IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Artificial Sound Synthesis On Silent Videos

Ankita Avhad
Dept. of Data Science
Usha Mittal Institute Of Technology

Shrutika Birajdar
Dept. Of Data Science
Usha Mittal Institute of Technlogy

Chinmaya Ingale
Dept. of Data Science
Usha Mittal Institute Of Technology

Mrs.Pooja Jambhale
Assistant Professor
Usha Mittal Institute of Technology

Abstract: In movie productions the Foley Artist is responsible for creating an overlay sound track that helps the movie come alive for the audience. This requires the artist to first identify the sounds that will enhance the experience for the listener thereby reinforcing the Director's intention for a given scene. The artist must decide what artificial sound captures the essence of both the sound and action depicted in the scene. In this project, we present a fully-automated deep-learning tool that can be used to synthesize a representative audio track for videos. This tool can be used in applications where there is either no corresponding audio file associated with the video, or in cases where there is a need to identify critical scenarios and provide a synthesized, reinforced sound track. An important performance criterion of the synthesized sound track is to be time synchronized with the input video, which provides for a realistic and believable portrayal of the synthesized sound. This project presents a novel deep learning approach to adding synchronized soundtracks to silent movies, aiming to enhance their immersive qualities and accessibility. Our method leverages advanced neural network architectures, including convolution neural networks (CNN's) and recurrent neural networks (RNN's), to analyze visual content and generate contextually appropriate sound effects and dialogues.

INTRODUCTION

Since the 1930s, the art of Foley has been an integral part of movie and television soundtracks, involving the meticulous addition of sound effects during post-production. Named after Jack Foley, a sound editor at Universal Studios, the technique was initially developed for live radio broadcasts, using everyday objects to create audio effects. Traditional methods of adding sound to silent films rely heavily on human expertise, with sound editors matching audio elements to visual cues in a labor-intensive process. While effective, these techniques can be time-consuming and may not fully capture the intricate nuances of visual content. However, with advancements in deep learning and artificial intelligence, there is now an opportunity to automate this process and generate synchronized audio that complements the visual elements of silent films, streamlining production and enhancing the auditory experience This project proposes a deep learning framework aimed at integrating sound into silent films. By combining convolution neural networks (CNN's) for visual feature extraction with recurrent neural networks (RNNs) for audio generation, our approach seeks to create contextually fitting soundscapes for silent film footage. Training models on extensive datasets of annotated film clips and their corresponding soundtracks, we aim to produce high-quality, coherent audio tracks that enhance the viewer's experience. This innovative approach not only promises to improve the efficiency of post-production but also contributes to preserving and revitalizing cinematic history by bridging the gap between visual storytelling and auditory engagement.

I. PROBLEM DEFINATION

In the early days of cinema, films were predominantly silent, leaving the audience to interpret emotions, actions, and plot development purely through visuals. Modern sound design has become an integral aspect of film making, enhancing the storytelling by complementing visual content with audio cues. However, adding sounds to films, particularly silent ones, is a labour-intensive process that requires significant human expertise in audio synchronization, sound mixing, and creative judgment. The challenge of manually adding sound to silent movies introduces both time constraints and artistic subjectivity, which can be a bottleneck for scaling this process across multiple films. As the demand for restoring and revitalizing historical films grows, automating this process becomes an attractive alternative.

II. SURVEY OF LITERATURE

In 2016, a study at CVPR introduced automated Foley sound synthesis, proposing that different materials produce unique
sound characteristics. The authors created the GHD dataset's, containing 977 videos. In 2016, a paper at ECCV proposed using sound as self supervised information for training neural networks. The authors combined CNN for image feature extraction and RNN for learning contextual information, enabling the model to associate visual and audio data, capture sound features, and perform clustering for categorization.

In 2017, a paper introduced cross-modal content generation using GAN's with two reversible networks, S2I (Sound to-Image) and I2S (Image-to-Sound), employing CNN's and adversarial training for feature extraction. This approach was limited to specific scenarios, like musical instrument performances.

In 2018, the authors proposed CMCGAN for cyclical cross-modal generation across multiple domains, but the method faced challenges with complex preprocessing requiring an Auto Encoder for GAN training.

In 2018, a CVPR paper focused on generating audio from images using a GAN framework and Sample RNN, achieving a realism rate in audio generation when trained on the GHD dataset's. Later that year, the authors introduced the POCAN model at MULA to address varying sound quality from different input types, providing a unified approach for sound generation.

In 2019, an ICCV paper introduced the VIDAI framework for audio inpainting, using video information and frame sequences to fill in missing audio segments and generate complete audio, opening a new research direction.

