FoleyGAN: Visually Guided Generative Adversarial Network-Based Synchronous Sound Generation in Silent Videos

Sanchita Ghose, Graduate Student Member, IEEE, and John J. Prevost, Senior Member, IEEE

Abstract—Deep learning based visual-to-sound generation systems have been developed that identify and create audio features from video signals. However, these techniques often fail to consider the time-synchronicity of the visual and audio features. In this paper we introduce a novel method for guiding a class-conditioned GAN to synthesize representative audio with temporally-extracted visual information. We accomplish this visual-to-sound generation task by adapting the synchronicity traits between the audio-visual modalities. Our proposed FoleyGAN model is capable of conditioning action sequences of visual events leading to the generation of visually aligned realistic soundtracks. We expanded our previously proposed Automatic Foley data set. We evaluated FoleyGAN’s synthesized sound output through human surveys that show noteworthy (on average 81 %) audio-visual synchronicity performance. Our approach outperforms other baseline models and audio-visual data sets in statistical and ablation experiments achieving improved IS, FID and NDB scores. In ablation analysis we showed the significance of our visual and temporal feature extraction method as well as augmented performance of our generation network. Overall, our FoleyGAN model showed sound retrieval accuracy of 76.08% surpassing existing visual-to-audio synthesis deep neural networks.

Index Terms—Deep neural network, foley generation, generative adversarial network, multi-modal learning, sound synthesis, video class prediction, visual guidance, visual-to-sound.

I. INTRODUCTION

Foley recording, a component of the film production process, provides added realism and clarity to movie scenes by overlaying artificial sounds that emphasizes important events and actions.

Today’s film production teams are dependent on Foley tracks for movie scenes where the background sound is either not present or where the original recording does not come through well. In these situations the Foley artist looks for available recorded Foley tracks, or records the required sounds in special studios. The Foley artist’s skill is being able to know how to accurately generate and record the required sound effect. Since the latter option is often costly, filmmakers often prefer to acquire pre-recorded tracks from online or other sources at a lower cost. Although this seems like an easy solution, they often encounter a lack of synchronicity between the video and the overlaying sound. One solution is to use deep learning algorithms that can learn the temporal-correspondence between audio and video signals, then generate the sound accordingly for the given video clip. In our previous work [1], we addressed traditional Foley generation problems and proposed two unique deep learning models for automatic Foley generation. While generating sounds for the Foley process, our goal is to enhance the visual effect of the film by creating a sound overlay that audiences will immediately associate with the related visual. A major goal therefore is sensory augmentation, meaning that the correct synthesized sound allows the audience to be more engaged than normal. This was the focus in our prior research. This research seeks to address the cross modal problem of the time synchronization between the generated sound and the video file (Fig. 1).

In this paper, we propose a visually guided class conditioned deep adversarial Foley generation network called “FoleyGAN,” which we present as an advancement in automatic Foley-sound synthesis from silent video clips. Since sound plays a crucial information role in the perception of the inherent action in most of the visual scenarios of the real world, and auditory guidance can assist a person or a device in analyzing the surrounding events more effectively, our proposed network also has the potential to serve as an IoT (Internet of Things) system that is able to learn the correspondence between visual and audio modalities along with synthesizing actionable synchronous sound tracks from visual signals.

Generative Adversarial Networks (GANs) [2] have started to become widely used by researchers as a deep generating model, particularly for high quality image generation applications (e.g. [3]–[7]). Notable advances are found in utilizing GANs for audio and music generation [8]–[13] as well, though adversarial audio generation still remains a highly challenging task because of intrinsic differences between sound waveforms and image signals. Sound waves generally show higher periodicities than image signals, which leads to the use of more sophisticated filters with large receptive fields. In addition, generated audios are more likely to be affected by annoying “checkerboard” artifacts, which can be avoided in a generated image using GAN. Recent research uses spectral representations of audio files for adversarial generation. However, none of these approaches have...
Fig. 1. Problem description: Visually synchronous sound synthesis capturing temporal action information. Our proposed model locates the temporal action changes in subsequent frames of a video and generates the sound accordingly.

considered time-action synchronicity traits as a visual guidance to condition the sound generator of GAN, along with sound class information, which is the key novelty of our “FoleyGAN” network. In addition, we efficiently incorporate scaled-up (512 × 512) BigGAN [14] architecture as our base generative network, which enables the synthesizing of high resolution spectrograms that are inverted to sound tracks via ISTFT [15]. In addition to using latent space and class information as inputs, we condition the BigGAN generator with visual guidance. Furthermore, we also expand on our previously proposed “AFD” dataset [1]. The discriminator network is pretrained with the spectrogram soundfile images of the updated dataset to differentiate between generated and actual samples.

Fig. 2 shows the proposed FoleyGAN network, which consists of two major neural network blocks: a video action recognition network (the upper block, consisting of CNN and Temporal Relational Network (TRN) architectures [16]) followed up with a visually guided class conditioned GAN network (the lower block) for sound generation. The first block provides the action category prediction of the respective input video as well as prediction weights of the action occurrences over the video time duration from which we are generating action spectrograms. These two outputs are forwarded to our next sound generative network using the GAN principle. Finally, the generated spectrogram is inverted via ISTFT to obtain the visually synced sound track for the respective video clip.

Previously in AutoFoley [1], we proposed two separate deep neural networks (e.g. Frame Sequence and Frame Relation Networks) for predicting action in video frames. Since the overall performance of both models are quite similar, we can use either of these models for the later expansion of this research. However, in this work, we want to focus on reducing computational complexity, as we are integrating a scaled up GAN architecture (e.g. BigGAN) for high resolution spectrogram generation.

Additionally, we aim to advance the earlier proposed automatic sound synthesis system with time synchronicity features. Therefore, we intentionally select the Frame Relation Network, which is not only capable of capturing the temporal relations between two consecutive video frames (leading to the prediction of the action happening in the scene), but also uses limited video frames
as inputs that are fed into a simpler multilayer perceptron (MLP) structure significantly reducing the computational loads. In addition, we are able to condition our generator network with relational reasoning information between two sequential frames with the help of temporal relation statistics.

