RD-FGFS: A Rule-Data Hybrid Framework for Fine-Grained Footstep Sound Synthesis from Visual Guidance

Qiutang Qi
Communication University of China
Beijing, China
qiqiutang@cuc.edu.cn

Haonan Cheng*
Communication University of China
Beijing, China
haonancheng@cuc.edu.cn

Yang Wang
North China Institute of Science and Technology
Beijing, China
wy0926go@163.com

Long Ye
Communication University of China
Beijing, China
yelong@cuc.edu.cn

Shaobin Li
Communication University of China
Beijing, China
shaobin@cuc.edu.cn

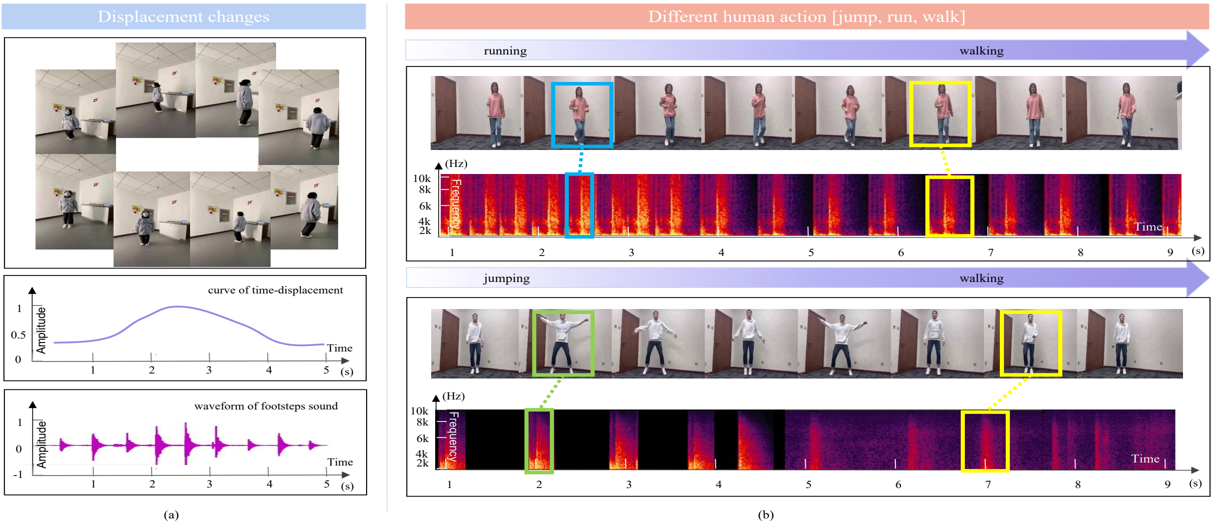


Figure 1: Two RD-FGFS synthesis examples. (a) shows the visual motion displacement images, the generated displacement trajectory cues and the synthesized sound waveforms based on the cues, from top to bottom. (b) illustrates the generated sound spectrograms under two different motions (running to walking and jumping to walking). In particular, we mark the video frames and the corresponding spectra for running, walking and jumping with blue, yellow and green boxes.

ABSTRACT

Existing methods are difficult to synthesize fine-grained footsteps based on video frames only. This is due to the complicated nonlinear mapping relationships between motion states, spatial locations and different footstep sounds. Aiming to address this issue, we propose a **R**ule-**D**ata guided **F**ine-**G**rained **F**ootstep **S**ound (RD-FGFS) synthesis method. To the best of our knowledge, our work takes the first step in integrating data-driven and rule modeling approaches for visually aligned footstep sound synthesis. Firstly, we design a learning-based footstep sound generation network (FSGN) architecture driven by pose and flow features. The FSGN is proposed for generating an initial target sound which captures timing cues. Secondly, a rule-based fine-grained footstep sound adjustment (FGFSA) method is designed based on the visual guidance, namely ground material, movement type, and displacement distance. The proposed FGFSA effectively constructs a mapping relationship between different visual cues and footstep sounds, enabling fine-grained variations of footstep sounds. Experimental results show that our method improves the visual and sound synchronization results of footsteps and achieves impressive performance in footstep sound fine-grained control.

*indicates corresponding authors.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613765

CCS CONCEPTS

• **Computing methodologies** → *Scene understanding*; • **Applied computing** → **Sound and music computing**.

KEYWORDS

Sound Synthesis, Footstep Sound, Rule-Data Hybrid Framework, Visual Guidance, Procedural Audio

ACM Reference Format:
Qiutang Qi, Haonan Cheng, Yang Wang, Long Ye, and Shaobin Li. 2023. RD-FGFS: A Rule-Data Hybrid Framework for Fine-Grained Footstep Sound Synthesis from Visual Guidance. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3613765

1 INTRODUCTION

Footstep sound effects, as one of the core sound effects, play an important role in enhancing the sense of realism, immersion, and presence [19]. With different movement states and different spatial positions, the footstep sound shows various variations. Consequently, professional sound engineers manually perform labor-intensive time alignment and content management to generate fine-grained variations of footstep sounds. In this paper, we propose a method for the automatic synthesis of fine-grained footstep sound effects that takes advantage of both cross-modal learning and rule modeling. Two examples of our method are shown in Figure 1.

Visually aligned sound synthesis can be divided into three categories: manual-based methods, rule-based modeling methods, and data-driven methods. Initially, Foley artists generate sound for film post-production manually [40]. Later, the emergence of a large number of dubbing tasks makes it difficult to complete the manual method alone. Signal-based [2, 33] and physics-based [6, 7, 29] rule-based sound synthesis approaches are gradually becoming a hot research topic. Signal-based methods utilize sample signals for analytical modeling to achieve sound synthesis, while physics-based methods are usually based on animation information such as object position and motion state. However, such methods rely on visual parameters that need to be obtained with the help of graphical simulations, which are difficult to obtain in movies. In recent years, the increasing development of deep learning has made data-driven sound synthesis methods [1, 4, 12, 28, 39] mainstream. This class of methods explores the idea of training neural networks to achieve Foley automation, which uses large amounts of video-audio data pairs to learn synchronized mapping relationships and thus achieve sound effect generation. However, these methods cannot distinguish fine-grained features such as motion and material.

