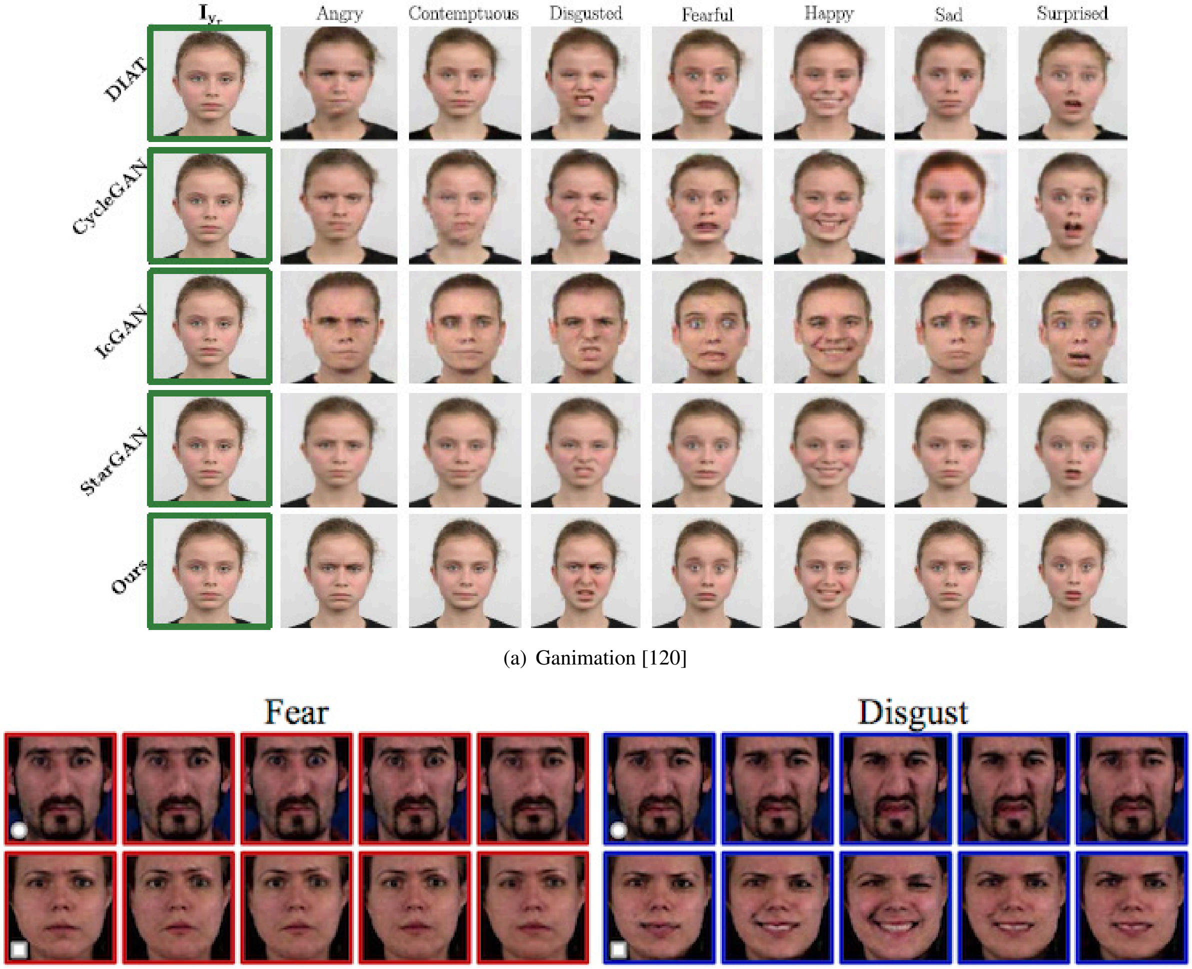


(a) DyadGAN [88], top to bottom: joy, anger, surprise, fear, contempt, disgust, sad and neutral.

(b) ExprADA[91], left to right: input face, angry, disgusted, fearful, happiness, neutral, sadness and surprised, respectively.

(c) CollaGAN[101]

FIGURE 8: Visual comparison of the GAN models for image generation of facial expression, images are in courtesy of the reviewed papers.



(a) Ganimation [120]

(b) MocoGAN [24]

FIGURE 9: Visual comparison of the GAN models for video generation of facial expressions, images are in courtesy of the reviewed papers.

3) Evaluation Metrics for Image

Evaluation metrics of the generative models are different from one research to another due to several reasons [8]. First, the quality of the synthesized sample is a perceptual concept and, as a result, it cannot be accurately expressed. Usually, researchers provide the best-synthesized samples for visual comparison and thus problems like mode drop are not covered qualitatively. Second, employing human annotators to judge the visual quality can cover only a limited number of data samples. Specifically, in topics such as human emotion, experts are required for accurate annotation and having the least possible labeling error. Hence, approaches like Amazon Mechanical Turk are less reliable considering classification based on those labels. Third, general metrics like photometric error, geometric error, and inception score are not reported in all publications [135]. These problems cause the comparison among papers either unfair or impossible.

One widely evaluative metric is the Inception Score (IS) that can be computed as follows (8):

$$\text{IS} = \exp(\mathbb{E}_{\boldsymbol{x}_g}[\text{KL}(p(y|x_g) \,||\, p(y))]), \tag{8}$$

where x_g denotes the generated sample, y is the label predicted by an arbitrary classifier, and KL$(.)$ is the KL divergence to measure the distance between probability distributions as defined in Eq. (1). Based on this score, an ideal model produces samples that have close congruence to real data samples as much as possible. In fact, KL divergence is the de-facto standard for training and evaluating generative models.

Other used evaluative metrics are Structural Similarity Index Measure (SSIM) (9) [136] and Peak Signal to Noise Ratio (PSNR).

$$\text{SSIM}(x, y) = I(x, y)^{\alpha} C(x, y)^{\beta} S(x, y)^{\gamma}, \tag{9}$$

where I, C, and S are luminance, contrast, and structure and they can be formulated as (10):

$$
\begin{aligned}
I(x,y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\\
C(x,y) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\\
S(x,y) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}
\end{aligned}
\tag{10}
$$

Here μ_x, μ_y, σ_x, and σ_y denote mean and standard deviations of pixel intensity in a local image patch where the patch is superimposed so that its center coincides with the center of the image. Typically, a patch is considered as a square neighborhood of $n \times x$ pixels. Also, σ_{xy} is the sample correlation coefficient between corresponding pixels in that patch. $C1$, $C2$, and $C3$ are small constants values added for numerical stability.

PSNR or the peak signal-to-noise ratio assesses the quality between two monochrome images $\boldsymbol{x}_g$ and $\boldsymbol{x}_r$. Let $\boldsymbol{x}_g$ and $\boldsymbol{x}_r$ be the generated image and the real image, respectively. Then, PSNR is can be measured in dB as indicated in (11):

$$
\text{PSNR}(\boldsymbol{x}_g, \boldsymbol{x}_r) = 20\log_{10}\left(\frac{\text{MAX}_{\boldsymbol{x}_r}}{\text{MSE}(\boldsymbol{x}_g, \boldsymbol{x}_r)}\right), \tag{11}
$$

where $\text{MAX}_{\boldsymbol{x}_r}$ is the maximum possible pixel value of the image and MSE stands for Mean Square Error. PSNR is measured in dB, generated images with a better quality result in higher PSNR.

In addition to the metrics that evaluate the generated image, the Generative Adversarial Metric (GAM) proposed by Im et al. (2016) compares two GAN models by engaging them in a rivalry. In this metric, first GAN models M_1 and M_2 are trained. Then, model M_1 competes with model M_2 in a test phase by having M_1 trying to fool discriminator of M_2 and vice versa. In the end, two ratios are calculated using the discriminative scores of these models as (12):

$$
r_{\text{test}} \overset{\text{def}}{=} \frac{\epsilon(D_1(X_{\text{test}}))}{\epsilon(D_2(X_{\text{test}}))}, \text{ and } r_{\text{sample}} \overset{\text{def}}{=} \frac{\epsilon(D_1(G_2(Z)))}{\epsilon(D_2(G_1(Z)))} \tag{12}
$$

where G_1, D_1, G_2, and D_2, are the generators and the discriminators of M_1 and M_2, respectively. In Eq. (12), $\epsilon(.)$ outputs the classification error rate. The test ratio or r_{test} shows which model generalizes better because it discriminates based on X_{test}. The sample ratio or r_{sample} shows which model fools the other more easily because discriminators classify based on the synthesized samples of the opponent. The sample ratio and the test ratio can be used to decide the winning model:

$$
\text{winner} = \begin{cases} M_1 & \text{if } r_{\text{sample}} < 1 \text{ and } r_{\text{test}} \simeq 1\\ M_2 & \text{if } r_{\text{sample}} > 1 \text{ and } r_{\text{test}} \simeq 1\\ \text{Tie} & \text{otherwise} \end{cases} \tag{13}
$$

To measure the texture similarity, Peng and Yin [100] simply calculated correlation coefficients between T_g and T_r that are the texture of the synthesized image and the texture of its corresponding ground truth, respectively. Let ρ be the texture similarity score. Then, the mathematical representation is as (14):

$$
\rho = \frac{\sum_i\sum_j(T_r(i,j) - \mu_r)(T_g(i,j) - \mu_g)}{\sqrt{\sum_i\sum_j(T_r(i,j) - \mu_r)^2 \sum_i\sum_j(T_g(i,j) - \mu_g)^2}}, \tag{14}
$$

where (i,j) specifies pixel coordinates in the texture images, and μ_g and μ_r are the mean value of T_g and T_r, respectively.