The significant contributions made by this paper are:

- We take the initial step toward automatic Foley generation in a silent video clip using a visually guided class conditioned generative adversarial network, taking into consideration the time-action synchronicity requirement in the highly diverse movie-sound-effects domain.
- We introduce a concept of conditioning the generated samples of a video frame sequence that can be deployed for automatic Foley synthesis as well as other multi-modal applications.
- We expand our previously proposed “Automatic Foley Dataset (AFD)” for future research and training.
- We present an image generating BigGAN architecture trained on AFD for realistic and synchronous three second duration sound synthesis for the multimedia applications field.
- For the performance analysis of our generated sounds, we perform qualitative, numerical, ablation experiments comparing with baseline models and conduct a human survey on our generated sound quality as well as the video/sound alignment in the respective visual events.

This paper is structured as follows. In Sections II and III, we present related works and a brief review of GAN. In Section IV, we describe our detailed methodology and present the complete algorithm used in this research. In Sections V and VI, we provide the explanation of our extended AutoFoley dataset, training details with specifications on hyper-parameter tuning, and model evaluation results analysis through numerical, qualitative and ablation experiments to assess the overall performance. Finally, Section VII concludes with the summarization of substantial points and future directions of this work.

II. RELATED WORK

A. Foley Generation

Automatic sound effect creation from 3D models has been approached in [17] through dynamic simulation and user interaction. In our recent work [1], deep learning is deployed in the application of automatic Foley generation, where we propose a unique deep learning solution to predict sound in silent video clips and then synthesize Foley from the predicted features. In this paper, we utilize conditional generative adversarial training on our predicted video categories to generate Foley of that respected class.

B. Audio-Visual Correlation

Everyday we observe audio-visual events happening around us where sound plays a vital role. The human ability to quickly correlate between these two modalities simultaneously allows us to react appropriately to real-time events. Taking inspiration from this fact, [1], [18]–[22] utilizes these audio-video correspondence properties for training their neural networks with unlabeled video data. The audio-visual relationship is employed to develop deep neural networks in various fields of applications, e.g. for the material recognition task [23], sound source localization task in video [18], [20], [24]–[29], audio source separation tasks [30], audio event identification tasks for video analysis [31], and video action recognition to automatic foley generation tasks [1]. Likewise, advanced research approaches proposed in [26], [28], [29] have assisted in localizing a sound source against visual data in 3D space by utilizing our ability to observe audio-visual events. In [21], an automatic video sound recognition and visualization framework is proposed, where nonverbal sounds in a video are automatically converted into animated sound words and are placed close to the sound source of that video for visualization. In addition, an attention mechanism learning network for the sound source proposed in [32] and semantic guided modules (SGMs) performed in [33] for action recognition to extract spatial-temporal features from videos show promising applicability in audio-visual association properties. We are motivated by this research on audio-visual relevance, and aim for improved mapping of audio-video features by expanding our AutoFoley deep neural network with an efficient generative adversarial model.

C. Sound Synthesis From Videos

Understanding the capability of the human brain to synchronize audio and visual modalities simultaneously, [1], [13], [23], [34]–[39] propose different neural networks for sound synthesis from visual inputs. Research in [40] use audio generation for the full viewing sphere, when a 360° video and corresponding mono audio are given, whereas in their later work [36], they leveraged object configurations in videos for transforming mono channel to binaural audio. Similar video-based audio spatialization research is shown in [41]. Prior work in [34] shows natural sound generation from videos captured in the wild, whereas the AutoFoley framework [1] synthesizes Foley tracks in silent video frames. Another approach for sound generation from visual inputs is presented in [13] using conditional generative adversarial networks. Recent work in [38] proposed a spectrogram based sound generation model named REGNET, where authors introduced an audio forwarding regularizer to pass missing information while training. REGNET research focuses on eliminating irrelevant sound component to prevent incorrect audio-visual mapping whereas in our proposed visual-to-audio generator we focus on learning temporal action relation for audio-visual mapping. To reduce the computational cost while synthesizing realistic rain sound, a different sound generation approach have been adopted in [42] where authors proposed a physically-based statistical simulation method to capture dynamic variations of rain sound. Authors in [43] presented a real time sound synthesis system based on extracting foreground sounds from background textures using double layer Markov Models capable of capturing different properties of foreground and background units. Their proposed hierarchical grid
scheme generates Head-Related Transfer Function filters to localize sound clues represented as area sources. Authors in [44] considered positional variations of multiple sound sources from listeners perspective in complex scenes. They introduced event loudness density (ELD) that is able to relate the rapidity of received events to their loudness in order to compose and encode acoustic texture. In our work, we develop a deep learning model comprising of a visual action recognition and adversarial audio synthesis network to generate realistic Foley tracks for silent movie clips.

D. Audio Generation With GAN

GANs have extensive potentials in the computer vision and image generation field (e.g. [3], [4], [5], [7], [45]), which encourages researchers to deploy principles of GAN in the audio generation domain as well. Being inspired by image inpainting, authors in [37] have recently performed audio inpainting as a form of spectromorphs with GAN. Earlier works in [8], [9], [10], [11], [12], [13] show a clear direction of using generative adversarial training with audio signals. However [11], [10] portray the challenges in training GAN with audio waveforms compared to image matrices. Therefore, spectral representations of audio are preferred while training adversarial audio generation. The phase-gradient heap integration (PGHI) [46] algorithm proposed in TiFGAN’s paper [47], represents an improved reconstruction technique of the audio from the spectrogram with minimal loss. Authors in [47] trained GANs on short-time Fourier features to mitigate the problems of generating audio in the short-time Fourier domain. GAN has been used for efficient and high-fidelity speech synthesis task in [48] where the proposed HiFi-GAN performs with the melspectrogram and can generate faster samples with comparable quality to an autoregressive counterpart. Likewise, MelGAN architecture in [49] showed the potential of a non-autoregressive feed-forward convolutional model for faster audio generation using mel-spectrogram. MelGAN architecture is applied on text-to-speech generation using VQGAN with a new perceptual loss for spectrogram generation. They used BN-Inception model for feature extraction whereas our proposed system utilized TRN models to extract visual temporal action features to condition BigGAN. In our previous work [51], we first proposed an IoT System of Systems framework of audio generation for visual inputs exploiting BigGAN [14]. Recently, authors in [52] utilized BigGAN architecture for adversarial audio generation in a guided manner. Our proposed FoleyGAN architecture is a novel approach to apply BigGAN in the movie sound production domain where we are synthesizing the audio for silent movie clips using visual and temporal guidance.