In summary, existing works on controllable sound synthesis are unsuitable for the fine-grained footstep sound synthesis task due to the following two reasons: (1) the lack of fine-grained representation of footstep sound, as existing methods only synthesize a single category of footstep sound without further fine-grained representation. (2) the unclear audio-visual mapping relationships of footstep sound, as the audio-visual mapping relationships of footsteps are ambiguous, lacking definite audio-visual mapping cues.

To address the aforementioned problems, this paper proposes a **R**ule-**D**ata guided **F**ine-**G**rained **F**ootstep **S**ound (RD-FGFS) synthesis method. The proposed method consists of a rule-based modeling module and a data-driven sound synthesis module to achieve fine-grained content consistency. Firstly, a learning-based footstep sound generation network (FSGN) is designed to initially ensure temporal coherence between vision and sound. This FSGN module enables an effective audio-visual alignment. This is benefiting from a combination of features that we constructed to represent the fine-grained variation of footsteps, namely pose and flow. Then, in order to further improve the synchronization effect and realism, we propose a rule-based fine-grained footstep sound adjustment (FGFSA) module. This module explores the mapping rules between foot pose and parametrically control of footstep sound. By this means, we can generate fine-grained controllable footstep sound. Since the existing Foley sound synthesis datasets have limited footstep types and the frame does not contain a full pose, we construct a dataset for footstep sounds, named VAFoot. This dataset contains a total of 7000 seconds of video with different ground materials, motion types and motion directions. To sum up, the major contributions of this paper are as follows:

- We propose a rule-data hybrid framework for fine-grained footstep sound synthesis in a coarse-to-fine manner. To the best of our knowledge, this is the first controllable visually-guided sound synthesis framework which solves the lack of representation problem associated with existing works.
- The designed FGFSA module that integrates a visual pose analyzer, timber and loudness adjustment effectively constructs audio-visual mappings, making parametric sound control applicable.
- The proposed RD-FGFS method achieves outstanding performance in comparison with the state-of-the-art methods on the VAFoot dataset.

2 RELATED WORK

The visually aligned sound synthesis is always a focus of researchers. In 1992, Takala and Hahn [34] propose one of the earliest sound rendering pipelines, aiming to produce the composite soundtrack from component sound objects. Later, various advanced sound synthesis methods developed. The sound synthesis methods can be aggregated into rule-based and data-driven methods.

Rule-based modeling sound synthesis methods include signal-based and physics-based methods. Signal-based methods [2, 25, 41] adopt signals such as motion, texture, and spectrum for analysis and modeling to achieve sound synthesis. This type of method can produce perceptually reasonable sounds, but most require laborious manual control. And that is difficult to synchronize with visual animations. In contrast, physics-based methods [21, 22, 29, 30] usually calculate sound pressure based on the animation's object motion and acoustic equations to render visual synchronization without additional manual editing. For the synthesis of footsteps sound, Cook [7] proposes an algorithm for automatic parameter synthesis based on physically informed stochastic model (PhiSM) [6]. However, it requires additional parameter tweaking. Subsequently, further footstep sound synthesis methods [8, 10, 27, 36, 37] focus on exploring aspects such as action, material,

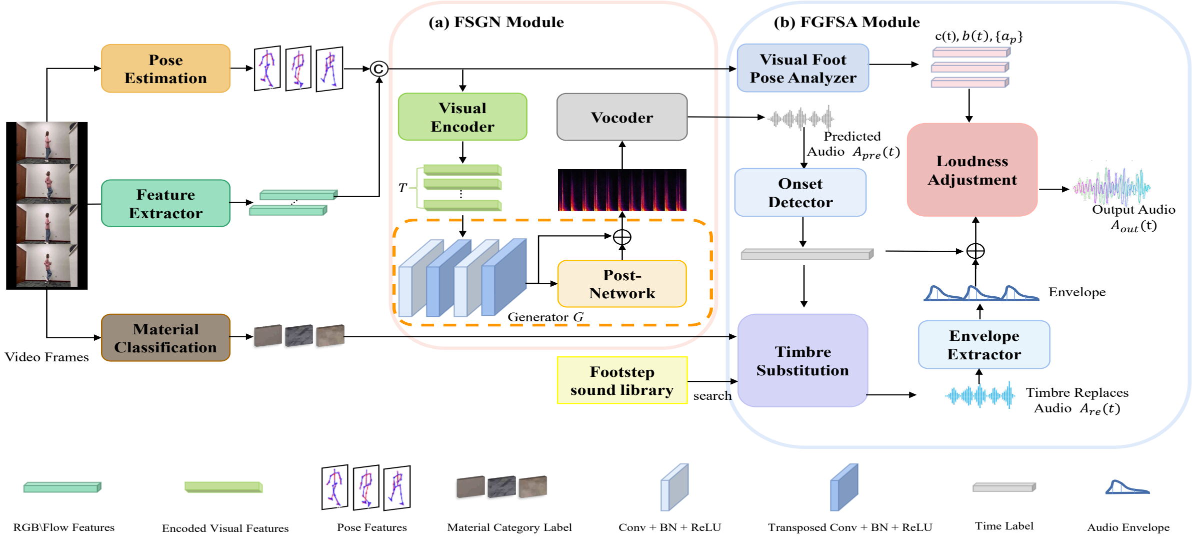


Figure 2: The pipeline of RD-FGFS method. Our model takes footstep sound and video frames as input. The audio and video data are processed by STFT and feature extraction, pose estimator and meterial classfication, respectively. In the next step, the initial footstep sound is generated by visual encoder, generator, and vocoder in the FSGN module. After then, in the FGFSA module, visual foot pose analyzer, timbre substitution and loudness adjustment are applied to the initial footstep sound to synthesize the final footstep sound.

or providing multimodal cues. Nevertheless, the above methods cannot achieve the task of generating fine-grained footstep sounds given a video well. This is due to the significant differences in expression and features between audio and video signals, making it difficult to build correlation.

Data-driven deep learning methods fortunately provide new insights for visually aligned sound synthesis by directly learning the mapping between vision and sound. Owens et al. [28] propose an algorithm that adopts the neural network to synthesize the sound of drumstick hitting. Then, Chen et al. [3] propose a method of generating a single class of sound using perceptual loss. However, these algorithms cannot directly generate raw audio signals due to differences in audio-visual time-space scales and feature structures. Zhou et al. [42] propose an end-to-end sound generation method based on samplerRNN [26] on an unconstrained dataset. However, experiments by Owens and Mehri [26, 28] show that RNNs have limitations in learning visual content, making it difficult to achieve ideal synchronization. To achieve better audio-visual synchronization and sound quality, more methods have been proposed, such as RegNet [5], V2RA-GAN [23], SpecVQGAN [16], and SPMNet [24]. For Foley sound production, Ghose and Prevost [12] propose a video-based automatic sound effects synthesis algorithm - AutoFoley, which was later improved using generative adversarial networks (GAN) [14] named FoleyGAN [13]. However, almost all the aforementioned methods only learn the category mapping, which often fails to fine-grained control of footstep sound synthesis.