Other important metrics include Fréchet Inception Distance (FID), Maximum Mean Discrepancy (MMD), the Wasserstein Critic, Tournament Win Rate and Skill Rating, and Geometry Score. FID works based on embedding the set of synthesized samples into a new feature space using a certain layer of a CNN architecture. Then, mean and covariance are estimated for both the synthesized and the real data distributions based on the assumption that the embedding layer is a continuous multivariate Gaussian distribution. Finally, FID or Wasserstein-2 distance between these Gaussians is then used to quantify the quality of generated samples as calculated in (15):

$$
\text{FID}(r,g) = \parallel \mu_r - \mu_g \parallel_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}). \tag{15}
$$

Here, (μ_g, Σ_g) and (μ_r, Σ_r) represent the mean and covariance of generated and real data distributions, respectively. Lower FID score indicates a smaller distance between the two distributions. MMD focuses on the dissimilarity between the two probability distributions by taking samples from each distribution independently. The kernel MMD is expressed as (16):

$$
\begin{aligned}
M_K(P_r, P_g) = &\mathbb{E}_{\boldsymbol{x}_r, \boldsymbol{x}'_r \sim P_r}[k(\boldsymbol{x}_r, \boldsymbol{x}'_r)]\\
&+\mathbb{E}_{\boldsymbol{x}_g, \boldsymbol{x}'_g \sim P_g}[k(\boldsymbol{x}_g, \boldsymbol{x}'_g)]\\
&-2\mathbb{E}_{\boldsymbol{x}_r \sim P_r, \boldsymbol{x}_g \sim P_g}[k(\boldsymbol{x}_r, \boldsymbol{x}_g)]
\end{aligned}
\tag{16}
$$

where k is some fixed characteristic kernel function like Gaussian kernel: $k(x_r, x_g) = \exp(\parallel x_r - x_g \parallel^2)$ that measures MMD dissimilarity between the generated and real data distributions. Also, $\boldsymbol{x}_r$ and $\boldsymbol{x}'_r$ are randomly drawn samples from real data distribution, i.e. P_r. Similarly, $\boldsymbol{x}_g$ and $\boldsymbol{x}'_g$ are randomly drawn from model distribution, i.e. P_g.

The Wasserstein Critic provides an approximation of the Wasserstein distance between the model distribution and the real data distribution. Let P_r and P_g be the real data and the model distributions, then:

$$
\begin{aligned}
W(P_r, P_g) \propto \max_f \{&\mathbb{E}_{\boldsymbol{x}_r \sim P_r}[f(\boldsymbol{x}_r)]\\
&-\mathbb{E}_{\boldsymbol{x}_g \sim P_g}[f(\boldsymbol{x}_g)]\},
\end{aligned}
\tag{17}
$$

where $f : \mathbb{R}^D \longrightarrow \mathbb{R}$ is a Lipschitz continuous function. In practice, the critic f is a neural network with clipped weights and bounded derivatives [136]. In practice, this is approximated by training to achieve high values for real samples and low values for generated ones:

$$\hat{W}(X_{\text{test}}, X_g) = \frac{1}{N}\sum_{i=1}^{N}\hat{f}(X_{\text{test}}[i]) - \frac{1}{N}\sum_{i=1}^{N}\hat{f}(X_g[i]), \quad (18)$$

where X_{test} is a batch of testing samples, X_g is a batch of generated samples, and $\hat{f}$ is the independent critic. An alternative version of this score is known as Sliced Wasserstein Distance (SWD) that estimates the Wasserstein-1 distance (see Eq. (7)) between real and generated images. SWD computes the statistical similarity between local image patches extracted from Laplacian pyramid representations of the images [43].

In the case of the metrics for video generation, evaluating content consistency based on Average Content Distance (ACD) is defined as calculating the average pairwise L_2 distance of the per-frame average feature vectors. In addition, the motion control score (MCS) is suggested for assessing the motion generation ability of the model. Here, a spatio-temporal CNN is first trained on a training dataset. Then, this model classifies the generated videos to verify whether the generated video contained the required motion (e.g. action/expression).

Other metrics include but are not limited to identification classification, true/false acceptance rate [98], expression classification accuracy/error [22], real/fake classification accuracy/error [22], attribute editing accuracy/error [106], and Fully Convolutional Networks. List of evaluative metrics used in the reviewed publications is given in Table 4. For a comprehensive list on evaluative metrics of GAN models, we invite the reader to study “Pros and Cons of GAN Evaluation Measures" by [136].

Synthesizing models are proposed with different aims and purposes. Texture synthesis, super-resolution images, and image in-painting are some applications. Considering face synthesis, the most important goal is the data augmentation for improved recognition performance. A complete list of such purposes and the model properties are given in Table

Author	Based on	Model	Loss	Data	M	RS	RM
			Original GANs				
Latif et al.	GAN	-	La_1	Da_3 Da_4	Ma_7	47.87 36.49	Pa_4, Pa_5
Gao et al.	GAN	-	La_1, La_2, La_5	Da_3	Ma_2 Ma_{11} Ma_{12}	41.25 40.75 47.00	Pa_3, Pa_6
Pascual et al.	CGAN	SEAGAN	La_1	Da_1	Ma_1 Ma_2 Ma_3 Ma_4 Ma_5 Ma_6	2.16 3.18 3.48 2.94 2.80 7.73	Pa_2
Sheng et al.	CGAN	-	La_1, La_{10}	Da_9	Ma_{18}	8.16	Pa_1
Sahu et al.	CGAN	-	La_1, La_4	Da_3	Ma_{17}	60.29	Pa_1, Pa_2
Chatziagapi et al.	CGAN (BAGAN)	-	La_1, La_2, La_6, La_{11}	Da_3	Ma_2 Ma_{12} Ma_{14} Ma_{19}	2.88 66.05 50.62 3.41	Pa_1, Pa_{14}
Hsu et al.	WGAN CVAE GAN	VAWGAN	La_1	Da_2	Ma_2	3.00	Pa_3
			GAN Variants by Generator				
Mathur et al.	CycleGAN	Mic2Mic	La_1, La_2, La_6	Da_8	Ma_8	89.00	Pa_2
Kameoka et al.	StarGAN CycleGAN	StarGAN- VC	La_1, La_2, La_3, La_6	Da_5	Ma_{13} Ma_{14}	82.00 67.00	Pa_{11}

Continue on the next page

TABLE 7: Comparison of speech Emotion synthesis models, description of loss functions (L), metrics (M), databases (D) and purposes (P) used in the reviewed publications are given in Tables 8, 9, 10, and 11.

Author	Based on	Model	Loss	Data	M	RS	RM
Kaneko and Kameoka	CycleGAN	CycleGAN-VC	La_1, La_2, La_6	Da_2	Ma_2 Ma_{14}	2.4 39.5	Pa_3
Kaneko et al.	CycleGAN	CycleGAN-VC2	La_1, La_2, La_6	Da_5	Ma_2 Ma_{11} Ma_{13} Ma_{15} Ma_{16}	3.1 4.8 69.5 6.26 1.45	Pa_7 Pa_{10}
Tanaka et al.	CycleGAN SEAGAN	Wave-CycleGAN-VC	La_1, La_2	Da_6	Ma_2	4.18	Pa_8
Tanaka et al.	CycleGAN	Wave-CycleGAN-VC2	La_1, La_2, La_6	Da_6	Ma_2	4.29	Pa_8, Pa_9
Other Generative Models							
Macartney and Weyde	WaveUNet	-	La_1	Da_1	Ma_2 Ma_3 Ma_4 Ma_5 Ma_6	2.41 3.54 3.24 2.97 9.98	Pa_2
Kameoka et al.	CVAE	CVAE-VC	La_1, La_3, La_4	Da_5	Ma_{13} Ma_{14}	92.00 77.00	Pa_5
Kameoka et al.	-	ConvS2S	La_1, La_5, La_7, La_8	Da_7	Ma_{13} Ma_{14}	50.00 47.00	Pa_{10}, Pa_{11}, Pa_{12}
Tanaka et al.	-	AttS2S	La_1, La_7, La_9	Da_7	Ma_{13} Ma_{14}	54.00 52.00	Pa_{10}, Pa_{11}, Pa_{12}, Pa_{13}

- M: Metric, RS: Results, RM:Remarks

TABLE 7: Comparison of speech Emotion synthesis models, description of loss functions (L), metrics (M), databases (D) and purposes (P) used in the reviewed publications are given in Tables 8, 9, 10, and 11.

5.

Despite the numerous publications on image and video synthesis, yet some problems are not solved thoroughly. For example, generating high-resolution samples is an open research problem. The output is usually blurry or impaired by checkered artifacts. Results obtained for video generation or synthesis of 3D samples are far from realistic examples. Also, it is important to highlight that the number of publications focused on expression classification is greater than that of those employing identity recognition.

B. SPEECH EMOTION SYNTHESIS

Research efforts focusing on synthesizing speech with emotion effect has continued for more than a decade now. One application of GAN models in speech synthesis is speech enhancement. A pioneer GAN-based model developed for raw speech generation and enhancement is called the Speech Enhancement GAN (SEGAN) proposed by Pascual et al. [27]. SEGAN provides a quick non-recursive framework that works End-to-End (E2E) with raw audio. Learning from different speakers and noise types and incorporating that information to a shared parameterizing system is another contribution of the proposed model. Similar to SEAGAN, Macartney and Weyde [150] proposes a model for speech enhancement based on a CNN architecture called Wave-UNet. The Wave-UNet is used successfully for audio source separation in music and speech de-reverberation. Similar to section III-A, we compare the results of the reviewed papers in Table 7. Additionally, Tables 8 to 11 represent databases, loss functions, assessment metrics and characteristics used in speech synthesis.

Sahu et al. [141] followed a two-fold contribution. First, they train a simple GAN model to learn a high-dimensional feature vector through the distribution of a lower-dimensional representation. Second, CGAN is used to learn the distribution of the high-dimensional feature vectors by conditioning on the emotional label of the target class. Eventually, the generated feature vectors are used to assess the improvement of emotion recognition. They report that using synthesized samples generated by CGAN in the training set is helpful. Also it is concluded that using synthesized samples in the test set suggests the estimation of a lower-dimensional distribution is easier than a high-dimensional complex distribution. Employing the synthesized feature vectors from IEMOCAP database in a cross-corpus experiment on emotion classification of MSP-IMPROV database is reported to be successful.

Mic2Mic [144] is another example of a GAN-based model for speech enhancement. This model addresses a challenging problem called microphone variability. The Mic2Mic model disentangles the variability problem from the downstream speech recognition task and it minimizes the need for training data. Another advantage is that it works with unlabeled and unpaired samples from various microphones. This model defines microphone variability as a data translation from one microphone to another for reducing domain shift between the train and the test data. This model is developed based on CycleGAN to assure that the audio sample [144] from microphone A is translated to a corresponding sample from microphone B.