III. GENERATIVE ADVERSARIAL NETWORK (GAN) BASICS

Generative Adversarial Networks (GANs) proposed in [2] include a generator network $G$ and a discriminator network $D$. The two networks play in an adversarial manner taking part in a min-max game throughout the training process. The training objective of the G network is to map random vector $z \in Z$ into generated samples by minimizing the following value function (1), whereas the D network, which judges between real and generated examples, is trained to maximize the value function. Here, $z$ belongs to random noise distribution $p_z$ and $p_{data}$ denotes the target data distribution.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))]$$

In a conditional GAN approach (2), conditional information (e.g. labels of images) is passed to the generator and discriminator networks where $y$ represents the condition variable.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z|y))]$$

In this work, we use hinge loss for updating the generator and the discriminator in our visually guided sound generation network. In a conditional GAN network, hinge loss for the discriminator and generator is calculated as,

$$L_D = L_{Dreal} + L_{Dfake}$$

$$\mathbb{E}_{(x,y) \sim p_{data}} [\max(0, 1 - D(x, y))] + \mathbb{E}_{z \sim p_z(z), y \sim p_{data}} [\max(0, 1 + D(G(z, y), y)))]$$

$$L_G = -\mathbb{E}_{z \sim p_z(z), y \sim p_{data}} [D(G(z, y))]$$

IV. PROPOSED RESEARCH METHOD

We separate our proposed architecture into two networks: A) a video action recognition network and B) a sound generation network. We explain these network details in the following subsections. The graphical representation of the complete FoleyGAN architecture is presented in Fig. 3.

A. Video Action Recognition Network

We pick the frame relation model from [1] for class prediction because of its superior performance in learning temporal dependencies from visual frames, with less computational complexity compared with other prediction models. The video action recognition network provides the prediction of the overall action category (along with the frame-by-frame identical action occurrence probabilities) of the entire video clip exploiting the multiscale and 2-frame temporal relational networks [16] principle respectively. The detailed methods are explained in following paragraphs.

1) Video Action Class Prediction: We use a fused network comprised of CNN and a multiscale temporal relation network (MTRN) proposed in [16] to identify the action occurring throughout the video clip. We compute the temporal relation composite functions $R_Q$ using the following equation, where $Q = [2, 3, \ldots, 8]$ represents the number of video frames under
Here, $F_j$, $F_k$, and $F_l$ represent the activation output obtained from the pretrained ResNet-50 [53] CNN architecture at the $j^{th}$, $k^{th}$, and $l^{th}$ frame of the video. We train the ResNet-50 model with $n$ number of soundless video frames $[I_1, I_2, \ldots, I_n]$ from each video ($V$) of our trained dataset. In this equation, $h_{\phi}$ is a single layer and $g_{\theta}$ is a double layer multilayer perceptron (MLP) associated with 256 units per layer. These functions compile features of video frames at different temporal orders and are unique for each $R(V)$. In this way, we calculate the composite temporal function over time up to 8 frames as $R_8(V)$ since, up to this frame number, we achieve optimum results through ablation studies on TRN networks in our earlier work [1]. Finally, we sum all the temporal relation functions (2) to compute the action category $\mathbf{C}(V)$ happening in the entire video clip.

$$C(V) = R_2(V) + R_3(V) + \cdots + R_8(V)$$ (5)

2) Video Action Spectrogram Generation: Our intuition is to obtain the relational reasoning information between two sequential frames over the complete time duration of the video with the help of a temporal relation equation of $R_2$. We plot these values over time to get a time-series graph representing the probabilities $P_{\text{act}}$ of similar action occurrence of two subsequent frames over the whole time period of the video. Next, we convert this time series plot into a spectral representation by computing STFT and reshape it into the required dimension ($512 \times 512 \times 3$). Therefore, we get a 3D matrix, $S_{\text{act}}$ that we name a video action spectrogram since it contains the frame-by-frame similar action occurrence probabilities of each video (Fig. 4). Next, we condition our sound generation network with this visual guidance to create the temporal synchronicity between the audio and visual inputs.
B. Sound Generation Network

1) Preprocessing of Sound Data for Training: In our objective to train our generative model (image generating GAN) with sound data, we have to represent our sound files as three dimensional matrices without losing the magnitude and phase information contained within the individual tracks. To do so, we first extract audio from video recordings and clip them into 3 s durations. We then convert audio files into mono-wave files and compute their spectrograms by calculating STFT with the help of TensorFlow’s built-in functions. We use the Hanning window and sample frequency of 44kHz and select a stride of 256 and frame size of 1024, allowing windows to overlap 75% with 513 frequency bins. In order to obtain a three dimensional image-like matrix, we use padding in the time axis. Finally, our complex spectrogram of each sound file becomes a matrix (512,512,3), containing both the magnitude and phase information of the original audio in the 1st and 2nd channel respectively. For the 3rd channel, we again apply zero padding, which we extract later through depadding during the reconstruction process. Finally, we prepare the sound spectrogram features, applying a mel-filter bank to convert the frequency scale into the mel-scale. Since our generator network applies a tanh nonlinearity function, we scale the log magnitudes and phase angles within the -1 to 1 range to comply with the generator model.

2) Generation of Visually Guided Sound: Similar to the SpecGAN model proposed in [10], our deep sound generation network is a frequency-domain sound synthesis GAN architecture. The proposed generation network is trained with spectrogram inputs by performing short-time Fourier Transform (STFT) [54] on audio samples. The generated output spectrograms are then inverted using the (ISTFT) method [15]. The objective of feeding spectrogram inputs to the generation network is to leverage the efficiency of GAN in high resolution image generation tasks. In this proposed model, we adopt BigGAN [14] for adversarial sound synthesis by generating high fidelity spectrogram images of multiple categories through large scale GAN training. The generator and discriminator network follows a BigGAN (512 × 512) image generation architecture capable of generating high resolution spectrogram images of multiple sound classes.