3 METHOD

The designed rule-data hybrid method for visually guided fine-grained footstep sound synthesis is illustrated in Figure 2. To be specific, we first propose a data-driven FSGN module (as shown in Figure 2 (a)) to explore the synchronization between visual and audio, and synthesize rough footstep sounds. In the FSGN module, a novel visual fusion feature captures better audio-visual timing cues. Then, to further fine-grained adjustment of the footstep sound, we also design a rule-based FGFSA module (as shown in Figure 2 (b)). This module constructs a mapping relationship between vision and sound. By this means, the method can control footstep sound fine-grained, thereby resulting in better audio-visual synchronization and sound quality. More details of the aforementioned modules utilized in our method are presented in the following sections.

3.1 Data-driven FSGN Module

To obtain audio-visual synchronization cues from footsteps, we design a FSGN module based on a data-driven approach using deep learning for audio-visual generation. Figure 2 (a) illustrates the structure of the module proposed in this section. The module mainly consists of a visual encoder, a sound generation network, and a vocoder.

The FSGN module aims to generate a synchronized footsteps sound $\{A_{pre}(t)|t = 1, ..., T_L\}$ given the corresponding sequence of video frames $\{I_n|n = 1, ..., N\}$ as visual inputs, where T_L is total time, N is total frame number. The visual feature vector V_n is obtained by concatenating the pose feature vector VP_n, the flow feature vector VF_n, and the RGB feature vector VR_n of the n-th frame. The concatenation process of visual features can be described as follows:

$$VF_n = \psi(J_n), \tag{1}$$

$$VR_n = \psi(I_n), \tag{2}$$

$$VP_n = PCA(\phi(I_n)), \tag{3}$$

$$V_n = Concat(VF_n, VR_n, VP_n), \tag{4}$$

where $\psi(\cdot)$ represents feature extraction operation through the BN-inception network [17]. $\phi(\cdot)$ is the operation to extract the pose feature of visual pose estimation network [20]. $PCA(\cdot)$ is the operation that reduces the dimensionality of features through principal component analysis (PCA). I_n and J_n are the video frame and space-time image of the n-th frame. Finally, we obtain a 3072-dimensional visual vector V_n in n-th frame, where the vector dimensions of VF_n, VR_n, and VP_n are all 1024. For the entire video, the visual feature vector can be obtained as $\{V_n\}_{n=1}^{N}$.

Different from prior works [5, 16] where visual features only adopt RGB and flow features, we complement the pose feature for the characteristics of pedestrian movement. To reduce the computational cost and ensure the balance between features, we reduce the dimensionality of pose features to 1024. In the training process, the raw audio $A_{GT}(t)$ is converted to a series of spectrograms $\{S_m\}_{m=1}^{M}$ via a Short Time Fourier Transform (STFT), where M is the number of total spectrograms.

3.1.1 *Audio-visual Feature Fusion*.

After obtaining the audio and visual features, we further fuse the features. To compensate the difference between audio and video sampling rates, we employ a visual encoder and an audio encoder. The visual encoder consists of three one-dimensional convolutional layers and a Bidirectional Long Short-Term Memory (Bi-LSTM) [15], which can better capture bidirectional long-distance audio-visual dependencies. The outputs from the forward and backward paths of the Bi-LSTM at each time step are concatenated as T encoded visual features $F^v = \{F_1^v, F_2^v, ..., F_T^v\}$. For sound encoding, we adopt a two-layer Bi-LSTM to process the ground truth spectrogram with the cell dimension D_m. The outputs from the two paths of Bi-LSTMs are concatenated to form a $2D_m \times T'$ feature map, where T' is the time dimension of the input spectrogram. The feature map is then uniformly down-sampled by a factor of S_p. To match the temporal dimension of the encoded visual features, we upsample it by copying and generate a $2D_m \times T$ output. The output dimensions D_m and down-sampling rate S_p can be adjusted to control the richness of sound information. Afterwards, the encoded sound features $F^s = \{F_1^s, F_2^s, \ldots, F_T^s\}$ are obtained. Finally, we obtain the fused audio-visual features $F^c = \{F_1^c, F_2^c, ..., F_T^c\}$.

$$F^c = Concat(F^s, F^v), \tag{5}$$

3.1.2 *Sound Generation Network*.

In order to predict the generated audio spectrogram $\{S'_m\}_{m=1}^{M}$, we introduce a generative adversarial network (GAN) [14], which consists of a generator and a discriminator, as the audio generation network. Specifically, we first predict the audio spectrogram SG' by two one-dimensional convolutional layers and two one-dimensional transposed convolutional layers. Furthermore, we add a post-network [31] to supplement fine structures SP'. Finally, the generator $G(\cdot)$ obtained $S' = SG' + SP'$. The discriminator $D(\cdot)$ takes the extracted visual features and spectrogram as input to distinguish whether the spectrogram comes from a real or fake video. To maintain high-resolution structures over the temporal scale, we introduce PatchGANs [18] in discriminator. During training, the generator tries to minimize the following loss:

$$L_{rec} + L_G = \mathbb{E}[\|G(F^c) - F^s\|_2^2] + \mathbb{E}[log(1 - D(G(F^c), F^v))], \tag{6}$$

where the first term is the L2 reconstruction error, and the second term is the adversarial loss. The discriminator minimizes the following loss:

$$L_D = -\mathbb{E}[log(D(F^s, F^v))] - \mathbb{E}[log(1 - D(G(F^c), F^v))], \tag{7}$$

In addition, we add the $L2$ constraint on the predicted spectrogram SG', which is expressed as follows:

$$L'_{rec} = \mathbb{E}[\|SG' - F^s\|_2^2], \tag{8}$$

In total, the whole loss be summarized as follows:

$$L_{total} = L_{rec} + \alpha L'_{rec} + \beta(L_D + L_G), \tag{9}$$

In the testing phase, we do not involve the ground truth sound. Instead, we consider the $G(F^v)$ as the predicted results. Finally, we adopt WaveNet [38] as the vocoder to convert the synthesized spectrogram S'_m into waveform $A_{pre}(t)$.