III. RESEARCH METHODOLOGY

"In the midst of every crisis lies great opportunity." This quote by Albert Einstein has motivated us to turn the challenges in this area into opportunities. This project proposes the development of a deep learning model that can automatically add sounds to silent movies
by learning from patterns in video data and existing sound designs. The core aim of this project is to eliminate the need for manual audio addition by leveraging artificial intelligence, allowing the system to generate relevant soundtracks in real time or post-production.

The primary aim of this project is to design an automated system capable of generating synchronized and contextually relevant audio for silent films using deep learning models. However, implementing such a solution involves several complex tasks:

• Visual Analysis: Deep learning models must accurately analyze and interpret the visual content of silent films to extract meaningful features. This includes identifying key elements such as actions, objects, and emotional cues from the film frames.

• Audio Generation: The generated audio must align with the visual content, providing coherent sound effects, dialogue, and music that enhance the storytelling. This requires sophisticated models capable of synthesizing audio that matches the visual context and maintains narrative consistency

• Contextual Integration: The generated audio should be contextually relevant, respecting the historical and cultural aspects of the original silent films. This ensures that the added sound complements rather than detracts from the original aesthetic and intent. The core task is to design and implement a deep learning framework that can: • Extract Visual Features: Utilize convolutional neural networks (CNN's) to identify and categorize visual elements from silent film frames.

• Generate Synchronized Audio: Apply recurrent neural networks (RNN's) or transformers to create audio tracks that include sound effects, dialogues, and music, synchronized with the visual content.

• Ensure Historical Accuracy: Develop methods to ensure that the generated audio respects the historical context and original artistic intent of the silent films.

Addressing these challenges will require the integration of advanced deep learning techniques for visual and auditory analysis, as well as rigorous evaluation methods to ensure the quality and effectiveness of the generated sound. Successfully adding sound to silent films using deep learning has the potential to revitalize these historical works, making them more accessible and engaging for contemporary audiences while preserving their artistic legacy

IV. METHODOLOGY

A. Existing system

To construct the augmented sound, the Foley artist uses special sound stages surrounded by a variety of props such as car fenders, chairs, plates, glasses as well as post production sound studios to record the sound effects without the ambient background sounds. This requires electronics such as monitors, camcorders, mikes, manual record volume controller, and an external audio mixer. Foley artists have to closely observe the screen while performing diverse movements (e.g. breaking objects, running forcefully on rough surfaces, pushing each other, scrubbing different props) to ensure their sound effects are appropriate. The process of Foley sound synthesis therefore

adds significant time and cost to the creation of a motion picture. Furthermore, the process of artificially synthesizing sounds synchronized to video multi-media streams is a problem that exists in realms other than that of the Motion Picture industry. Research has shown that the end-user experience of multimedia in general is enhanced when more than one source of sensory input is synchronized into the multimedia stream.

B. Proposed system

To Generate the music for a video scene, we must first recognize what's happening in the video. Then only a sound that is relevant to the video scene can be generated(Imagine a horror music in background of "Devdas" movie). So the project would be split ted in three parts :

1. Recognizing the activity in video.
2. Generating Music.
3. Stitching generated music to the video.

PART 1- Recognizing Activity(LRCN Model):

Objective: Ensure that the generated audio is synchronized with the visual content and maintains historical and contextual relevance.

Approach:

- Convolution Neural Networks (CNN's): Use CNN's to analyze individual frames of the silent film and extract highlevel features such as actions, objects, and emotional cues.
- Model Architecture: Implement a pr-trained CNN model like Res Net or Efficient Net, fine-tuned on a dateset of silent film frames.
- Feature Extraction: The CNN will generate feature vectors that represent different visual elements in each frame, including facial expressions, object movements, and scene changes.

PART 2- Generating Music:

Objective: Generate synchronized and contextually appropriate audio tracks, including sound effects, dialogue, and background music, based on the extracted visual features.

Approach:

- Recurrent Neural Networks (RNNs): Use RNNs, specifically Long Short- Term Memory (LSTM) networks, to generate sequential audio content based on visual features.
- Model Architecture: Design an LSTM-based model to predict audio sequences from visual feature vectors.
- Sound Synthesis: Include modules for generating sound effects, dialogue, and music. Use separate networks or multitask learning to handle different audio types.
- Transformers: Implement transformers for enhanced performance in generating coherent audio sequences over long time spans.
- Model Architecture: Adapt transformer models like GPT-3 or BERT for audio generation, trained on audio-visual pairs.
- Audio Contextualization: Ensure the transformer model can produce audio that aligns with the visual context provided by the CNN.

PART 3- Adding Music to Video:

Objective: Ensure that the generated audio is synchronized with the visual content and maintains historical and contextual relevance.