	Corpus	#Sbj.	#Smp.	#Cls.
Da_1	voice Bank [154]	500	-	$6^{\bullet}$
Da_2	VC Challenge 2016 [155]	10	216	-
Da_3	IEMOCAP [156]	10	7,142	$9^{\star}$
Da_4	FAU-AIBO [157]	51	18,216	$5^{\dagger}$
Da_5	VC Challenge 2018 [158]	20	-	-
Da_6	Japanese speech [159]	1	13,100	-
Da_7	CMU Arctic [160]	4	1,132	-
Da_8	RAVDESS [161]	24	1,440	$8^{\ddagger}$
Da_9	Aurora 4 [162]	-	-	-

- Sbj.: Subject, Smp.: Samples, Cls.: Class
$9^{\star}$: anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state
$5^{\dagger}$: anger, emphatic, neutral, positive, rest
$7^{\diamond}$: anger, anxiety, boredom, disgust, neutral, sadness
$8^{\ddagger}$: neutral, calm, happy, sad, angry, fearful, surprise, and disgust

TABLE 8: List of databases used for speech synthesis in the reviewed publications.

Gao et al. [139] decomposed each speech signal into two codes: a content code that represents emotion-invariant information and a style code that represents emotion-dependent information. The content code is shared across emotion domains and should be preserved while the style code carries domain-specific information and it should change. The extracted content code of the source speech and the style code of the target domain are combined at the conversion step. Finally, they use the GAN model to enhance the quality of the generated speech.

Another widely extended research direction in speech synthesis is Voice Conversion (VC). Hsu et al. [143] proposed a non-parallel VC framework called Variational Autoencoding Wasserstein Generative Adversarial Network (VAWGAN). This method directly incorporates a non-parallel VC criterion into the objective function to build a speech model from unaligned data. VAWGAN improves the synthesized samples with more realistic spectral shapes. Even if VAE-based approaches can work free of parallel data and unaligned corpora, yet they have three drawbacks. First, it is difficult to learn time dependencies in the acoustic feature sequences of source and target speech. Second, the decoder of the VAEs tends to output over-smoothed results. To overcome these limitations, Kameoka et al. [151] adopted fully convolutional neural networks to learn conversion rules that capture short-term and long-term dependencies. Also, by transplanting the spectral details of input speech into its converted version at the test phase, the proposed model avoids producing buzzy speech. Furthermore, in order to prevent losing class information during the conversion process, an information-theoretic regularizer is used.

	Name	Remarks
La_1	$\mathcal{J}_{\text{adv}}$	adversarial loss
La_2	$\mathcal{J}_{\text{cyc}}$	cycle-consistency loss
La_3	$\mathcal{J}_{\text{cls}}$	classification loss
La_4	$\mathcal{J}_{\text{cnd}}$	conditional loss
La_5	$\mathcal{J}_{\text{rec}}$	speech reconstruction loss
La_6	$\mathcal{J}_{\text{id}}$	speaker identity loss
La_7	$\mathcal{J}_{\text{att}}$	guided attention loss
La_8	$\mathcal{J}_{\text{post}}$	PostNet loss
La_9	$\mathcal{J}_{\text{s2s}}$	sequence-to-sequence loss
La_{10}	$\mathcal{J}_{\text{KLD}}$	KLD of output distribution and labels
La_{11}	$\mathcal{J}_{\text{li}}$	linguistic-information loss

TABLE 9: Different losses used for speech synthesis in the reviewed publications.

In 2018, Kaneko and Kameoka made two modifications in CycleGAN architecture to make it suitable for voice conversion task and so the name CycleGAN-VC is selected for the modified architecture. Representing speech by using Recurrent Neural Networks (RNN) is more effective due to the sequential and hierarchical structure of the speech. Howsoever, RNN is computationally demanding considering parallel implementations. As a result, they used gated CNNs that are proven to be successful both in parallelization over sequential data and achieving high performance. The second modification is made by using identity loss to assure preserving linguistic information. Here, a 1D CNN is used as a generator and a 2D CNN as a discriminator to focus on 2D spectral texture.

Later, they released CycleGAN-VC2 [147] which is an improved version of CycleGAN-VC to fill the large gap between the real target and converted speech. Architecture

	Measurement	Remarks
Ma_1	PESQ	Perceptual Evaluation of Speech Quality
Ma_2	MOS	Mean Opinion Score (MOS) for voice quality
Ma_3	CSIG	MOS for prediction of the signal distortion
Ma_4	CBAK	MOS for prediction of the intrusiveness of background noise
Ma_5	COVL	MOS for prediction of the overall effect
Ma_6	SSNR	Segmental Signal to Noise Ratio (SSNR)
Ma_7	SER$_{err}$	speech emotion classification (Error)
Ma_8	SER$_{acc}$	speech emotion classification (Accuracy)
Ma_9	ID$_{err}$	identity classification (Error)
Ma_{10}	ID$_{acc}$	identity classification (Accuracy)
Ma_{11}	-	MOS for speaker similarity
Ma_{12}	-	Preference of Emotion Conversion (%)
Ma_{13}	-	Preference of voice quality (%)
Ma_{14}	-	Preference of speaker similarity (%)
Ma_{15}	MCD	measures the distance between the target and converted speech
Ma_{16}	MSD	measures the local structural differences
Ma_{17}	UAR	Unweighted Average Recall
Ma_{18}	WER	Weighted Error Rate
Ma_{19}	FAD	a VGGish model to evaluate similarity metric

TABLE 10: List of evaluative metrics used for speech synthesis in the reviewed publications.

is altered by using 2-1-2D CNN for the generator and PatchGAN for the discriminator. In addition, the objective function is improved by employing a two-step adversarial loss. It is known that downsampling and upsampling have a severe degradation effect on the original structure of the data. To alleviate this, a 2-1-2D CNN architecture is used in the generator where 2D convolution is used for downsampling and upsampling, and only 1D convolution is used for the main conversion process. Another difference is that while CycleGAN-VC uses a fully connected CNN as its discriminator, CycleGAN-VC2 uses PatchGAN. The last layer of PatchGAN employs convolution to make a patch-based decision for the realness of samples. The difference in objective functions between these two models is reported in Table 7. They report that CycleGAN-VC2 outperforms its predecessor on the same database.

To overcome the shortcomings of CVAE-VC[151] and CycleGAN-VC [146], the StarGAN-VC [145] method combines these two methods to address nonparallel many-to-many voice conversion. While CVAEVC and CycleGAN-VC require to know the attribute of the input speech at the test time, StarGAN does not need any such information. Other GAN-based methods for VC like WaveCycleGAN-VC [148] and WaveCycleGAN-VC2 [149] rely on learning based on filters that prevents quality degradation by overcoming the over-smoothing effect. The over-smoothing effect causes degradation in resolution of acoustic features of the generated speech signal. WaveCycleGAN-VC uses cycle-consistent adversarial networks to convert synthesized speech to natural waveform. The drawback of WaveCycleGAN-VC is aliasing distortion that is avoided in WaveCycleGAN-VC2 by adding identity loss.

Conventional methods like VAEs, cycle-consistent GANs, and StarGAN have a common limitation. Instead of focusing on converting prosodic features like fundamental frequency contour, they focus on the conversion of spectral features frame by frame. A fully convolutional sequence-to-sequence (seq2seq) learning approach is proposed by [152] to solve this problem. Generally, all inputs of a seq2seq model must be encoded into a fixed-length vector. In order to avoid this

	Purpose or characteristic		Purpose or characteristic
Pa_1	tested for data augmentation	Pa_8	generates vocoder-less sounding speech
Pa_2	designed for speech enhancement	Pa_9	alleviates the aliasing effect
Pa_3	designed for non-parallel identity preserving VC	Pa_{10}	voice conversion (VC)
Pa_4	designed for defense against Malicious Adversary	Pa_{11}	fully convolutional sequence-to-sequence
Pa_5	designed for non-parallel many-to-many identity VC	Pa_{12}	modifies prosodic features of voice
Pa_6	performs emotion conversion	Pa_{13}	uses an attention-based mechanism
Pa_7	performance improvement	Pa_{14}	generates spectrograms with high quality

TABLE 11: List of purposes and characteristics used for speech synthesis by reviewed publications.

general limitation of seq2seq models, the authors used an attention-based mechanism that learns where to pay attention in the input sequence for each output sequence. The advantage of seq2seq models is that one can transform a sequence into another variable-length sequence. The proposed model is called ConvS2S [152] and its architecture comprises a pair of source and target encoders, a pair of source and target reconstructors, one target decoder, and a PostNet. The PostNet aims to restore the linear frequency spectrogram from its Mel-scaled version.

Similar to ConvS2S is ATTS2S-VC [153] that employs attention and context preservation mechanisms in a Seq2Seq-based VC system. Although this method addresses the aforementioned problems, yet it has a lower performance in comparison to CVAE-VC, CycleGAN-VC, CycleGAN-VC2, and StarGAN. An ablation study is required to evaluate each component of seq2seq methods considering performance degradation.

Despite the promising performance of deep neural networks, they are highly susceptible to malicious attacks that use adversarial examples. One can develop an adversarial example through the addition of unperceived perturbation with the intention of eliciting wrong responses from the machine learning models. Latif et al. [28] conducted a study on how adversarial examples can be used to attack speech emotion recognition (SER) systems. They propose the first black-box adversarial attack on SER systems that directly perturbs speech utterances with small and imperceptible noises. Later, the authors perform emotion classification to clean audio utterances by removing that adversarial noise using a GAN model to show that GAN-based defense stands better against adversarial examples. Other examples of malicious attacks are simulating spoofing attacks [163] and cloning Obama's voice using GAN-based models and low-quality data [164].

The next target application of speech synthesis is data augmentation. Data augmentation is the task of increasing the amount and diversity of data to compensate for the lack of data in certain cases. Data augmentation can improve the generalization behavior of the classifiers. Despite its importance, only a few papers contributed fully toward this concept.