3.2 Rule-based FGFSA module

Although the data-driven FSGN method can generate synchronized footstep sound $A_{pre}(t)$, this sound can only ensure consistency in terms of categories and cannot achieve more fine-grained distinctions. To achieve the goal of fine-grained sound control based on visual clues and obtain more realistic sound, we design the FGFSA module. Based on the two invariants of footsteps namely surface material and human motion characteristics, our FGFSA module is divided into two parts: 1) timbre substitution based on visual ground material, and 2) loudness adjustment based on visual motion analysis.

3.2.1 *Visual Ground Material based Timbre Substitution*.

The perception of realism in a scene by humans is affected by the feedback of ground material sounds during walking. To ensure consistency between the sound timbre and the visual ground material, and to increase the realism of the generated sound, we design the timbre substitution for modifying sound timbre and built a small footsteps sound (SFoot) library.

Specifically, to achieve recognition and annotation of visual ground materials while balancing efficiency and classification accuracy, we design a material classification network using EfficientNet [35]. For the n-th video frame I_n, the material classification network outputs a visual material label $M_V = \{concrete, tile, carpet\}$. Then, an onset detector is adopted to annotate the landing time of the footstep sound $A_{pre}(t)$ generated by the FSGN module, obtaining the time labels $Onset_{time} = \{o_1, o_2, \ldots, o_z\}$, where z is total onset number. The onset detector obtains the final onsets through peak detection and adaptive thresholding, utilising spectral flux as the detection density. Next, we input the visual material label M_V and time labels $Onset_{time}$ into the sound timbre substitution. To implement visual and audio material mapping, the first step of the model retrieves the corresponding footstep sound samples $Sample_M$ from the SFoot library based on the visual ground material label M_V and the SFoot library material label $M_A = \{concrete, tile, carpet\}$. In the second step, we segment and adjust the SFoot library samples based on the time labels $Onset_{time}$. We replace the sound according to the duration of a single footstep and synthesize the replaced footstep sound $A_{re}(t)$. Our SFoot library is based on existing sample banks of footstep sounds (such as Adobe sound library). In addition, users can substitute their own sound samples in the experiment. The high-quality footstep sound samples ensure the quality of the final synthesized sound.

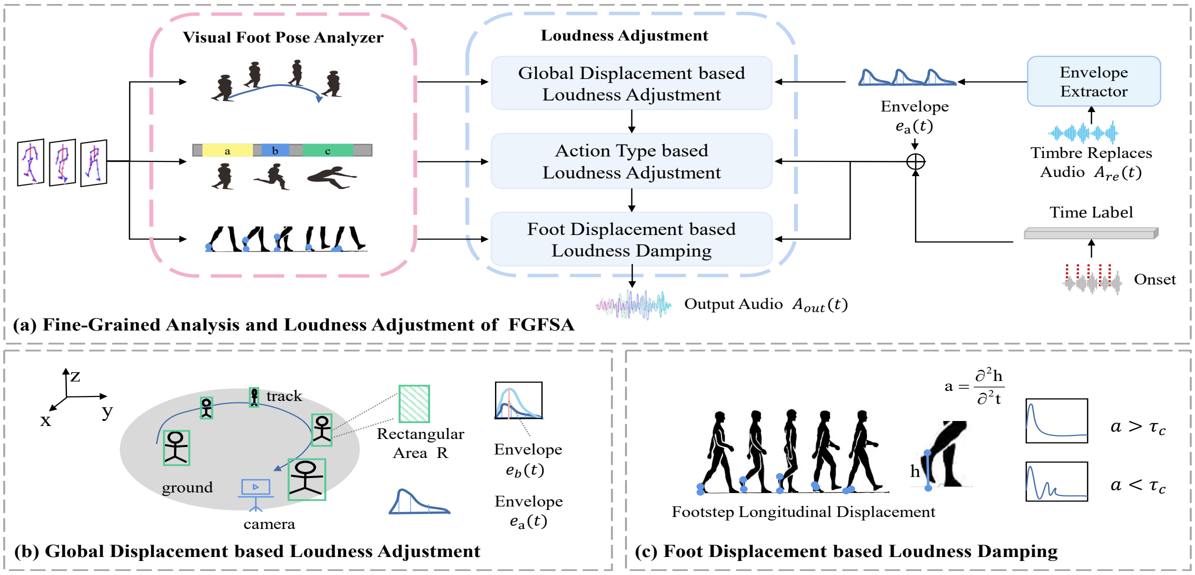

Visual Foot Pose Analyzer
a
b
c
Loudness Adjustment
Global Displacement based Loudness Adjustment
Action Type based Loudness Adjustment
Foot Displacement based Loudness Damping
Envelope $e_a(t)$
Envelope Extractor
Timbre Replaces Audio $A_{re}(t)$
Time Label
Onset
Output Audio $A_{out}(t)$
(a) Fine-Grained Analysis and Loudness Adjustment of FGFSA
z
x
y
track
ground
camera
Rectangular Area R
Envelope $e_b(t)$
Envelope $e_a(t)$
(b) Global Displacement based Loudness Adjustment
$a = \frac{\partial^2 h}{\partial^2 t}$
h
$a > \tau_c$
$a < \tau_c$
Footstep Longitudinal Displacement
(c) Foot Displacement based Loudness Damping

Figure 3: The architecture of FGFSA module. (a) is a demonstration of the overall fine-grained analysis and loudness adjustment structure of the FGFSA module, (b) is a demonstration of adjusting the loudness according to the global displacement, and (c) is a demonstration of adjusting the loudness damping to the single foot displacement.

3.2.2 ***Foot Position Analysis based Loudness Adjustment.*** In order to achieve fine-grained control of footstep sounds based on visual motion cues, we conduct fine-grained visual foot position analysis and loudness adjustment in the FGFSA module (as shown in Figure 3).