In brief, BigGAN is a high resolution and high-fidelity class-conditional image generating GAN model that significantly improves the inception score by using higher batch sizes with increased width in each layer. Being a class conditional GAN, it takes image class information and a point from latent space as an input. Rather than using the pretrained weights of BigGAN trained on natural images from the ImageNet dataset, we train the model with our generated spectrograms to follow our goal for adversarial sound synthesis. As previously mentioned, the class output C resulted from the prediction network, and the action spectrogram S_{act}, are both fed into the generator. Being conditioned by the video action information, the generative network produces a spectrogram of the predicted class, taking random noise z as an input. Next, the generated image S_{gen} is passed to the discriminator block pretrained with the original spectrogram image S_{real} of that predicted class.

The discriminator network distinguishes between the real spectrogram S_{real} and the synthesized spectrogram S_{gen}. Using BigGAN, we adopt an orthogonal regularization technique and a truncation trick to boost performance and improve the generated spectrogram quality. Using a “truncation trick,” our generator uses less random numbers, which leads to the output of more realistic images. Finally, L_D and L_G losses are calculated (as 5) and fed back to the generator and discriminator blocks to update their weights at the end of each training epoch.

As the training proceeds, the generator moves closer to synthesizing a spectrogram that misguides the discriminator that is identifying the differences between the original and generated images. At the end of the training, the generator learns the pattern and features of the original spectrograms and generates representative spectrogram images of 512 × 512 resolution classified as real by the discriminator. For the complete architecture and parameter details of BigGAN’s generator and discriminator blocks, we direct readers to the appendix section of the original paper [14].

V. EXPERIMENTAL DETAILS

A. Dataset

In the context of generating artificial foley tracks from silent video, in our previous work we proposed an Automatic Foley Dataset (AFD) [1] that is carefully prepared to avoid external noise focusing on popular foley categories. Since GAN training requires a large set of training samples for improved learning, we expand our dataset with more diverse video samples to be used in FoleyGAN training. In Table I, we show the data percentages of individual classes of our updated AFD dataset. The total number of video samples is 27,800 (of 3 s duration each). In addition, as an ablation analysis, we compare the generated audio sample performance (Table III) by training the proposed FoleyGAN architecture with a subset of AudioSet [55] and YouTube8M [56] datasets, as they closely comply with our data requirements for

---

Algorithm 1: Visually Guided Adversarial Foley Generation

| Input: | Silent video frames (I_1, I_2, ... I_N), training audio tracks (A_1, A_2, ... A_N) and random noise z. |
| Output: | Generated audio tracks (A_gen). |

1: V_t ← CNN(I_N)
2: C ← MTRN(V_t)
3: Prob_seq ← 2TRN(V_t)
4: S_{act} ← Spectrogram(Prob_seq)
5: S_{real} ← Spectrogram(A_N)
6: for number of training iterations do
7: S_{gen} ← BigGAN_G(z, C, S_{act})
8: R ← BigGAN_D(S_{real}, S_{gen})
9: Calculate L_G and L_D
10: Update BigGAN_G and BigGAN_D
11: end for
12: A_gen ← ISTFT(S_{gen})

---

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
that this task. We prepare the subsets by collecting videos of 12 similar categories contained in the AFD. In all cases, our training set comprises of 80% and testing set comprises the remaining 20% of the whole dataset.

### B. Experimental Protocols

We trained the event class prediction MTRN (Multi-Scale Temporal Relation Network), the consecutive action prediction 2TRN network, and the sound generating GAN network separately on training datasets. We collected image features from the output of the conv5 layer of the ResNet-50 network. The TRN models have two layers of MLP (256 units in each) for \( g_{\theta} \) and a single layer MLP (12 units) for \( h_{\phi} \). The training of 100 epochs was completed in less than 24 hours on a NVIDIA Tesla V100 GPU. We used a minibatch gradient descent with the Adam optimizer [25]. The minibatch size was 128 and the learning rate was 0.001.

To implement our audio generation network, we adopted the 512 × 512 BigGAN [14] architecture (which is a Self-Attention GAN [57] based model) trained on our AutoFoley spectral data. In most cases, we followed similar hyper-parameters and optimization techniques for the discriminator and generator while training. The entire implementation was done using TensorFlow. Similar to BigGAN, we applied an orthogonal Initialization [58] strategy (e.g. introducing a random orthogonal matrix weight in each layer maintaining their orthogonal property) on both the generator and the discriminator. The generator model used the skip-z technique to directly link the input latent vector \( z \) to specific layers deep in the network where the full dimensionality of \( z \) is set to 160 for a 512 × 512 spectrogram image generation. We set the learning rate to \( 2 \times 10^{-4} \) and \( 5 \times 10^{-5} \) for the discriminator and generator respectively. We obeyed the truncation trick [14] by resampling the \( z \) values to arbitrate between image quality and variety. The overall model was trained by calculating the hinge loss. We used the Adam optimizer [25] for optimization. BigGAN performance greatly depends on increasing the batch size – specifically, BigGAN requires high batch size training to provide better gradient information while updating the weights through training epochs. However, training with larger batches requires GPUs of higher memories. To handle the memory constraints, we implemented a gradient accumulation technique during our training session. We trained our sound generation network on a single NVIDIA Tesla V100 GPU of 32GB VRAM. Our intuition was to train with a total batch size of 2048. To avail this large batch size without facing a “OOM” error (e.g. out of memory error), we used a mini batch size of 128 for 16 gradient accumulations. Each training session took 8 days to complete for 500 epochs with 12k iterations. We then added a post-processing filter of 512 length to the generator output for decreasing the noisy artifacts of the generated spectrogram samples.

### VI. Model Evaluation

In this section, we first describe different numerical evaluation matrices adopted to assess the performance of our proposed method in a quantitative manner, and explain the calculated results comparing with state-of-the-art models (subsection A-E). Next, in subsection F and G we show an extensive ablation analysis and a phase coherence study respectively. Later in subsection H, we present human survey results to evaluate the generated sound quality in a comprehensive way in accordance with the video clips.

#### A. Sound Retrieval Accuracy

We adopted the similar sound retrieval experiment system as in AutoFoley [1]. This time we prepared a sound classifier by training a ResNet-50 [53] CNN model with spectrogram images of our updated AFD training data. We calculated the classifier’s performance (i.e., accuracy with real data) by testing it with AFD test spectrogram samples. Next, we measured the prediction accuracy of our generated spectrogram samples with proposed and other baseline models. The average accuracy was measured over all event classes (shown in Table II).