One of the researches on data augmentation for the purpose of SER improvement is the work of Sheng et al. [140]. The authors used a variant of CGANs model that works at frame level and uses two different conditions. The first condition is the acoustic state of each input frame that is combined as a one-hot vector with the noise input and fed into the generator. The same vector is combined with real noisy speech and fed into the discriminator. The second condition is the pairing of speech samples during the training process. In fact, parallel paired data is used for training. For example, original and clean speech is paired with manually added noisy speech or close-talk speech sample is paired with far-field recorded speech. The discriminator learns the naturalness of the sample based on the paired data.

Another study with more focus on the improvement of SER is done by Chatziagapi et al. [142]. They adopt a CGAN called Balancing GAN (BAGAN) [165] and improve it to generate synthetic spectrograms for the minority or under-represented emotion classes. The authors modified the architecture of BAGAN by adding two dense layers to the original generator. These layers project the input to higher dimensionality. Also, the discriminator is changed by using double strides to increase the height and width of the intermediate tensors which affect the quality of the generated spectrogram remarkably.

Other interesting applications like cross-language emotion transfer and singing voice synthesis are also investigated by various researches. However, these applications are not thoroughly studied and they have plenty of potential for further research. One such example is ET-GAN [166]. This model uses a cycle-consistent GAN to learn language-independent emotion transfer from one emotion to another while it does not require parallel training samples.

Also, some works are dedicated to speech synthesis in the frequency domain. Long-range dependencies are difficult to model in the time domain. For instance, MelNet model [167] proves that such dependencies can be more tractably modeled in two dimensional (2D) time-frequency representations such as spectrograms. By coupling the 2D spectrogram representation and an auto-regressive probabilistic model with a multi-scale generative model, they synthesized high fidelity audio samples. This model captures local and global structures at time scales that time-domain models have yet to achieve. In a Mean Opinion Score (MOS) comparison between MelNet and WaveNet, MelNet won with a 100% vote for a preference of the quality of the sample.

In the case of speech synthesis in feature-domain, several pieces of research are represented under VC application. For instance, Juvela et al. [168] proposed generating speech from filterbank Mel-frequency cepstral coefficients (MFCC). The method starts by predicting the fundamental frequency (F_0) and the intonation information from MFCC using an auto-regressive model. Then, a pitch synchronous excitation model is trained on the all-pole filters obtained in turn from spectral envelope information in MFCCs. In the end, a residual GAN-based noise model is used to add a realistic high-frequency stochastic component to the modeled excitation signal. Degradation Mean Opinion Score (DMOS) is used to evaluate the quality of synthesized speech samples.

1) Evaluation Metrics for Speech

In order to evaluate the local and global structures of the generated samples, various metrics are employed by the researchers. In general, metrics like MOS and Perceptual Evaluation of Speech Quality (PESQ) are used widely, while other efficient metrics like Mel-cepstral distortion (MCD) and modulation spectra distance (MSD) are less employed in the literature. Following, we provide a brief explanation of each metric with the hope of a more cohesive future comparison in the literature.

	SGI	BAK	OVRL
1	very unnatural/degraded	very conspicuous/intrusive	bad
2	fairly unnatural/degraded	fairly conspicuous, somewhat intrusive	poor
3	somewhat natural/degraded	noticeable but not intrusive	regular
4	fairly natural, little degradation	somewhat noticeable	good
5	very natural, no degradation	not noticeable	excellent

TABLE 12: Description of SIG, BAK, and OVRL scales used in the subjective listening tests.

MOS is a quality rating method that works based on the subjective quality evaluation test. The quality is assessed by human subjects for a given stimulus. The user rates its Quality of Experience (QoE) as a number within a categorical range with “bad" being the lowest perceived quality 1 and 5 being “Excellent" or the highest perceived quality. MOS is expressed as the arithmetic mean over QoEs.

$$\text{MOS} = \frac{1}{N}\sum_{n=1}^{N} r_n, \tag{19}$$

where N is the total number of subjects contributed to the evaluation and r_n is the QoE of the subject considering the stimuli. MOS is subject to certain biases. Number of subjects in the test and content of the samples under assessment are some of the problems. The ordinal categories code a wide range of perceptions. That is why MOS is considered to be an absolute measure of total quality, regardless of any specific quality dimension. This is useful in many applications related to communications. However, for other applications, a measure that could be more sensitive to specific quality dimensions is more suitable. Other biases include, changing the user expectation about quality over time, and the value of the smallest MOS difference that is perceptible to users and can actually claim if one method is better over another. For example, Pascual et al. [27] and Macartney and Weyde [150] achieved an MOS of 3.18 and 2.41 on the same database which provides a naive comparison of .77 MOS difference in favor of the former method. Howsoever, one question here is that whether the sample tests and the number of subjects were the same. This becomes more interesting by comparison of the methods proposed in Tanaka et al. [148] and Tanaka et al. [149] where the authors achieved only a .11 MOS difference using the same database.

MOS is time-consuming though, it is applicable to different quantities. For instance, likewise MOS of voice quality, the MOS of signal distortion and MOS for intrusiveness of background noise are used as a metric. PESQ is an objective speech quality assessment based on subjective quality ratings. In fact, it automatically calculates what is MOS of subjective perception. PESQ integrates disturbance over several time frequency scales. This is applied by using a method that take soptimal account of the distribution of error in time and amplitude [169, 170]. The disturbance values are aggregated using the L_p (or simply p-norm) as (20):

$$L_p = \left(\frac{1}{N}\sum_{m=1}^{N} \left|\text{disturbance}[m]\right|^p\right)^{\frac{1}{p}} \tag{20}$$

where N is the total number of frames summing disturbance across frequency using an L_p norm gives a frame-by-frame measure of perceived distortion. The subjective listening tests were designed to reduce the effect of uncertainty arising from the listener's decision by highlighting which of the three components of a noisy speech signal should form the basis of their ratings of overall quality.

These components are the speech signal, the background noise, or both. In this method the listener successively attends and rates the synthesized speech sample on: a) the speech signal alone using a five-point scale of signal distortion (SIG), b) the background noise alone using a five-point scale of background intrusiveness (BAK), c) the overall quality using the scale of the mean opinion score (OVRL). The SIG, BAK, and OVRL scales are described in Table 12.

Two widely used objective measurements for speech intelligibility are the Signal to Noise Ratio (SNR) and Segmental Signal to Noise Ratio (SSNR/SegSNR). The SNR can be expressed as (21):

$$\text{SNR} = 10\log_{10}\frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N}(x_i - y_i)^2}, \tag{21}$$

where $x(i)$ and $y(i)$ are the ith real and synthesized samples and N is the total number of samples.

The SSNR/SegSNR can be expressed as the average of the SNR values of short segments (15 to 20 ms). It can be expressed as (22):

$$\text{SSNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\sum_{i=Nm}^{Nm+N-1}\left(\frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N}(x_i - y_i)^2}\right) \tag{22}$$

where N and M are the segment length and the number of segments, respectively. SSNR tends to provide better results than SNR for waveform encoders and generally SSNR results are poor on vocoders.

Other objective measurements include MCD that evaluates the distance between the target and converted Mel-cepstral coefficients (MCEP) sequences. Also, MSD assesses the local structural differences by calculating the root mean square error between the target and converted logarithmic modulation spectra of MCEPs averaged over all MCEP dimensions and modulation frequencies. For both metrics, smaller values

Author	Based on	Model	Loss	Data	M	RS	RM
Original GANs							
Wiles et al.	GAN	X2Face	Lav_2, Lav_3, Lav_4	Dav_2	Mav_3	0.0521	Pav_2
Duarte et al.	CGAN (SEAGAN)	Wav2Pix	Lav_1, Lav_2	Dav_4	Mav_4	-	Pav_1
GAN Variants by Discriminator							
Vougioukas et al.	TGAN RNN	-	Lav_3	Dav_3	Mav_5 Mav_6	27.98 0.844	Pav_1
Other Generative Models							
Ephrat and Peleg	-	-	Lav_2	Dav_1	Mav_1	79.9	Pav_1
Ephrat et al.	-	-	Lav_2	Dav_1	Mav_2	1.97	Pav_1
Suwajanakorn et al.	RNN	-	-	Dav_1	Mav_4	-	Pav_3
Jamaludin et al.	-	Vid2Speech	Lav_1, Lav_3, Lav_7	Dav_2	Mav_3	327	Pav_4

- M: Metric, RS: Results, RM:Remarks

TABLE 13: Comparison of cross-modal Emotion synthesis models, description of loss functions (L), metrics (M), databases (D) and purposes (P) used in the reviewed publications are given in Tables 14 to 17.

indicate higher distortion between the real and converted speech. It is important to highlight that some of the successful methods like GANSynth [171] are not mentioned in this paper as it focuses on the musical note synthesis.

C. AUDIO-VISUAL EMOTION SYNTHESIS

Although GAN models have an impressive performance on single-domain and cross-domain generation, yet they did not achieve much success in cross-modal generation due to the lack of a common distribution between heterogeneous data. In a cross-domain generation, one generates data samples of various styles from the same modality. As a result, the generated sample and its original counterpart have a common shape structure. However, in a cross-modal generation, the pair of samples have heterogeneous features with quite different distributions.

In this section, we investigate the cross-modal research line where audio and video provide applications like talking heads, audio-video synchronization, facial animations, and visualizing the face of an unseen subject from their voice. Note that other modalities like text [70, 177, 178] and biological signals [179] are used in combination with audio and video. However, those modalities are beyond the scope of this review paper. Ephrat and Peleg [174] proposed the Vid2Speech model that uses neighboring video frames to generate sound features for each frame. Then, speech waveforms are synthesized from the learned speech features. In 2017, the authors designed a two-tower CNN [175] framework that reconstructs a natural-sounding speech signal from silent video frames of the speaking person. Their model shows that using one modality to generate samples of another modality is indeed useful because it provides the possibility of natural supervision which means segmentation of the recorded video frames and the recorded sound is not required. The two-tower CNN relies on improving the performance of a Residual neural network (ResNet) that is used as an encoder and redesigning a CNN-based decoder.

	Dataset	#Sbj.	#Smp.
Dav_1	GRID [180]	4	-
Dav_2	VoxCeleb [181]	1,251	21,245
Dav_3	Obama's video footage	-	-
Dav_4	Youtubers [32]	62	4,860

- Sbj.: Subject, Smp.: Samples

TABLE 14: List of databases used for cross-modal synthesis in the reviewed publications.