Envelope Extractor. Control of sound loudness can be achieved by adjusting the sound envelope. Therefore, we introduce a nonlinear low-pass filter [7] to extract and analyze the envelope of the walking sound at the first stage. We extract the envelope $e_a(t)$ from the sound $A_{re}(t)$:

$$e_a(t) = (1 - k(t))|A_{re}(t)| + k(t)e_a(t-1), \tag{10}$$

where $k(t)$ is the regulating parameter. If $|A_{re}(t)| > e_a(t-1)$, $k(t) = k_{up}$, otherwise $k(t) = k_{down}$. And t and $t-1$ indicate respectively the current and previous time. Typical values for a 22,050-Hz sample rate clapping/walking file are $k_{up} = 0.8$ and $k_{down} = 0.995$. This nonlinear filter contributes to ensure accurate tracking of peaks while eliminating false high-frequency components.

Visual Foot Pose Analyzer. In order to better analyze the visual motion of pedestrians, we design a visual foot pose analyzer. According to the magnitude of influence, we divide the analysis of vision into three layers: global displacement, action type, and single foot displacement. Firstly, the sound source position is a macro factor that affects the sound volume. Without camera calibration, it is difficult to obtain the accurate human position. However, changes in the body contour of people in the visual scene can be used to estimate changes in human displacement. In the same visual scene, the larger the human bounding box area, the closer the distance. Based on the visual pose estimation, we calculate and obtain the human bounding box area R. After curve fitting and normalization, we finally obtain the global displacement control factor $c(t)$.

Secondly, the amplitude of footstep sounds increases in the order of walking, running, and jumping. Therefore, when there are multiple actions in the same video, the sound amplitude should be distinguished accordingly. To obtain action labels, we employ the SlowFast Network [9] for detecting human motion. We obtain the action control factor $b(t)$ by clustering the time intervals of the different actions.

Finally, the landing force of a single foot also has an impact on the loudness. In footstep actions, the greater the landing force, the greater the ground reaction force (GRF), and the faster the loudness decays and the shorter the duration. However, GRF cannot be directly obtained from the video, so we approximate it using displacement acceleration. The second derivative of the vertical displacement distance between the two feet h is used to calculate the displacement acceleration ac. For each segment according to $Onset_{time}$, we take the mean and normalize it to obtain the single-foot force control factor $a = \{a_p\}_{p=1}^{P}$, where P is the fragments number. Through the above analysis of human motion, we finally obtain the visual control factors from displacement to action to single-foot strength.

Loudness Adjustment. Since visual parameters extracted by the visual foot pose analyzer are difficult to precisely match existing synthesis engines (Sound Design Toolkit or Nemisindo) for complex parameters such as reverberation, distortion, compressor, etc. The loudness adjustment is designed in order to achieve a mapping of visual features to sound. Corresponding to the visual foot pose analysis, we divide the parameter control of the loudness adjustment into three steps: First, the overall amplitude of the loudness is adjusted by the global displacement change. Secondly, the amplitude of the loudness for a single footstep is adjusted based on the action type. Thirdly, the loudness damping is adjusted by single foot displacement.

Firstly, overall displacement changes have a global effect on sound loudness and therefore it can be used as a baseline for sound loudness. In the visual foot pose analysis, we obtain the global displacement control factor $c(t)$. Based on this, we obtain the sound envelope $e_b(t)$ after global displacement adjustment as follows:

$$e_b(t) = e_a(t)c(t), \tag{11}$$

Next, we perform more fine-grained adjustments to the footstep sound volume based on the action control factor $b(t)$. The sound envelope $e_c(t)$ after the motion adjustment is as follows:

$$e_c(t) = e_b(t)b(t), \tag{12}$$

where

$$b(t) = \begin{cases} b_{jump}, & if \quad mt = jump \\ b_{run}, & if \quad mt = run \\ b_{walk}, & if \quad mt = walk \end{cases} \tag{13}$$

where mt is the motion type, b_{jump}, b_{run} and b_{walk} is 1.34, 1.17 and 1 respectively, according to the amplitude ratio of the three actions.

Then, we adjust the footstep sound volume based on the single-foot force control factor a. First, the $e_c(t)$ is split into fragments $\{e_c^i(t), i \in 1, ..., P\}$ according to $Onset_{time}$. Then depending on a, we adjust damping ρ by setting the threshold τ_c. The modified envelope fragment $e_d^i(t)$:

$$e_d^i(t) = e_c^i(t) sin(2\pi f t)e^{-\rho t}, \tag{14}$$

where f is the frequency (random assignment between 1 kHz and 16 kHz). And if $a > \tau_c$, $\rho = f/0.1$, otherwise $\rho = f$. The threshold value τ_c is determined based on the calculation of the acceleration. The $e_d(t)$ can be obtained by concatenating $e_d^i(t)$. Finally, we obtain the final output audio $A_{out}(t)$ through the envelope $e_d(t)$.

4 EXPERIMENTS

4.1 Dataset

Although existing audio-visual datasets [11, 32, 42] have footstep categories, the data contents do not contain full gestural motion and have limited data types. To address these issues, we construct a dataset VAFoot for footstep sound synthesis (as shown in Figure 4).

In order to adapt the footstep sound data for more situations with generalization capability, different settings are made for footstep motion type, motion direction, and ground material. Moreover, we record audiovisual data with a mobile phone where the sound source contains only footsteps and process the data for denoising. Finally, we construct the VAFoot dataset containing 7000 seconds of video in MP4 format. The recorded videos are further split into 700 video segments, each of which has a duration of 10 seconds. Specifically, we combine three motion types (walking, running and jumping) with three motion directions (straight line, circle and in-situ). In order to increase the diversity of data, we add a group of random motion modes. In summary, there are a total of 10 motion combinations in the VAFoot dataset.

In addition, there is a correlation between footstep sound timbre and ground material type. In order to achieve fine-grained timbre adjustment, we set 3 common ground material types during recording: concrete, tile, and carpet. The VAFoot dataset can meet the needs for fine-grained synthesis of footstep sounds.

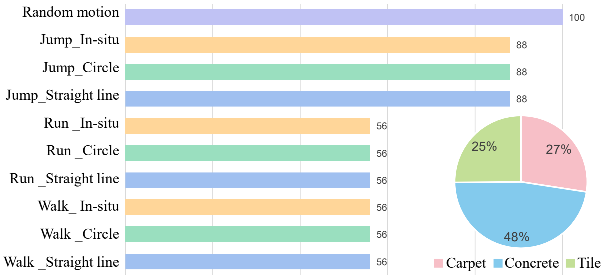


Figure 4: The dataset statistics of VAFoot.