#### B. Inception Score (IS)

To evaluate the semantic diversity of generated samples, we calculated the inception score (IS) proposed in [59] using the following equation:

\[
\exp(\mathbb{E}_x D_{KL}(P(y|x)||P(y)))
\]

Here, \( P(y|x) \) represents the conditional class distribution for image sample \( x \) predicted by the Inception Network [60], and \( P(y) \) gives the marginal class distribution. The equation computes the IS score by calculating the Kullback-Leibler (KL) Divergence between these two distributions. The Inception features are extracted from the Inception Network [60] trained on the ImageNet dataset. A high IS value is preferred when evaluating generation quality. Since the Inception Score evaluation matches with human judgements at a high level, we wanted to evaluate our generated spectrograms on this basis. Therefore, we used our pretrained sound retrieval CNN classifier features mentioned in the previous subsection to compute the score (shown in Tables II and III).

#### C. Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) measures the Fréchet Distance (FD) between two multivariate Gaussian distributions.

| Class Group | Data (%) |
|-------------|----------|
| A (car racing, clock ticking, fire rainfall, thundering, typing and waterfall videos) | 10.79 |
| B (chopping, footsteps, gunshots and horse running videos) | 5.4 |
| C (breaking videos) | 2.88 |

### TABLE I

**AUTOFOLEY DATASET STATISTICS**

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE II

PERFORMANCE COMPARISON OF GENERATED SAMPLES FROM SOUND GENERATIVE BASELINE MODELS WITH AFD DATASET

| Samples                  | IS   | FID  | NDB  | Average Accuracy (%) |
|--------------------------|------|------|------|-----------------------|
| Real Data                | 11.42| 11   | 3.23 | 78.32                 |
| FoleyGAN                 | 10.97| 67   | 18.47| 76.08                 |
| FoleyGAN without visual  | 9.22 | 181  | 26.53| 64.61                 |
| guidance                 |      |      |      |                       |
| AutoFoley (Frame Sequence | 10.40| 127  | 20.94| 65.79                 |
| Network)                 |      |      |      |                       |
| AutoFoley (Frame Relation | 10.72| 119  | 20.03| 63.40                 |
| Network)                 |      |      |      |                       |
| GANSYNTH (IF-Mel + H)    | 10.87| 115  | 22.14| 73.12                 |
| Taming VGSG              | 10.47| 121  | 20.63| 68.16                 |
| SpecGAN                  | 8.62 | 271  | 30.07| 61.75                 |
| WaveGAN                  | 7.36 | 322  | 34.91| 59.93                 |

TABLE III

PERFORMANCE COMPARISON OF GENERATED SAMPLES FROM FOLEYGAN WITH AUDIO-VISUAL DATASETS

| Dataset                  | IS   | FID  | NDB  | Average Accuracy (%) |
|--------------------------|------|------|------|-----------------------|
| Real Data                | 11.42| 11   | 3.23 | 78.32                 |
| AFD                      | 10.97| 67   | 18.47| 76.08                 |
| YouTube8M Subset         | 10.04| 114  | 20.03| 70.16                 |
| AudioSet Subset          | 9.72 | 102  | 21.16| 68.71                 |

TABLE IV

GENERATED SAMPLE QUALITY COMPARISON WITH AFD DATASET FOR DIFFERENT SOUND FEATURES USING FOLEYGAN

| Class        | NDB (k = 50) |
|--------------|--------------|
|              | STFT  | CQT   | MFCC | LMS   |
| Break        | 31.6  | 30.1  | 28.4 | 29.3  | 23.5 |
| Car          | 22.8  | 27.5  | 30.2 | 21.8  | 21.6 |
| Clock        | 15.3  | 20.4  | 19.1 | 14.0  | 11.2 |
| Chopping     | 21.6  | 18.5  | 22.1 | 17.2  | 15.7 |
| Fire         | 15.1  | 17.4  | 15.3 | 13.6  | 12.0 |
| Footstep     | 18.9  | 21.0  | 19.1 | 18.2  | 13.3 |
| Gunshot      | 26.7  | 22.1  | 30.4 | 25.6  | 24.5 |
| Horse        | 19.9  | 20.3  | 21.2 | 19.1  | 17.3 |
| Rain         | 12.8  | 13.4  | 13.9 | 12.4  | 12.1 |
| Thunder      | 34.1  | 31.7  | 36.5 | 35.3  | 33.8 |
| Typing       | 27.6  | 29.8  | 31.0 | 28.1  | 27.2 |
| Waterfall    | 10.7  | 11.5  | 12.3 | 11.2  | 9.4  |
| Average      | 21.43 | 22.48 | 23.29| 20.48 | 18.47|

for synthesized and real samples, configuring the mean and covariance of intermediate layer inception features as follows:

$$FID(r, g) = ||\mu_r - \mu_g||^2 + Tr$$

$$\left( \sum r + \sum g - 2 \left( \sum r \sum g \right)^{1/2} \right)^2$$ (7)

Here, $\mu_r$ and $\sum r$ represent the mean and covariance of real samples respectively. Likewise, $\mu_g$ and $\sum g$ represent the mean and covariance of generated samples. The FID score is considered a good evaluation metric to compare between real data and generated outputs. A low FID score is preferred when evaluating generation quality. The existing Inception features pretrained with Imagenet or S09 data will not match the requirements for our specific audio spectrogram generation associated with the video clip. Therefore, we again used the same sound retrieval CNN classifier pretrained on AutoFoley spectrograms to compute the FID scores (shown in Tables II and III).

D. Number of Statistically-Different Bins (NDB)

We followed another effective quantitative evaluation metric called number of statistically-different bins (NDB), proposed in [61]. This method takes up two sets of samples from the same distribution and indicates that the number of samples that fall into a given bin should be the same up to sampling noise. Thus, a NDB score shows the number of cells where the training sample number is statistically different from the generated sample number through a two-sample binomial test. Here, we cluster our training sample $k = 50$ Voronoi cells into a log-spectrogram by $k$-means clustering. Next, we assign the generated samples to the nearest cell by mapping them into the log-spectrogram space. Certainly, a low NDB score is preferred in the case of evaluating good generation quality. The existing Inception features pretrained with Imagenet or S09 data will not match the requirements for our specific audio spectrogram generation associated with the video clip. Therefore, we again used the same sound retrieval CNN classifier pretrained on AutoFoley spectrograms to compute the FID scores (shown in Tables II and III).