	Name	Remarks
Lav_1	$\mathcal{L}_{\text{adv}}$	adversarial loss
Lav_2	$\mathcal{L}_{\text{mse}}$	mean squared error loss
Lav_3	$\mathcal{L}_{\text{pixel}}$	pixelwise L1 loss
Lav_4	$\mathcal{L}_{\text{id}}$	identity loss
Lav_5	$\mathcal{L}_{\text{frame}}$	frame loss
Lav_6	$\mathcal{L}_{\text{seq}}$	sequential loss
Lav_7	$\mathcal{L}_{\text{cnt}}$	content loss

TABLE 15: Different losses used for cross-modal synthesis in the reviewed publications.

One of the foremost cross-modal GAN-based models is proposed by Chen et al. [35]. They explored the performance of CGANs by using various audio-visual encodings on generating sound/player of a musical instrument from the pose of the player or the sound of the instrument. This model is not tested on any emotional database and hence, it is not listed in Table 13. Another leading and interesting research work is conducted by Suwajanakorn et al. [176]. An RNN is trained on weekly audio footage of President Barack Obama to map raw audio features to mouth shapes. In the end, a high-quality video with accurate lip synchronization is synthesized. The model can control fine-details like lip texture and mouth-pose.

	Measurement	Remarks
Mav_1	SER_{acc}	speech emotion classification (Accuracy)
Mav_2	PESQ	automated quality evaluation
Mav_3	error	L1 reconstruction error
Mav_4	-	qualitative/visual representation
Mav_5	PSNR	Peak Signal to Noise Ratio
Mav_6	SSIM	measures image quality degradation
Mav_7	ACD	content consistency of a generated video

TABLE 16: List of evaluative metrics used for cross-modal synthesis in the reviewed publications.

Speech-driven video synthesis is the next application. The X2Face model proposed by Wiles et al. 2018 uses a facial photo or another modality sample (e.g. audio) to modify the pose and expression of a given face for video/image editing. They train the model in a self-supervised fashion by receiving two samples: a source sample (video) and a driving sample (video, audio or, a combination). The generated sample inherits the same identity and style (e.g. hairstyle) from the source sample and gets the pose, expression of the driving sample. The authors employed an embedding network that factorizes the face representation of the source sample and applies face frontalization. Unfortunately, the authors reported only the visual generated samples and, no further metric is used as of comparison.

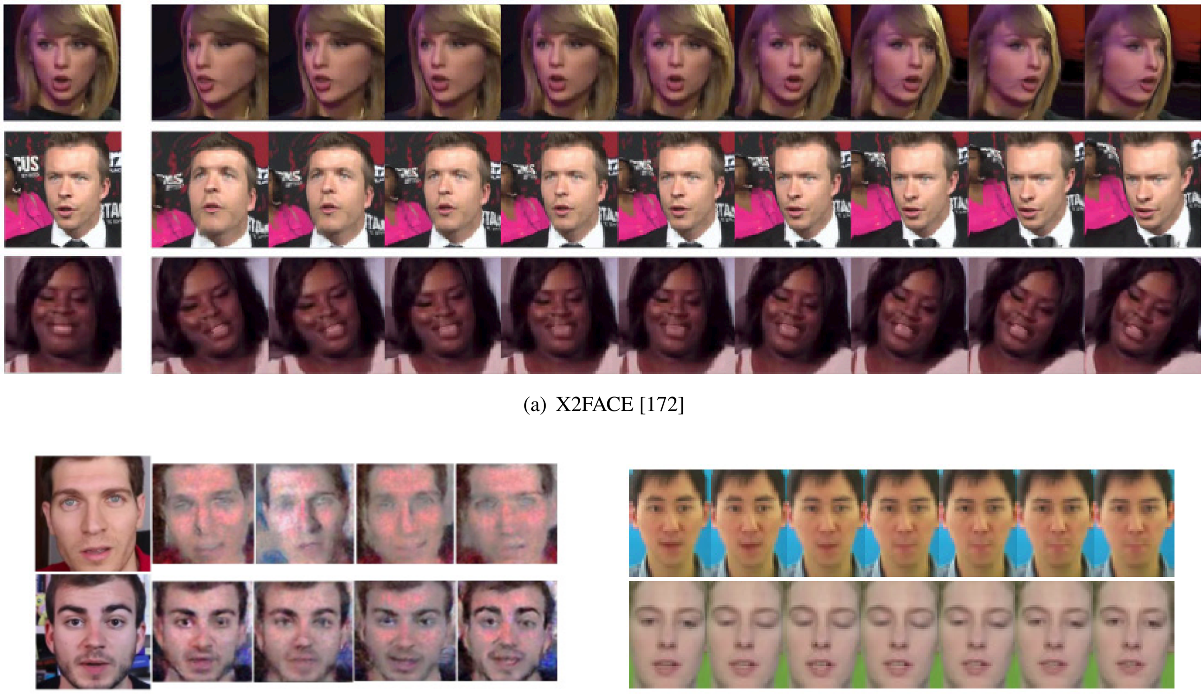
Another noteworthy work in speech-driven video synthesis is the one presented by Vougioukas et al. [173]. They suggested an E2E temporal GAN that captures the facial dynamics and generates synchronized mouth movements and fine-detailed facial expressions, such as eyebrow raises, frowns, and blinks. The authors paired still images of a person with an audio speech to generate subject independent realistic videos. They use raw speech signal as the audio input. The model includes one generator comprising an RNN-based audio encoder, an identity image encoder, a frame decoder, and a noise generator. Also, there exist two discriminators: frame discriminator that simply classifies the frames into real and fake, and sequence discriminator distinguishing real videos from the fake ones. Evaluating the generated samples in frame-level and video-level helps to generate high-quality frames while the video remains synchronized with audio. In 2019, Vougioukas et al. [182] modified their previous work to generate speech-driven facial animations. This E2E model has the capability of generating synchronized lip movements with the speech audio and it has fine control over facial expressions like blinking and eyebrow movement.

Duarte et al. [32] proposed Wav2Pix model generating the facial image of a speaker without prior knowledge about the face identity. This is done by conditioning on the raw speech signal of that person. The model uses Least-Square GAN and SEAGAN to preserve the identity of the speaker half of the time. The generated images by this model are of low quality (see Fig. 10). Also, the model is sensitive to several factors like the dimensionality and quality of the training images and the duration of the speech chunk. Likewise, [34] in “You said that? : Synthesising Talking Faces from Audio" designed the Speech2Vid model that gets the still images of the target face and an audio speech segment as input. The model synthesizes a video of the target face that has a synchronized lip with the speech signal. This model consists of a VGG-M as an audio encoder, a VGG-Face as an identity image encoder, and a VGG-M in reverse order as a talking face image decoder. Here instead of raw speech data, the audio encoder uses MFCC heatmap images. The network is trained with a usual adversarial loss between the generated image and the ground truth and a content representation loss.

One of the most important models developed in the cross-modal community is the SyncGAN model [183] capable of successfully generating synchronous data. A common problem of the aforementioned cross-modal GAN models is that they are one-directional because they learn the transfer between different modalities. This means they cannot generate a pair of synchronous data from both modalities simultaneously. SyncGAN addresses this problem by learning in a bidirectional mode and from synchronous latent space representing the cross-modal data. In addition to the general generator and discriminator of the vanilla GAN, the model

	Purpose or characteristic		Purpose or characteristic
Pav_1	read speech	Pav_4	generating unseen face of subject using raw speech
Pav_2	controlling the pose and expression of face based on audio modality	Pav_5	generating lip synchronized video of a talking face
Pav_3	generating high quality mouth texture	Pav_6	generating an intelligible speech signal from silent video of a speaking person

TABLE 17: List of purposes and characteristics used for cross-modal synthesis by reviewed publications.



(a) X2FACE [172]

(b) WAV2PIX [32]

(c) Vougioukas et al. [173]

FIGURE 10: Visual comparison of the cross-modal GAN models, images are in courtesy of the reviewed papers.

uses a synchronizer network for estimating the probability that two input data are from the same concept. This network is trained using synchronous and asynchronous data samples to maximize the following loss function (23):

$$\begin{aligned}\mathcal{L}_S = &\mathbb{E}_{\boldsymbol{x}_1\sim p_r(\boldsymbol{x}_1),\boldsymbol{x}_2\sim p_r(\boldsymbol{x}_2)}[\log S(\boldsymbol{x}_1^i,\boldsymbol{x}_2^j) \mid i = j] \\ &+\mathbb{E}_{\boldsymbol{x}_1\sim p_r(\boldsymbol{x}_1),\boldsymbol{x}_2\sim p_r(\boldsymbol{x}_2)}[\log(1 - S(\boldsymbol{x}_1^i,\boldsymbol{x}_2^j)) \mid i \neq j]\end{aligned} \tag{23}$$

Similarly, CMCGAN [184] is a cross-modal CycleGAN that handles generating mutual generation of cross-modal audio-visual videos. Given an image/sound sample from a musical instrument outputs a sound LMS or an image of the player. Unfortunately, neither SyncGAN nor CMCGAN are tested on any multi-modal emotional database. In Fig. 10, we compare the generated samples of the reviewed publications qualitatively.

IV. DISCUSSION

In this section, we discuss the concepts that are yet not explored thoroughly about GAN-based emotion synthesis within the literature. Also, despite the active development of GAN models, there exist open research problems like mode collapse, convergence failure, and vanishing gradients. Following we discuss these problems. Also, the evolution of GAN models is shown in Fig. 11 (see next page).

A. DISADVANTAGES

The most important drawback of GAN models is mode drop or mode collapse. Mode collapse occurs when a generator learns to generate a limited variety of samples out of the many modes available in the training data. Roth et al. [185] attempted to solve the mode collapse problem by stabilizing the training procedure using regularization. Numerical analysis of general algorithms for training GAN showed that not all training methods actually converge [186], which leads to the mode collapse problem. Several objective functions [51] and structures [187] are developed to tackle this problem, however, none have solved the problem thoroughly.