4.2 Implementation Details

In the pre-processing part of the audio and video data we refer to the RegNet [5] data pre-processing scheme. The raw waveform is converted to a spectrogram via an STFT with a sampling rate of 22,050 Hz. The STFT hop size is set to 256 with a window length of 1,024, and the time dimension of the spectrogram is 860. We train the proposed FSGN module for 10,000 epochs adopting the Adam optimization method with an initial learning rate of 0.002. We set α and β to 1 and 10,000 in Eq.9, respectively. In accordance with the original paper's setup, we retrain the three baseline methods on the VAFoot dataset. All the experiments are trained on a single NVIDIA GeForce RTX 2080 Ti GPU.

4.3 Evaluation Metrics

In order to quantitatively evaluate the audio-visual synchronization performance of the algorithm, we design five metrics to evaluate. To evaluate the number of synchronized audio-video matches, we first calculate the proportion of onset correct matches (NCM), redundant matches (NRM) and missing matches (NMM) between the real sound and the predicted sound. In particular, we perform onset detection on the real recorded sound A_{GT} and the predicted sound A_{pre} to obtain the corresponding $\{Onset_{A_{pre}}\}_{i=1}^{Num_{pre}}$ and $\{Onset_{A_{GT}}\}_{i=1}^{Num_{GT}}$. The onset detector utilizes the spectral flux as the detection density and obtains the final onset through peak detection and adaptive threshold processing. If the onset distance between the predicted sound and the real sound is less than 125 ms, the onset is accepted as a correct match and the number of correct matches is counted as Num_{corr}. The number of redundant matches Num_{redu} is the number of correctly matched onsets Num_{corr} subtracted from the predicted number of onsets Num_{pre}. The number of missing matches Num_{miss} is represented by subtracting the number of correctly matched onsets Num_{corr} from the real onset number Num_{GT}. The total number of matches $Num_{total} = Num_{corr} + Num_{redu} + Num_{miss}$. The NCM, NRM, and NMM are calculated as:

$$\begin{aligned} NCM &= Num_{corr}/Num_{total} \\ NRM &= Num_{redu}/Num_{total} \\ NMM &= Num_{miss}/Num_{total} \end{aligned} \tag{15}$$

Moreover, to evaluate the audio-visual synchronic distance difference, we calculate the cumulative distance (DIS) and the average distance (DISAV) between the predicted sound and the real sound.

Table 1: Quantitative evaluation for the results

Experiment	NCM↑	NRM↓	NMM↓	DIS↓	DISAV↓
RegNet [5]	0.676	0.175	0.149	0.669	0.044
SpecVQGAN [16]	0.306	0.377	0.317	0.806	0.053
SPMNet [24]	0.694	0.178	0.127	0.575	0.038
ours-FSGN	**0.712**	**0.169**	**0.120**	**0.566**	**0.037**

Specifically, based on the number of real sound onsets Num_{GT}, the same number of predicted sound onsets with the closest distance to the real sound $\{Onset_{A_p\sim G}\}_{i=1}^{Num_{A_p\sim G}}$ were selected. The cumulative and average distances between the real sound and the predicted sound onset:

$$\begin{aligned} DIS =& |Onset_{GT} - Onset_{A_p\sim G}| \\ DISAV =& |Onset_{GT} - Onset_{A_p\sim G}| / Num_{GT} \end{aligned} \tag{16}$$

4.4 Comparison with State-of-the-Arts

In this section, we compare our proposed method with state-of-the-art data-driven methods RegNet [5], SpecVQGAN [16] and SPMNet [24] on the VAFoot dataset for synchronization performance.

Quantitative and qualitative results are presented in Table 1 and Figure 5, respectively. The results show that our method has outstanding performance in audio-visual synchronization. Table 1 shows that our proposed method brings improvements of 1.80%, 0.90%, 0.70%, 0.90% and 0.10% in all metrics compared to the previous best synchronization method SPMNet [24]. Furthermore, we show the qualitative evaluation results (as shown in Figure 5). The first column shows the spectrogram results, and the second column shows the comparison of the sound envelope. RegNet loses the sound onset at the black circle, while SpecVQGAN loses a large area of sound onset at the blue circle. Although SPMNet do not lose the onset, there are more glitches in the envelope bottom as shown in the green circle. Our method performs better in both aspects. This is due to our method incorporating human pose features, which are more focused on unique human movements compared to other methods, resulting in better performance. This also demonstrates that our FSGN module can effectively achieve synchronization.

4.5 Ablation Study

Ablation experiment of FSGN module. We compare the effects of different feature combinations and time thresholds on the synchronization performance, as shown in Table 2. The first part of the experiment shows the results of three combinations of visual features (RGB, flow, and pose) and attention mechanisms. Among them, the combination of Flow+Pose achieves the best results in the NCM, NRM, NMM and DISAV indicators, reaching 0.712, 0.169, 0.119 and 0.037, respectively. The combination of Flow+RGB+Pose achieved the best results in the DIS indicator, with Flow+Pose being only slightly inferior by 0.002. After a comprehensive comparison, we select Flow+Pose (shown in Table 2 Red font) as the feature combination scheme. In the second part, we studied the comparison of synchronization results under different thresholds. We conduct experiments in the threshold range of [45-205]ms and find that NCM is positively correlated with the threshold size, and the larger the threshold, the better the result. In contrast, DIS gradually increase as the threshold expanded. To balance the two, we ultimately choose 125ms (shown in Table 2 Red font) as the synchronization time threshold.

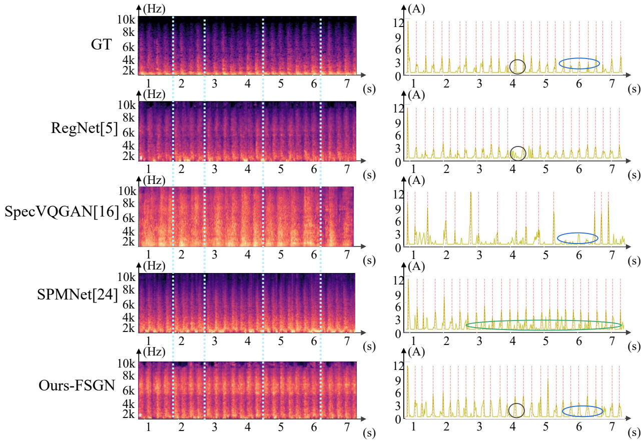


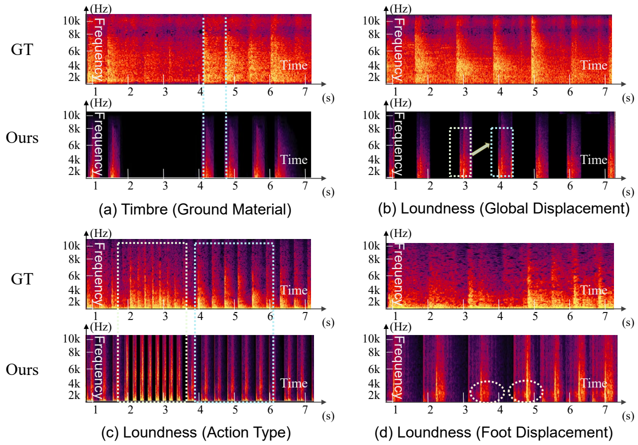
Figure 5: Qualitative evaluation for the results.