E. Quantitative Study Result Analysis

As mentioned in the above subsections, we performed quantitative experiments on the generated samples from our proposed FoleyGAN and other baseline audio generating networks and present the results in Table II, where all the models are trained on our AFD dataset. The FoleyGAN model with visual guidance achieves the highest IS score (10.97) and sound retrieval accuracy (76.08%), which are very close to the experiment results with real samples. However, the generated sample performance deteriorates (lower than AutoFoley, GANSYNTH and Taming VGSG samples) when FoleyGAN is not guided with visual action information. The same trend follows in the case of FID and NDB computations. Our proposed FoleyGAN with visual guidance results in the lowest (considered as better) scores – 67 and 18.47 for FID and NDB, respectively – which again represents good generation quality. Next, we wanted to evaluate our proposed model efficiency on the two most popular video datasets, YouTube8M and AudioSet – the comparative results are shown in Table III. Since most of the audio clips associated with YouTube8M and AudioSet video samples consist of background noise and occasionally sounds from multiple sources, it becomes difficult for the generator to learn the original pattern from latent $z$ from a similar number of training epochs used with AFD video samples. However, the scores are not too far from real data, which leads to the fact that despite foley
generation, our proposed model can be deployed in a general-
ized application of audio synthesis in silent video inputs as well. Later in Table IV, we present a NDB scores of generated sam-
ple of individual AFD classes on FoleyGAN models using 5
different sound encodings (eg. Short-Time-Fourier Transform
(STFT), Mel-Spectrum (MS), Mel-Frequency Cepstral Coef-
ficient (MFCC), Log-amplitude of Mel-Spectrum (LMS), and
Constant-Q Transform (CQT)) as GAN inputs. All class results
showed the lowest value of NDB is calculated for the generated
samples where FoleyGAN is trained with LMS audio features.

F. Ablation Analysis

We performed 2 separate ablation studies performed on AFD
dataset. In the first study (Table V), we showed the significance
of our proposed temporal action information extraction method
using TRN architecture. Here, we kept the generation architec-
ture (BigGAN 512) unchanged. Since in the ablation exper-
iment of our previous AutoFoley paper [1], with 8-scale TRN
we obtained higher sound retrieval accuracy with comparatively
less time, in FoleyGAN research we used the same method for
extracting action class information. But we trained the model
with different action spectrogram matrices where we evaluated
similar action occurring probabilities from subsequent 3 and
4 frames separately. In addition, we substituted our class and
action TRN models with BN-Inception network used in base-
line models ([38], [50]) for visual feature extraction purpose.
We evaluated average accuracy and inference time with differ-
ent extraction and similar generation technique. Table IV shows
close values for 2-frame and 3-frame action TRN techniques.
Whereas, when we are taking 4 consecutive frames in consider-
ation for getting similar action occurrence probabilities, it is tak-
ing longer computation time. In contrary, when we are using the
BN-Inception model [62] to condition our generator with only
action class information, we find significant accuracy degrada-
tion, that highlights the significance of conditioning the gener-
ator with temporal action information of visual scenes. More-
over, TRNs are taking less computation time than BN-Inception
model (the 3rd column of Table V shows the time required by
the models to generate a 3 s sound sample).

In the second ablation experiment, we evaluated FoleyGAN’s
sound retrieval accuracy using BigGAN and MelGAN [49]
arithmetic trained with 2 different losses (e.g. Hinge and
Wasserstein loss [63]). We kept the action extraction technique
unchanged to observe the impact of our proposed and state-of-
the-art GAN model for waveform generation. Table VI results
show that both BigGAN and MelGAN generated waveforms are
performing better with Hinge loss with a small margin. However,
our proposed BigGAN generator is significantly outperforming
MelGAN generator.

| Method | Accuracy (%) | Time (sec) |
|--------|--------------|------------|
| 2 Frame Action TRN + 8 scale Class TRN + BigGAN | 76.08 | 4.7 |
| 3 Frame Action TRN + 8 scale Class TRN + BigGAN | 75.99 | 4.9 |
| 4 Frame Action TRN + 8 scale Class TRN + BigGAN | 71.52 | 5.4 |
| BN Inception + BigGAN | 48.61 | 5.3 |

G. Phase Coherence

To envision the phase coherence between training and gener-
ated waveforms, we show Rainbowgram representations [11] of
each event class in Fig. 5, where the left column is the indicat-
ing rainbowgrams of the originals and the right column displays
the same for generated tracks. The comparison between two
rainbowgrams helps visualize both the phase consistency and
differences of the wave harmonics in a clearer way. In every
rainbowgram image, the brightness symbolizes the log magni-
tudes and the color depicts the instantaneous frequencies of the
respective waveform. Noticeably, the rainbowgrams of the fire,
footstep, rain, and waterfall class synthesized waves depicting
vigorous consistent colors and phase coherence like that of real
waves. Few deformities in color lines are noticed in the break-
ing, chopping, ticking clock, running horse, typing, and thunder-
categories. However, rainbowgrams of the car and gunshot
classes show more phase discontinuities since the wave harmo-
nicas are occasionally afflicted by noise components, which are
responsible for the additional color flecks, phase irregularities,
and aperiodicities.

H. Qualitative Study: Human Survey

We found that a human survey was an inevitable assessment
to judge both the audio quality and its synchronicity with the
video recording, since the human brain can inherently perceive
the correspondence between audio-visual modalities in such co-
inciding events. Therefore, we prepared a research study partic-
ipated in by our College of Engineering students and officials
to survey qualitative questions on our synthesized sound tracks
superimposed on real video clips. There, we set 2 queries for
videos of each event class. Every audience member was asked
to observe videos with our synthesized sounds and rate the gen-
erated sample on the basis of the overall quality of the audio
(question 1) and how much they perceived that the audio was
synchronous with the visual scene (question 2). The observance
score is marked out of a scale of 10 and the experiment was con-
ducted using 100 participants. Audiences are provided with the
ground truth samples as reference. Through this approach, we
intended to capture human’s natural intuition to assess the arti-
ficially synthesized sound quality to determine how accurately
our generated sound traits portray the original event.

| Method | Average Accuracy (%) |
|--------|----------------------|
| FoleyGAN with BigGAN + Hinge loss | 76.08 |
| FoleyGAN with BigGAN + Wasserstein loss | 73.08 |
| FoleyGAN with MelGAN + Hinge loss | 65.97 |
| FoleyGAN with MelGAN + Wasserstein loss | 64.13 |
Fig. 5. Phase coherence comparison between the original and generated sound samples through Rainbowgram representation. Horizontal and vertical axes are showing time and frequency, respectively.