GANs also suffer from convergence failure. Convergence failure happens when the model parameters oscillate and they cannot stabilize during training. In the minimax game, convergence occurs when the discriminator and the generator reach the optimal point under Nash equilibrium theorem. Nash equilibrium is defined as the situation where one player will not change action irrespective of opponent action.

It is known that if the discriminator performs too accurately, the generator fails due to the vanishing gradient. In fact, the discriminator does not leak/provide enough information for the generator to continue the learning process.

B. OPEN RESEARCH PROBLEMS

In addition to the theoretical problems mentioned in section IV-A, GANs have task-based limitations. For instance, GANs cannot synthesize discrete data like one-hot coded vectors. Although this problem is addressed partially in some research works [188, 189, 190], yet it needs more attention to unlock the full potential of GAN models. A series of novel divergence algorithms like FisherGAN [191] and the model proposed by Mroueh et al. [192] try to improve the convergence for training GANs. This area deserves more exploration by studying families of integral probability metrics.

The objective of a GAN model is to generate new samples that come from the same distribution as the training data. However, they do not generate the distribution that generated the training examples. As a result, they do not have a prior likelihood or a well-defined posterior. The question here is how can one estimate the uncertainty of a well-trained generator.

Considering the emotion synthesis domain, some problems are studied partially. First of all, data augmentation is not fully explored. At the time of writing this paper, there is no large scale image database generated artificially by using GAN models and released for public usage. Such a database could be compared in terms of classification accuracy and quality with the existing databases. Although methods like GANimation and StarGAN successfully generate all sorts of facial expressions, yet generating a fully labeled database requires further processing. For example, the synthesized samples should be annotated and tested against a ground truth like facial Action Units (AU) to confirm that the generated samples carry the predefined standards of a specific emotional class. This issue becomes very complicated when one deals with compound emotions and not only the basic discrete emotions. Also, generated samples are not evaluated within continuous space considering the arousal, valence, and dominance properties of the emotional state. Finally, despite the fact that some successful GAN models are proposed for video generation, the results are not realistic.

In the case of speech emotion synthesizing, the majority of papers focused on raw speech and spectrograms. As a result, feature-based synthesis is less explored. Human-likeliness of the generated speech samples is another open discussion in this research direction. Furthermore, evaluation metrics in this field are less developed and merely the ones from the traditional speech processing are used on the generated results. Research works that are focusing on cross-modal emotion generation do not exceed from few publications. This research direction requires both developing new ideas and improving the result of previous models.

C. APPLICATIONS

One important application of GAN models in the computer vision society includes synthesizing Super Resolution (SR) or photo-realistic images. For example, SRGAN [193] and Enhanced SRGAN [194] are generating photo-realistic natural images for an upscaling factor. Considering facial synthesis, these applications include manipulation of facial pose using DRGAN [195] and TPGAN [196], generating a facial portrait [197], generating the face of an artificial subject or manipulating the facial attributes of a specific subject [46], [23], and synthesizing/manipulating fine detail facial features like skin, lip or teeth texture [176]. Generally speaking, the application of GANs considering the visual modality could be categorized into texture synthesis, image super-resolution, image inpainting, face aging, face frontalization, human image synthesis, image-to-image translation, text-to-image, sketch-to-image, image editing, and video generation. Some specific applications with respect to the emotional video generation include the synthesizing of talking heads [24], [120].

In the case of speech emotion synthesis, as mentioned before, these applications can be categorized into speech enhancement, data augmentation, and voice conversion. Other research directions like feature learning, imitation learning, and reinforcement learning are important research directions for the near future.

V. CONCLUSION

In this paper, we survey the state-of-the-art of GAN models applied to human emotion synthesis GAN models proposed first in 2014 by Goodfellow et al.. The core idea of GAN models is based on a zero-sum game in game theory. Generally, a GAN model consists of a generator and a discriminator, which are trained iteratively in an adversarial learning manner, approaching Nash equilibrium. Instead of estimating the distribution of real data samples, GANs learn to synthesize samples that adapt to the distribution of real data samples. More specifically, facial expression synthesis can be performed as an image-based or a video-based downstream task. It is observed that using conditional information like the identity or facial features of the subject leads to improved quality. In addition, video-based methods are more realistic when motion content is employed. In the case of speech emotion synthesis, preserving the identity of the subject helps to improve the results while problems like robotic and noisy sound effects in the generated samples require to be addressed. Cross-modal emotion synthesis is a fresh topic in this field that can be improved by answering tempting questions about intrinsic relationships between various modalities. Finally, this paper discussed some open research problems and the great potential of GANs in terms of research and development in human emotion synthesis.

ACKNOWLEDGEMENTS

We would like to thank Eastern Mediterranean University for supporting this research work through the BAP-C project under the grant number BAP-C-02-18-0001. This work was partially supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grants 2018/26455-8, 2020/01928-0, 2018/12579-7 and 2019/07665-4.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

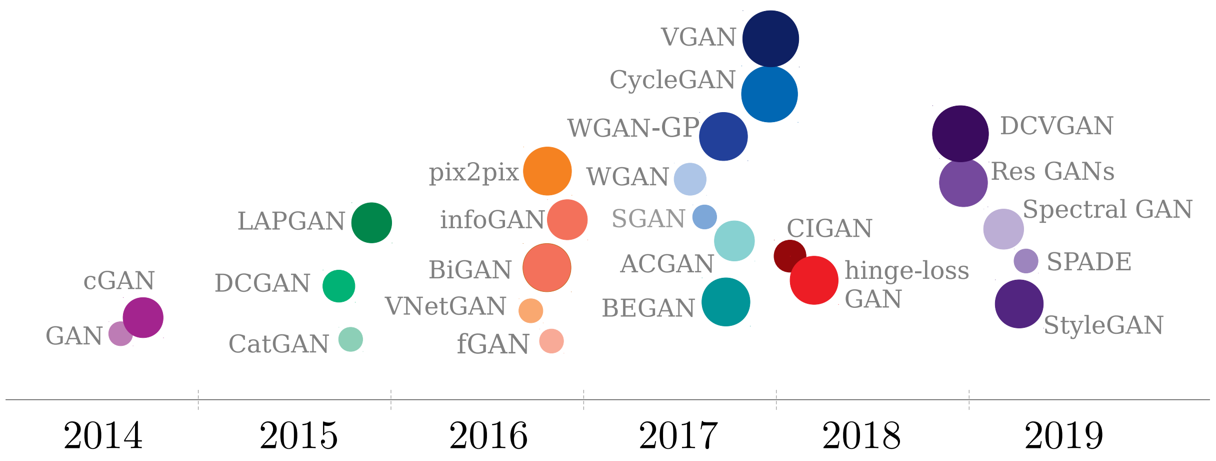


FIGURE 11: Evolution of GAN models, the horizontal line shows the release year of the model, each specific year is shown using shades of one color

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[4] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, “A review on generative adversarial networks: Algorithms, theory, and applications,” arXiv preprint arXiv:2001.06937, 2020.

[5] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, “Generative adversarial networks: introduction and outlook,” IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 588–598, 2017.

[6] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, “Recent progress on generative adversarial networks (GANs): A survey,” IEEE Access, vol. 7, pp. 36 322–36 333, 2019.

[7] M. Zamorski, A. Zdobylak, M. Zieba, and J. Świątek, “Generative adversarial networks: recent developments,” in International Conference on Artificial Intelligence and Soft Computing. Springer, 2019, pp. 248–258.

[8] S. Hitawala, “Comparative study on generative adversarial networks,” arXiv preprint arXiv:1801.04271, 2018.

[9] Z. Wang, Q. She, and T. E. Ward, “Generative adversarial networks: A survey and taxonomy,” ArXiv, vol. abs/1906.01529, 2019.

[10] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, “How generative adversarial networks and their variants work: An overview,” ACM Computing Surveys (CSUR), vol. 52, no. 1, pp. 1–43, 2019.

[11] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53–65, 2018.

[12] H. Huang, P. S. Yu, and C. Wang, “An introduction to image synthesis with generative adversarial nets,” arXiv preprint arXiv:1803.04469, 2018.

[13] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly, “The GAN landscape: Losses, architectures, regularization, and normalization,” ArXiv, 2018.

[14] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” Medical image analysis, p. 101552, 2019.

[15] N. Torres-Reyes and S. Latifi, “Audio enhancement and synthesis using generative adversarial networks: A survey,” International Journal of Computer Applications, vol. 975, p. 8887, 2019.

[16] X. Wu, K. Xu, and P. Hall, “A survey of image synthesis and editing with generative adversarial networks,” Tsinghua Science and Technology, vol. 22, no. 6, pp. 660–674, 2017.

[17] J. Agnese, J. Herrera, H. Tao, and X. Zhu, “A survey and taxonomy of adversarial neural networks for text-to-image synthesis,” arXiv preprint arXiv:1910.09399, 2019.

[18] A. Schirmer and R. Adolphs, “Emotion perception from face, voice, and touch: comparisons and convergence,” Trends in Cognitive Sciences, vol. 21, no. 3, pp. 216–228, 2017.

[19] P. Ekman, W. V. Friesen, and M. O’sullivan, “Smiles when lying.” Journal of Personality and Social Psychology, vol. 54, no. 3, p. 414, 1988.

[20] M. Zuckerman, D. T. Larrance, N. H. Spiegel, and R. Klorman, “Controlling nonverbal displays: Facial expressions and tone of voice,” Journal of Experimental Social Psychology, vol. 17, no. 5, pp. 506–524, 1981.

[21] A. Mehrabian and S. R. Ferris, “Inference of attitudes from nonverbal communication in two channels.” Journal of Consulting Psychology, vol. 31, no. 3,

p. 248, 1967.

[22] H. Ding, K. Sricharan, and R. Chellappa, “ExprGAN: Facial expression editing with controllable expression intensity,” in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.

[24] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, “MocoGAN: Decomposing motion and content for video generation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1526–1535.

[25] C. Kervadec, V. Vielzeuf, S. Pateux, A. Lechervy, and F. Jurie, “Cake: Compact and accurate k-dimensional representation of emotion,” arXiv preprint arXiv:1807.11215, 2018.

[26] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, “Deep video portraits,” ACM Transactions on Graphics (TOG), vol. 37, no. 4, pp. 1–14, 2018.

[27] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” arXiv preprint arXiv:1703.09452, 2017.