Table 2: Impact of different feature combination and different threshold for FSGN model

Experiment	NCM↑	NRM↓	NMM↓	DIS↓	DISAV↓
F	0.637	0.202	0.161	0.621	0.041
F+R	0.676	0.175	0.149	0.669	0.044
F+P	**0.712**	**0.169**	**0.119**	0.566	**0.037**
F+A	0.636	0.206	0.158	0.602	0.040
F+R+P	0.682	0.177	0.141	**0.564**	**0.037**
F+R+A	0.655	0.202	0.142	0.667	0.044
F+P+A	0.664	0.199	0.138	0.578	0.038
F+R+P+A *	0.678	0.181	0.141	0.601	0.040
45(ms)	0.286	0.376	0.339	**0.114**	**0.013**
65	0.429	0.306	0.265	0.261	0.023
85	0.549	0.248	0.203	0.419	0.029
105	0.602	0.222	0.176	0.495	0.032
125	0.712	0.169	0.119	0.566	0.037
145	0.758	0.147	0.095	0.824	0.046
165	0.807	0.124	0.069	0.957	0.055
185	0.832	0.115	0.053	1.117	0.061
205	**0.860**	**0.102**	**0.037**	1.273	0.066

* where R: RGB, F: Flow, P: Pose and A: Attention.

Ablation experiment of FGFSA module. The ablation experiments of the FGFSA module are shown in Figure 6, and (a)-(d) are the four parts with fine-grained tuning of footstep sound. Among them, Figure 6 (a) shows the onset of the sound without significant displacement before and after the timbre adjustment. Figure 6 (b) shows the result of footstep sound adjustment due to the change of visual displacement distance when a person walks forward. In the two boxes before and after the arrow in the figure, the amplitude of the sound increases. Figure 6 (c) shows the result of footstep sound adjustment during the change of walk-run-walk action. The two boxes correspond to two actions, and the sound will change with



(a) Timbre (Ground Material)

(b) Loundness (Global Displacement)

(c) Loundness (Action Type)

(d) Loundness (Foot Displacement)

Figure 6: The result of performance of different components of the FGFSA module. (a) shows the timbre substitution, (b) shows the global displacement-based loudness adjustment, (c) shows the action type-based loudness adjustment and (d) shows the foot displacement-based loudness adjustment. The first row of each subplot is the ground truth and the second row shows the experimental results of our work.

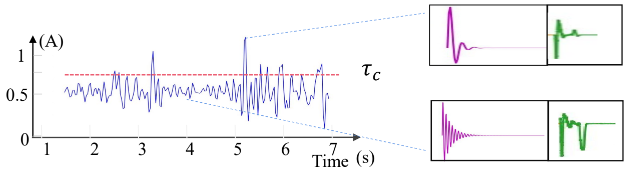


Figure 7: Impact of the threshold value τ_c for FGFSA

the action change. The last, Figure 6 (d) is the result of adjusting the footstep sound attenuation by changing the landing force of a single footstep. It can be found that there is a significant difference in the degree of sound attenuation between the two circles before and after.

Figure 7 shows the effect of the threshold setting on the audio. The blue curve in the left image is the normalized acceleration over time, and the threshold is estimated based on the waveform to distinguish the parts of the acceleration with significant differences. In the right graph, the magenta curve shows the damping decay curve that exceeds or falls within the threshold, and the green curve shows the adjustment result for the corresponding audio.

4.6 User study

To further evaluate the effectiveness of our method, we conduct a user study to qualitatively assess the proposed method. In the experiment, we compare the generated results of our method with those of the data-driven method and the real recorded videos in the dataset. Specifically, twenty people are involved in our user study, whose ages are around 20. For each test video, we show participants the video with audio generated by: RegNet [5], SpecVQGAN [16], SPMNet[24], our method and the real record. In each scenario, participants are asked three questions: “How do you rate the sound quality of the video?”, “How do you rate the audio-visual synchronization of the video?”, and “How do you rate the overall quality of the video?”. Each segment are scored on a scale of 1 to 5, where 1 represents “very poor” and 5 represents “very good”.

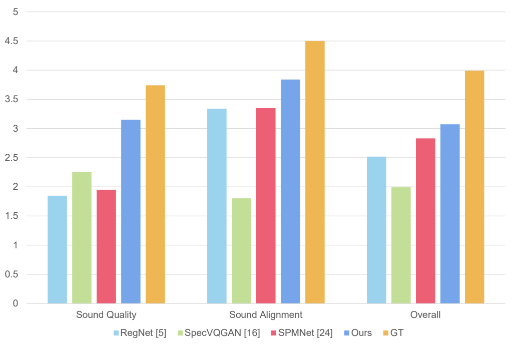


Figure 8: User study results

The user study results are shown in Figure 8, which demonstrate that our method produces results well in terms of sound quality and audio-visual synchronization. Although there is still some gap between our method and real recordings, it performs better compared to existing methods, which proves the effectiveness of our designed algorithm.

5 CONCLUSIONS

In this paper, we propose a novel RD-FGFS method, which is the first step to integrate data-driven and rule modeling approaches for visually aligned footstep sound synthesis. Among them, the learning-based FSGN module is able to achieve the temporal synchronization performance between vision and sound. To further improve synchronisation performance, novel combinations of visual features have been designed. The rule-based FGFSA module constructs a complex mapping relationship between visual motion states, spatial locations and footstep sounds. It realizes the parametric control of visual information on the footstep sound loudness and timbre. In addition, an audio-visual dataset dedicated to footstep is constructed for better fine-grained footstep sound generation. The quantitative and qualitative comparison results show that our proposed method is able to achieve fine-grained footstep sounds.