Table VII presents the average ratings for individual classes separately on both queries. The best result for both queries comes from the waterfall sound. We think it is because the training sound clips contain a similar continuous pattern for this category and thus the generator learns it more accurately. Regarding audio-visual synchronicity, the rainfall (9.6), fire (9.4), and clock ticking (9.2) event classes are three other classes that captured the continuous pattern of sound. Additionally, asynchronous event classes (e.g. chopping on kitchen board (9.5), footstep (9.3), horse running (9.2), breaking (9.0), and car (8.8)) also provided outstanding syncing scores. Since object movements are more visible due to close up video recordings (mostly in chopping, footstep, and breaking videos), we assume this helps in generating more synchronous sound with visual guidance.
However, in generated horse clips, we find some variation in sound intensity when the horse is hitting the ground while running or walking. This shows that depending on the action speed, the sound intensity and pitch change, which is a challenging property to learn. In a few cases, we observed that the model was unable to capture this trait and instead learned to generate a more general form of horse running sound. Despite this, the generated tracks are well synced with the test clips, indicating the success in visual guidance introduced to the GAN.

For gunshot and thundering videos, we must rely on videos that are available online for use, as we are not able to record them in person. The thundering category is the most challenging part - in most cases, the lightening visuals were unable to provide action info coherently with audio features while generating sound. However, if we consider the audio quality, the generated thundering audio clips sound similar to the raining sound. In the case of the gunshot sound generation, we find the action of shots are not clearly visible in most recordings due to distant object placement. This may hinder providing temporal action updates to the GAN while generating the sound.

We have the least number of training examples in the breaking category (mostly collected from online sources). We assume this class needs to be developed with the more training samples, with inclusion of a variety of object materials to expect better audio quality. According to this study, people perceive ticking clock, footstep, fire, running horse, rain, and water audio quality well, car, chopping, gunshot, and typing sounds as average, and thundering sound as the least similar to the originals. Averaging all class results, our generated sound score was 7.1 and 8.1 out of 10 in terms of quality and synchronicity with video, respectively.

### VII. Conclusion and Future Scope

In this paper, we address the time synchronization setback in the task of visual-to-audio generation and take the first attempt to exploit conditional GANs with visual guidance of an event to synthesize visually aligned sound. For efficient adversarial training, we expand the AFD dataset with adequate diverse video samples in each class. In order to evaluate our models, we conducted numerical, qualitative, and ablation evaluations and compared them with baseline models with leading results. Our experiments reveal that our proposed FoleyGAN system has the capability of successful synchronous sound synthesis, maintaining good audio quality that can indeed be used as automatic Foley generators for silent movie scenes as well as for other audio-visual intersensory applications. One shortcoming in this work is the requirement that the subject of classification is present in the entire video frame sequence. Furthermore, in our approach, we have not dealt with video clips containing multiple sound sources, and we want to work with more sound categories. These are the targeted directions of our future work.

### References

[1] S. Ghose and J. J. Prevost, "Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning," *IEEE Trans. Multimed.*, vol. 23, pp. 1895–1907, 2021.

[2] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[3] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 10542–10552, 2019.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.

[5] V. Dumoulin et al., "Adversarially learned inference," in *Proc. 5th Int. Conf. Learn. Representations*, Toulon, France, Apr. 24-26, 2017.

[6] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.

[7] T. Karras et al., "Analyzing and improving the image quality of stylegan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.

[8] C. Donahue, J. McAuley, and M. Puckette, "Synthesizing audio with GANs," in *Proc. Int. Conf. Learn. Representations*, 2018.

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

[10] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019.

[11] J. Engel et al., "GANSynth: Adversarial neural audio synthesis," in *Int. Conf. Learn. Representations*, 2019.

[12] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, "Midinet: A convolutional generative adversarial network for symbolic-domain music generation," 2017, arXiv:1703.10847.

[13] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 349–357.

[14] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019.

[15] R. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 99–102, Feb. 1980.

[16] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 803–818.

[17] K. van den Doel, P. G. Kry, and D. K. Pai, "Foleyautomatic: Physically-based sound effects for interactive simulation and animation," in *Proc. 28th Annua. Conf. Comput. Graph. Interactive Techn.*, 2001, pp. 537–544.

[18] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 609–617.

[19] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 801–816.

[20] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 892–900.

[21] F. Wang, H. Nagano, K. Kashino, and T. Igarashi, "Visualizing video sounds with sound word animation to enrich user experience," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 418–429, Feb. 2017.

[22] K. Chen et al., "Visually indicated sound generation by perceptually optimized classification," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018.

[23] A. Owens et al., "Visually indicated sounds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2405–2413.

### TABLE VII

| Class     | Audio Quality | Audio-Visual Synchronicity |
|-----------|---------------|-----------------------------|
| Break     | 7.7           | 9.0                         |
| Car       | 5.1           | 8.8                         |
| Clock     | 7.5           | 9.2                         |
| Chopping  | 5.6           | 9.5                         |
| Fire      | 8.2           | 9.4                         |
| Footstep  | 9.1           | 9.3                         |
| Gunshot   | 5.9           | 4.3                         |
| Horse     | 8.4           | 9.2                         |
| Rain      | 8.9           | 9.6                         |
| Thunder   | 3.2           | 3.6                         |
| Typing    | 6.3           | 5.5                         |
| Waterfall | 9.2           | 9.8                         |
| Average   | 7.1           | 8.1                         |


[24] W. W. Gaver, “What in the world do we hear?: An ecological approach to auditory event perception,” Ecological Psychol., vol. 5, no. 1, pp. 1–29, 1993.

[25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Represent., 2015.