[28] S. Latif, R. Rana, J. Qadir, and J. Epps, “Variational autoencoders for learning latent representations of speech emotion: A preliminary study,” arXiv preprint arXiv:1712.08708, 2017.

[29] J. Gideon, M. McInnis, and E. M. Provost, “Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG),” IEEE Transactions on Affective Computing, 2019.

[30] X. Zhou and W. Y. Wang, “Mojitalk: Generating emotional responses at scale,” arXiv preprint arXiv:1711.04090, 2017.

[31] K. Wang and X. Wan, “SentiGAN: Generating sentimental texts via mixture adversarial networks.” in IJCAI, 2018, pp. 4446–4452.

[32] A. Duarte, F. Roldan, M. Tubau, J. Escur, S. Pascual, A. Salvador, E. Mohedano, K. McGuinness, J. Torres, and X. Giro-i-Nieto, “Wav2Pix: Speech-conditioned face generation using generative adversarial networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, 2019.

[33] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, “Audio-driven facial animation by joint end-to-end learning of pose and emotion,” ACM Transactions on Graphics (TOG), vol. 36, no. 4, pp. 1–12, 2017.

[34] A. Jamaludin, J. S. Chung, and A. Zisserman, “You said that?: Synthesising talking faces from audio,” International Journal of Computer Vision, vol. 127, no. 11-12, pp. 1767–1779, 2019.

[35] L. Chen, S. Srivastava, Z. Duan, and C. Xu, “Deep cross-modal audio-visual generation,” in Proceedings of the on Thematic Workshops of ACM Multimedia 2017, 2017, pp. 349–357.

[36] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” arXiv preprint arXiv:1312.6114, 2013.

[37] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” arXiv preprint arXiv:1401.4082, 2014.

[38] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.

[39] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016.

[40] A. Yadav, S. Shah, Z. Xu, D. Jacobs, and T. Goldstein, “Stabilizing adversarial nets with prediction methods,” arXiv preprint arXiv:1705.07364, 2017.

[41] E. L. Denton, S. Chintala, R. Fergus et al., “Deep generative image models using a laplacian pyramid of adversarial networks,” in Advances in Neural Information Processing Systems, 2015, pp. 1486–1494.

[42] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, “Stacked generative adversarial networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5077–5086.

[43] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” arXiv preprint arXiv:1710.10196, 2017.

[44] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.

[45] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8798–8807.

[46] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.

[47] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” arXiv preprint arXiv:1802.05957, 2018.

[48] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of Wasserstein GANs,” in Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.

[49] T. Miyato and M. Koyama, “cGANs with projection discriminator,” arXiv preprint arXiv:1802.05637, 2018.

[50] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," arXiv preprint arXiv:1511.06390, 2015.

[51] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," arXiv preprint arXiv:1703.10717, 2017.

[52] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in Advances in Neural Information Processing Systems, 2016, pp. 2172–2180.

[53] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in Proceedings of the 34th International Conference on Machine Learning, vol. 70. JMLR. org, 2017, pp. 2642–2651.

[54] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in Advances in Neural Information Processing Systems, 2017, pp. 3856–3866.

[55] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, "CapsuleGAN: Generative adversarial capsule network," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[56] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," arXiv preprint arXiv:1701.07875, 2017.

[57] J. Niu, Z. Li, S. Mo, and B. Fan, "CIGAN: A novel GANs model based on category information," in 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, 2018, pp. 522–529.

[58] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2794–2802.

[59] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in Advances in Neural Information Processing Systems, 2016, pp. 271–279.

[60] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," arXiv preprint arXiv:1609.03126, 2016.

[61] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," arXiv preprint arXiv:1605.09782, 2016.

[62] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Emotional facial expression transfer from a single image via generative adversarial nets," Computer Animation and Virtual Worlds, vol. 29, no. 3-4, p. e1819, 2018.

[63] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[64] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[65] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," arXiv preprint arXiv:1512.09300, 2015.

[66] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," arXiv preprint arXiv:1606.00704, 2016.

[67] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in Advances in neural information processing systems, 2017, pp. 700–708.

[68] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 565–571.

[69] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in Proceedings of the 34th International Conference on Machine Learning, vol. 70. JMLR. org, 2017, pp. 1857–1865.

[70] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," arXiv preprint arXiv:1605.05396, 2016.

[71] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1316–1324.

[72] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson, "Generating facial expressions with deep belief nets," Affective Computing, Emotion Modelling, Synthesis and Recognition, pp. 421–440, 2008.

[73] P. Ekman and W. Friesen, Facial Action Coding System. Consulting Psychologists Press, 1978, no. v. 1. [Online]. Available: https://books.google.com.cy/books?id=08l6wgEACAAJ

[74] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikälnen, "Facial expression recognition from near-infrared videos," Image and Vision Computing, vol. 29, no. 9, pp. 607–619, 2011.

[75] R. Gross, I. Matthews, J. Cohn, T. Kanade, and

S. Baker, "Multi-pie," Image and Vision Computing, vol. 28, no. 5, pp. 807–813, 2010.

[76] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2010, pp. 94–101.

[77] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in 7th International Conference on Automatic Face and Gesture Recognition (FGR06). IEEE, 2006, pp. 211–216.

[78] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the Radboud Faces Database," Cognition and Emotion, vol. 24, no. 8, pp. 1377–1388, 2010.

[79] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.

[80] C. F. Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez, "Emotionet challenge: Recognition of facial expressions of emotion in the wild," arXiv preprint arXiv:1703.01210, 2017.

[81] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2017.

[82] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., "Challenges in representation learning: A report on three machine learning contests," in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124.

[83] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE, 2011, pp. 2106–2112.

[84] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The Japanese female facial expression (JAFFE) database," in Proceedings of Third International Conference on Automatic Face and Gesture Recognition, 1998, pp. 14–16.

[85] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in International Conference on Neural Information Processing, 2008.

[86] W. Wang, Y. Fu, Q. Sun, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, Y.-G. Jiang, and X. Xue, "Learning to augment expressions for few-shot fine-grained facial expression recognition," arXiv preprint arXiv:2001.06144, 2020.

[87] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10. IEEE, 2010, pp. 1–4.

[88] Y. Huang and S. M. Khan, "DyadGAN: Generating facial expressions in dyadic interactions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 11–18.

[89] S. G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B. R. Abidi, A. Koschan, M. Yi, and M. A. Abidi, "Multiscale fusion of visible and thermal IR images for illumination-invariant face recognition," International Journal of Computer Vision, vol. 71, no. 2, pp. 215–233, 2007.

[90] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in 2005 IEEE International Conference on Multimedia and Expo. IEEE, 2005, pp. 5–pp.

[91] B. Bozorgtabar, D. Mahapatra, and J.-P. Thiran, "Exprada: Adversarial domain adaptation for facial expression analysis," Pattern Recognition, vol. 100, p. 107111, 2020.

[92] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska directed emotional faces (KDEF)," CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet, vol. 91, no. 630, pp. 2–2, 1998.

[93] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7093–7102.

[94] S. Cheng, M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, and S. Zafeiriou, "MeshGAN: Non-linear 3D morphable models of faces," arXiv preprint arXiv:1903.10384, 2019.

[95] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4DFAB: A large scale 4D database for facial expression analysis and biometric applications," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5117–5126.

[96] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3911–3919.

[97] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "TV-GAN: Generative adversarial network based thermal to visible face recognition," in 2018 International Conference on Biometrics (ICB). IEEE, 2018, pp. 174–181.

[98] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geome-

try guided adversarial facial expression synthesis," in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 627–635.

[99] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional CycleGAN," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 282–297.

[100] Y. Peng and H. Yin, "ApprGAN: Appearance-based GAN for facial expression synthesis," IET Image Processing, vol. 13, no. 14, pp. 2706–2715, 2019.

[101] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "CollaGAN: Collaborative GAN for missing image data imputation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2487–2496.

[102] T. Caramihale, D. Popescu, and L. Ichim, "Emotion classification using a tensorflow generative adversarial network implementation," Symmetry, vol. 10, no. 9, p. 414, 2018.

[103] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion classification with data augmentation using generative adversarial networks," in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2018, pp. 349–360.

[104] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 263–270.

[105] A. Lindt, P. Barros, H. Siqueira, and S. Wermter, "Facial expression editing with continuous emotion labels," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019, pp. 1–8.

[106] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," IEEE Transactions on Image Processing, vol. 28, no. 11, pp. 5464–5478, 2019.

[107] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 821–830.

[108] V. Vielzeuf, C. Kervadec, S. Pateux, and F. Jurie, "The many variations of emotion," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019, pp. 1–7.

[109] X. Wang, Y. Wang, and W. Li, "U-Net conditional GANs for photo-realistic and identity-preserving facial expression synthesis," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 15, no. 3s, pp. 1–23, 2019.

[110] Y. Shen, B. Zhou, P. Luo, and X. Tang, "FaceFeat-GAN: A two-stage approach for identity-preserving face synthesis," arXiv preprint arXiv:1812.01288, 2018.

[111] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas, "Expression flow for 3D-aware face component transfer," in ACM SIGGRAPH 2011 papers. Rutgers University, 2011, pp. 1–10.

[112] U. Mohammed, S. J. Prince, and J. Kautz, "Visiolization: generating novel facial images," ACM Transactions on Graphics (TOG), vol. 28, no. 3, pp. 1–8, 2009.

[113] R. Yeh, Z. Liu, D. B. Goldman, and A. Agarwala, "Semantic facial expression editing using autoencoded flow," arXiv preprint arXiv:1611.09961, 2016.

[114] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in International Conference on Machine Learning, 2014, pp. 1431–1439.

[115] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen, "Discovering hidden factors of variation in deep networks," arXiv preprint arXiv:1412.6583, 2014.

[116] Y. Zhou and B. E. Shi, "Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder," in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2017, pp. 370–376.

[117] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5541–5550.

[118] M. Li, W. Zuo, and D. Zhang, "Deep identity-aware transfer of facial attributes," arXiv preprint arXiv:1610.05586, 2016.

[119] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional GANs for image editing," arXiv preprint arXiv:1611.06355, 2016.

[120] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 818–833.

[121] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 704–720.