Though our RD-FGFS method is capable of visually controllable sound adjustment, this advantage is limited when visual footstep vocalization is inaccurately localized. Therefore, further exploration of the audio-visual mapping relationship to obtain accurate sound onset is a possible research direction. Moreover, the RD-FGFS method only explores the mapping relationship between displacement, motion, ground material and sound. More detailed exploration of the visual information to synthesize enriched and realistic sound is another possible direction.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China under grant No. 62201524, No. 61971383, No. 62271455.

REFERENCES

[1] Mohammed Habibullah Baig, Jibin Rajan Varghese, and Zhangyang Wang. 2018. MusicMapp: A deep learning based solution for music exploration and visual interaction. In *Proceedings of the ACM International Conference on Multimedia*. 1253–1255.

[2] Marc Cardle, Stephen Brooks, Ziv Bar-Joseph, and Peter Robinson. 2003. Sound-by-numbers: Motion-driven sound synthesis. In *Proceedings of the ACM SIGGRAPH Symposium on Computer Animation*. 349–356.

[3] Kan Chen, Chuanxi Zhang, Chen Fang, Zhaowen Wang, Trung Bui, and Ram Nevatia. 2018. Visually indicated sound generation by perceptually optimized classification. In *Proceedings of the European Conference on Computer Vision Workshops*. 0–0.

[4] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. 2017. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia*. 349–357.

[5] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. 2020. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing* (2020).

[6] Perry R Cook. 1997. Physically informed sonic modeling (phism): Synthesis of percussive sounds. *Computer Music Journal* (1997).

[7] Perry R Cook. 2002. Modeling Bill's gait: Analysis and parametric synthesis of walking sounds. In *Proceedings of the Audio Engineering Society Conference on Virtual, Synthetic, and Entertainment Audio*.

[8] Andy James Farnell and Obiwannabe Uk. 2007. Marching onwards: Procedural synthetic footsteps for video games and animation. In *Proceedings of The Pure Data Convention*.

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 6202–6211.

[10] Federico Fontana and Roberto Bresin. 2003. Physics-based sound synthesis and control: Crushing, walking and running by crumpling sounds. In *Proceedings of the Colloquium on Musical Informatics*. 109–114.

[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 776–780.

[12] Sanchita Ghose and John Jeffrey Prevost. 2020. Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning. *IEEE Transactions on Multimedia* (2020).

[13] Sanchita Ghose and John J Prevost. 2022. Foleygan: Visually guided generative adversarial network-based synchronous sound generation in silent videos. *IEEE Transactions on Multimedia* (2022).

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* (2020).

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* (1997).

[16] Vladimir Iashin and Esa Rahtu. 2021. Taming visually guided sound generation. In *Proceedings of the British Machine Vision Conference*.

[17] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*. 448–456.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134.

[19] Angelika C Kern and Wolfgang Ellermeier. 2020. Audio in VR: Effects of a soundscape and movement-triggered step sounds on presence. *Frontiers in Robotics and AI* (2020).

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5253–5263.

[21] Shiguang Liu, Haonan Cheng, and Yiying Tong. 2019. Physically-based statistical simulation of rain sound. *ACM Transactions on Graphics* (2019).

[22] Shiguang Liu and Si Gao. 2020. Automatic synthesis of explosion sound synchronized with animation. *Virtual Reality* (2020).

[23] Shiguang Liu, Sijia Li, and Haonan Cheng. 2022. Towards an end-to-end visual-to-raw-audio generation with gan. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[24] Xin Ma, Wei Zhong, Long Ye, and Qin Zhang. 2022. Visually aligned sound generation via sound-producing motion parsing. *Neurocomputing* (2022).

[25] Damián Marelli, Mitsuko Aramaki, Richard Kronland-Martinet, and Charles Verron. 2010. Time-frequency synthesis of noisy sounds with narrow spectral components. *IEEE Transactions on Audio, Speech, and Language Processing* (2010).

[26] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2017. SampleRNN: An unconditional end-to-end neural audio generation model. In *Proceedings of the International Conference on Learning Representations*.

[27] Rolf Nordahl, Luca Turchet, and Stefania Serafin. 2011. Sound synthesis and evaluation of interactive footsteps and environmental sounds rendering for virtual reality applications. *IEEE Transactions on Visualization and Computer Graphics* (2011).

[28] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. 2016. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2405–2413.

[29] Leevi Peltola, Cumhur Erkut, Perry R Cook, and Vesa Valimaki. 2007. Synthesis of hand clapping sounds. *IEEE Transactions on Audio, Speech, and Language Processing* (2007).

[30] Eston Schweickart, Doug L James, and Steve Marschner. 2017. Animating elastic rods with sound. *ACM Transactions on Graphics* (2017).

[31] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 4779–4783.

[32] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[33] Auston Sterling and Ming C Lin. 2016. Interactive modal sound synthesis using generalized proportional damping. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. 79–86.

[34] Tapio Takala and James Hahn. 1992. Sound rendering. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*. 211–220.

[35] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*. 6105–6114.

[36] Luca Turchet. 2016. Footstep sounds synthesis: Design, implementation, and evaluation of foot–floor interactions, surface materials, shoe types, and walkers' features. *Applied Acoustics* (2016).

[37] Luca Turchet, Stefania Serafin, Smilen Dimitrov, and Rolf Nordahl. 2010. Conflicting audio-haptic feedback in physically based simulation of walking sounds. In *Proceedings of the Haptic and Audio Interaction Design*. 97–106.

[38] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *Proceedings of the 9th ISCA Speech Synthesis Workshop*. 125–125.

[39] Yujia Wang, Wei Liang, Wanwan Li, Dingzeyu Li, and Lap-Fai Yu. 2020. Scene-aware background music synthesis. In *Proceedings of the ACM International Conference on Multimedia*. 1162–1170.

[40] David Lewis Yewdall. 2012. Foley: The art of footsteps, props, and cloth movement. In *Practical Art of Motion Picture Sound*. Routledge, 402–439.

[41] Zechen Zhang, Nikunj Raghuvanshi, John Snyder, and Steve Marschner. 2019. Acoustic texture rendering for extended sources in complex scenes. *ACM Transactions on Graphics* (2019).

[42] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3550–3558.

.