[26] P. Majdak, M. J. Goupell, and B. Laback, “3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training,” Attention, Perception, Psychophys., vol. 72, no. 2, pp. 454–469, 2010.

[27] B. Shelton and C. Searle, “The influence of vision on the absolute identification of sound-source position,” Percep. Psychophys., vol. 28, no. 6, pp. 589–596, 1980.

[28] R. S. Bolla, W. R. D’Angelo, and R. L. McKinley, “Aurally aided visual search under virtual and free-field listening conditions,” Hum. Factors, vol. 41, no. 4, pp. 664–669, 1999.

[29] D. R. Perrott, J. Cisneros, R. L. McKinley, and W. R. D’Angelo, “Aurally aided visual search in three-dimensional space,” Hum. Factors, vol. 38, no. 4, pp. 702–715, 1996.

[30] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, “Learning sight from sound: Ambient sound provides supervision for visual learning,” Int. J. Comput. Vis., vol. 126, no. 10, pp. 1120–1137, 2018.

[31] N. Takahashi, M. Gygly, and L. Van Gool, “A3Net: Learning deep audio features for video generation,” IEEE Trans. Multimedia, vol. 20, no. 3, pp. 513–524, Mar. 2017.

[32] K. Xu et al., “Show and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 2048–2057.

[33] T. Salimans et al., “Improved techniques for training GANs,” in Proc. Adv. Neural Inf. Process. Syst., 2016.

[34] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, “Visual to sound: Generating natural sound for videos in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3550–3558.

[35] C. Zhang et al., “Visually indicated sound generation by perceptually optimized classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 324–333.

[36] R. Gao and K. Grauman, “2.5D visual sound,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 4950–4959.

[37] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, “Vision-infused deep audio inpainting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 283–292.

[38] P. Chen et al., “Generating visually aligned sound from videos,” IEEE Trans. Image Process., vol. 29, pp. 8292–8302, 2020.

[39] S. Liu, S. Li, and H. Cheng, “Towards an end-to-end visual-to-raw-audio generation with GAN,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1299–1312, Mar. 2021.

[40] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, “Self-supervised generation of spatial audio for 360 video,” in Proc. Adv. Neural Inf. Process. Syst., vol. 31, 2018.

[41] D. Li, T. R. Langlois, and C. Zheng, “Scene-aware audio for 360 videos,” ACM Trans. Graph., vol. 37, no. 4, pp. 1–12, 2018.

[42] S. Liu, H. Cheng, and Y. Tong, “Physically-based statistical simulation of rain sound,” ACM Trans. Graph., vol. 38, no. 4, pp. 1–14, 2019.

[43] J. Zheng, S.-H. Hung, K. Hiebel, and Y. Zhang, “Real-time rendering of decorative sound textures for soundscapes,” ACM Trans. Graph., vol. 39, no. 6, pp. 1–12, 2020.

[44] Z. Zhang, N. Raghuvanshi, J. Snyder, and S. Marschner, “Acoustic texture rendering for extended sources in complex scenes,” ACM Trans. Graph., vol. 38, no. 6, pp. 1–9, 2019.

[45] J. Donahue, P. Krähenbühl, and T. Darrell, “Adversarial feature learning,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2227–2235.

[46] K. N. Haque, R. Rana, and B. W. Schuller, “High-fidelity audio generation and representation learning with guided adversarial autoencoder,” IEEE Access, vol. 8, pp. 225309–225328, 2020.

[47] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[48] J. Allen, “Short term spectral analysis, synthesis, and modification by discrete fourier transform,” IEEE Trans. Acoust., Speech, Signal Process., vol. 25, no. 3, pp. 235–238, Jun. 1977.

[49] J. P. Gemmke et al., “Audio set: An ontology and human-labeled dataset for audio events,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2017, pp. 776–780.

[50] S. Abu-El-Haija et al., “Youtube-8m: A large-scale video classification benchmark,” 2016, arXiv:1609.08675.

[51] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 7354–7363.

[52] A. M. Saxe, J. L. Mcclelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural network,” in Proc. Int. Conf. Learn. Representations, 2014.

[53] T. Salimans et al., “Improved techniques for training GANs,” in Proc. Adv. Neural Inf. Process. Syst., 2016.

[54] C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.

[55] E. Richardson and Y. Weiss, “On GANs and GMMS,” in Proc. Adv. Neural Inf. Process. Syst., 2018.

[56] S. Joffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 449–456.

[57] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 214–223.

Sanchita Ghose (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from the Ahsanullah University of Science and Technology, Dhaka, Bangladesh, in 2013, and the M.S. and Ph.D. degrees in electrical engineering from the University of Texas at San Antonio, San Antonio, TX, USA, in 2021 and 2022, respectively. Her research interests include developing deep learning algorithms for multimodal learning and cross-modal retrieval applications, focusing on computer vision, action recognition, sound synthesis, and video processing. During her doctoral study, she was the recipient of multiple awards, including the Outstanding Graduate Research Award, ECE Pioneer Competitive Award, Best Poster in Applied AI Research-AI SUMMIT 2019, ECE Professional Development Award, ECEDHA iREDIFINE Scholar Award, and NSF-SWEETER Grant Award. She is a Student Member of the IEEE and a Senior Member of the IEEE. She has previously authored or coauthored in IEEE TRANSACTIONS ON MULTIMEDIA and had been reviewer of the same and IEEE ACCESS.

John J. Prevost (Senior Member, IEEE) received the first B.S. degree in economics from Texas A&M University, in 1990, the second B.S. degree (magna cum laude) in electrical engineering from the University of Texas at San Antonio (UTSA), San Antonio, TX, USA, in December 2009, and the M.S. and Ph.D. degrees in electrical engineering from UTSA, in December 2012 and December 2013. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, UTSA. In 2015, he co-founded and became the Chief Research Officer and Executive Director of the Open Cloud Institute. He is also the VP for Secure Cloud Architecture with the Cyber Manufacturing Innovation Institute (CyManII) University of Texas at San Antonio. Prior to his academic appointment, he was the Director of product development and information systems and Chief Technical Officer of various technical firms. He remains an Active Consultant in the areas of complex systems and cloud computing, and maintains strong ties with industry leaders. His research interests include IoT/edge-computing security and optimization, applied machine learning, and quantum informatics.