[122] O. Litany, A. Bronstein, M. Bronstein, and A. Makadia, "Deformable shape completion with graph convolutional autoencoders," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 1886–1895.

[123] J. Geng, T. Shao, Y. Zheng, Y. Weng, and K. Zhou, "Warp-guided GANs for single-photo facial animation," ACM Transactions on Graphics (TOG), vol. 37, no. 6, pp. 1–12, 2018.

[124] H.-A. Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen, "Bringing portraits to life," ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017), vol. 36, no. 6, p. 196, 2017.

[125] Y. Nakahira and K. Kawamoto, "DCVGAN: Depth conditional video generation," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 749–753.

[126] X. Sun, H. Xu, and K. Saenko, "TwoStreamVAN: Improving motion modeling in video generation," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 2744–2753.

[127] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, "Pose guided human video generation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 201–216.

[128] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-GAN: Unsupervised video retargeting," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 119–135.

[129] L. Yu, D. Davaasuren, S. Rao, and V. Kumar, "StarGAN-EgVA: Emotion guided continuous affect synthesis," in Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis, 2020, pp. 53–61.

[130] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in Advances in Neural Information Processing Systems, 2016, pp. 613–621.

[131] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2830–2839.

[132] J. Lee, D. Ramanan, and R. Girdhar, "Metapix: Few-shot video retargeting," arXiv preprint arXiv:1910.04742, 2019.

[133] Y. Chen, Y. Pan, T. Yao, X. Tian, and T. Mei, "Mocycle-GAN: Unpaired video-to-video translation," in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 647–655.

[134] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation," arXiv preprint arXiv:1904.11685, 2019.

[135] 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, 2016, pp. 2234–2242.

[136] A. Borji, "Pros and cons of GAN evaluation measures," Computer Vision and Image Understanding, vol. 179, pp. 41–65, 2019.

[137] D. J. Im, C. D. Kim, H. Jiang, and R. Memisevic, "Generating images with recurrent adversarial networks," arXiv preprint arXiv:1602.05110, 2016.

[138] S. Latif, R. Rana, and J. Qadir, "Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness," arXiv preprint arXiv:1811.11402, 2018.

[139] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," arXiv preprint arXiv:1811.01174, 2018.

[140] P. Sheng, Z. Yang, H. Hu, T. Tan, and Y. Qian, "Data augmentation using conditional generative adversarial networks for robust speech recognition," in 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2018, pp. 121–125.

[141] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," arXiv preprint arXiv:1806.06626, 2018.

[142] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using GANs for speech emotion recognition," Proc. Interspeech 2019, pp. 171–175, 2019.

[143] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," arXiv preprint arXiv:1704.00849, 2017.

[144] A. Mathur, A. Isopoussu, F. Kawsar, N. Berthouze, and N. D. Lane, "Mic2Mic: using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems," in Proceedings of the 18th International Conference on Information Processing in Sensor Networks, 2019, pp. 169–180.

[145] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 266–273.

[146] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018, pp. 2100–2104.

[147] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6820–6824.

[148] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 632–639.

[149] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "WaveCycleGAN2: Time-domain neural post-filter for speech waveform generation," arXiv preprint arXiv:1904.02892, 2019.

[150] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," arXiv preprint arXiv:1811.11307, 2018.

[151] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," arXiv preprint arXiv:1808.05092, 2018.

[152] H. Kameoka, K. Tanaka, T. Kaneko, and N. Hojo, "Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion," arXiv preprint arXiv:1811.01609, 2018.

[153] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: sequence-to-sequence voice conversion with attention and context preservation mechanisms," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6805–6809.

[154] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). IEEE, 2013, pp. 1–4.

[155] T. Tomoki, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, J. Yamagishi et al., "The voice conversion challenge 2016," Annual Conference of the International Speech Communication Association, 2016.

[156] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, p. 335, 2008.

[157] B. W. Schuller and A. M. Batliner, "Emotion, affect and personality in speech and language processing," Signal Processing, 1988.

[158] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," arXiv preprint arXiv:1804.04262, 2018.

[159] K. Ito et al., "The lj speech dataset," 2017.

[160] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in Fifth ISCA workshop on speech synthesis, 2004.

[161] 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," PloS one, vol. 13, no. 5, 2018.

[162] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), 2000.

[163] W. Cai, A. Doshi, and R. Valle, "Attacking speaker recognition with deep generative models," arXiv preprint arXiv:1801.02384, 2018.

[164] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," arXiv preprint arXiv:1803.00860, 2018.

[165] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "BAGAN: Data augmentation with balancing gan," arXiv preprint arXiv:1803.09655, 2018.

[166] X. Jia, J. Tai, H. Zhou, Y. Li, W. Zhang, H. Du, and Q. Huang, "ET-GAN: Cross-language emotion transfer based on cycle-consistent generative adversarial networks," arXiv preprint arXiv:1905.11173, 2019.

[167] S. Vasquez and M. Lewis, "Melnet: A generative model for audio in the frequency domain," arXiv preprint arXiv:1906.01083, 2019.

[168] L. Juvela, B. Bollepalli, X. Wang, H. Kameoka, M. Airaksinen, J. Yamagishi, and P. Alku, "Speech waveform synthesis from MFCC sequences with generative adversarial networks," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5679–5683.

[169] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), vol. 2. IEEE, 2001, pp. 749–752.

[170] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Rec. ITU-T P. 862, 2001.

[171] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial neural audio synthesis," arXiv preprint arXiv:1902.08710, 2019.

[172] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 670–686.

[173] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-End speech-driven facial animation with temporal GANs," arXiv preprint arXiv:1805.09313, 2018.

[174] A. Ephrat and S. Peleg, "Vid2speech: speech reconstruction from silent video," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5095–

5099.

[175] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 455–462.

[176] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: learning lip sync from audio," ACM Transactions on Graphics (TOG), vol. 36, no. 4, pp. 1–13, 2017.

[177] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7181–7189.

[178] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 595–602.

[179] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah, "Generative adversarial networks conditioned by brain signals," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3410–3418.

[180] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421–2424, 2006.

[181] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017.

[182] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," International Journal of Computer Vision, pp. 1–16, 2019.

[183] W.-C. Chen, C.-W. Chen, and M.-C. Hu, "SyncGAN: Synchronize the latent spaces of cross-modal generative adversarial networks," in 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018, pp. 1–6.

[184] W. Hao, Z. Zhang, and H. Guan, "CMCGAN: A uniform framework for cross-modal visual-audio mutual generation," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[185] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," in Advances in Neural Information Processing Systems, 2017, pp. 2018–2028.

[186] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" arXiv preprint arXiv:1801.04406, 2018.

[187] A. Ghosh, V. Kulharia, V. P. Namboodiri, P. H. Torr, and P. K. Dokania, "Multi-Agent diverse generative adversarial networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8513–8521.

[188] M. J. Kusner and J. M. Hernández-Lobato, "GANs for sequences of discrete elements with the Gumbel-Softmax distribution," arXiv preprint arXiv:1611.04051, 2016.

[189] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," arXiv preprint arXiv:1611.01144, 2016.

[190] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," arXiv preprint arXiv:1611.00712, 2016.

[191] Y. Mroueh and T. Sercu, "Fisher GAN," in Advances in Neural Information Processing Systems, 2017, pp. 2513–2523.

[192] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng, "Sobolev GAN," arXiv preprint arXiv:1711.04894, 2017.

[193] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, 2017, pp. 4681–4690.

[194] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[195] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1415–1424.

[196] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2439–2448.

[197] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 743–10 752.

N. HAJAROLASVADI Noushin Hajarolasvadi is currently a Ph.D. candidate with the Department of Electrical and Electronic Engineering, Eastern Mediterranean University. Her research works involve human behavior analysis with focus on facial expressions and speech emotion recognition. She has been involved in several scientific short term missions and has served as a reviewer at various journals.

M. A. RAMíREZ Miguel Arjona Ramírez (M'76–SM'2000) became a Member (M) of IEEE in 1978, and a Senior Member (SM) in 2000. He received the B.S. degree in electronics engineering from Instituto Tecnológico de Aeronáutica, Brazil, in 1980, and the M.S. and the Ph.D. degrees in electrical engineering and the Habilitation degree in Signal Processing from University of São Paulo, Brazil, in 1992, 1997 and 2006, respectively, and the electronic design eng. degree from Philips International Institute, The Netherlands, in 1981.

He was Engineering Development Group Leader for Interactive Voice Response Systems (IVRs) for Itautec Informática, Brazil, where he served from 1982 to 1990. In 2008 he carried research in time-frequency speech analysis and coding in a research visit to the Royal Institute of Technology in Sweden. He is currently Associate Professor at Escola Politécnica, University of São Paulo, where he is a member of the Signal Processing Laboratory. He has authored or coauthored 4 book chapters and over 60 journal and conference papers in these areas. His research focuses on the application of novel signal processing and machine learning algorithms to signal compression and prediction, speech analysis, coding and recognition, speaker identification and audio analysis and coding. He is a member of the Brazilian Telecommunications Society (SBrT).

W. BECCARO Wesley Beccaro received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from University of São Paulo in 2008, 2012, and 2017, respectively. His research interests include digital signal processing, instrumentation, embedded systems, and fuel qualification.

H. DEMIREL Hasan Demirel received the B.Sc. degree from the Electrical and Electronic Engineering Department, Eastern Mediterranean University, in 1992. He received the M.Sc. degree from King's College London in 1993 and the Ph.D. degree from Imperial College London in 1998. He joined the Electrical and Electronic Engineering Department, Eastern Mediterranean University, in 2000, as an Assistant Professor, where he was appointed as an Associate Professor in 2009 and a Professor in 2014. He has served as the Vice Chairman of the Electrical and Electronic Engineering Department, Eastern Mediterranean University from 2003 to 2005 and from 2007 to 2014, where he has been the elected Chairman since 2014. He had been the Deputy Director of the Advanced Technologies Research and Development Institute. His main research is in the field of image processing and he is currently involved in research projects related to biomedical image processing, image/video resolution enhancement, face recognition/tracking, facial expression analysis, and low-bit rate video coding. Furthermore, he had served as a member for the Executive Council of the EMU Technopark.