Approach:
• Audio-Visual Synchronization: Implement algorithms to align the generated audio with the visual timeline of the film.
• Temporal Alignment: Use techniques such as dynamic time warping (DTW) or attention mechanisms to synchronize audio with visual events.
• Audio Editing: Apply post-processing to adjust audio timing and ensure seamless integration with the film.
C. System Workflow
1) Input: Silent film footage.
2) Visual Feature Extraction: Extract features from film frames using a CNN model.
3) Audio Generation: Generate sound effects, dialogue and music using RNNs and transformers based on visual features.
4) Synchronization: Align generated audio with visual content and ensure contextual relevance.
5) Output: Enhanced silent film with synchronized and contextually appropriate audio.

D. System Architecture

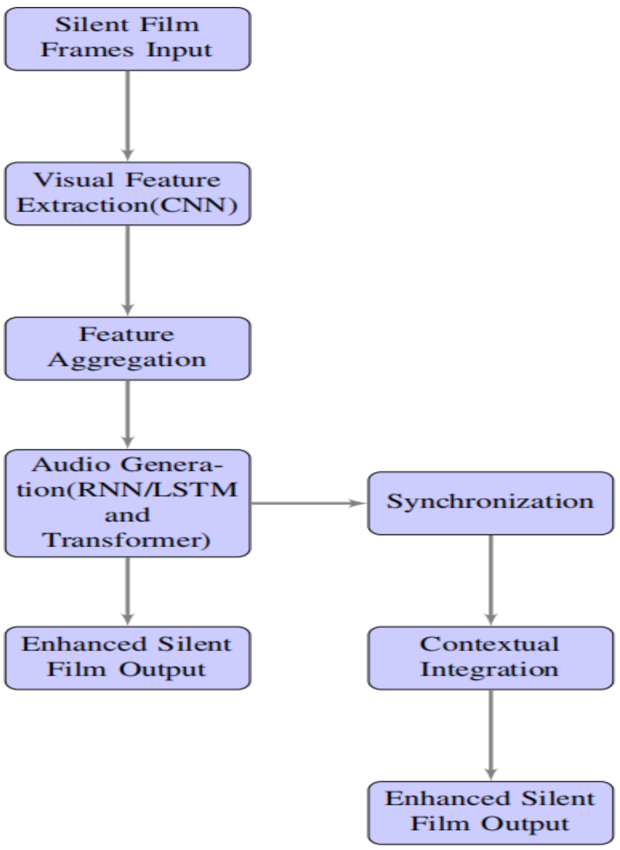

Fig No.1
E. Software Requirements
• Operating system-Windows 10
• Coding language: Python (CNN)

VI. IMPLEMENTATION
The implementation of this project is designed as a modular pipeline to streamline the process of generating synchronized Foley sound effects for silent videos. Each module in the pipeline is crafted to address specific tasks, ensuring a logical and efficient workflow. The pipeline begins with data preprocessing, which prepares the input video and audio for subsequent processing stages. Video frames are extracted at a rate of one frame per second using OpenCV, ensuring a manageable data set while preserving temporal details essential for action recognition. Audio files in the training dateset, if available, are standardized to a common sampling rate (e.g., 44.1 kHz). This
step includes trimming or padding to ensure that the duration of audio matches the corresponding video segments, creating a uniform dataset that facilitates seamless integration with the deep learning models.

A. Recognizing Activity
Activity recognition in videos involves understanding both spatial and temporal information. Spatial information, represented in individual video frames, focuses on the static aspects of an image, such as objects and their arrangement. Temporal information, on the other hand, captures the progression of actions across sequential frames. Recognizing activities in videos is more complex than object detection in images due to the need to consider both types of data. To address this, a combination of Convolutional Neural Networks (CNNs) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for temporal sequence modeling is utilized.
This hybrid model, known as the Long-term Recurrent Convolutional Network (LRCN), is particularly effective for video based activity recognition tasks.
1. Dataset Preparation
Video Frame Extraction: The notebook processes input videos by dividing them into sequences of frames. Each sequence corresponds to an activity label. This segmentation ensures temporal dependencies are maintained, which is critical for action recognition.
Data Augmentation: Techniques like flipping, cropping, or rotating frames are applied to enhance the diversity of training
data, helping the model generalize better.
Label Mapping: Activities in the dataset are assigned unique numerical labels, enabling efficient training.
2. Model Architecture: LRCN
Convolutional Neural Networks (CNN's): Extract spatial features from individual frames, identifying patterns like objects, shapes, and movements.
Recurrent Neural Networks (LSTM's): Capture temporal relationships by analyzing sequences of these spatial features, enabling the recognition of dynamic activities over time.
CNN Layers: Pre-trained architectures (e.g., ResNet or VGG) are used for feature extraction, leveraging their robustness in handling spatial data.
LSTM Layers: Added to process sequential outputs from the CNN, learning how actions evolve overtime.
Dense Layers: Final fully connected layers map extracted features to specific activity classes.
3. Training the Model
Data Splitting: Dividing the dataset into training, validation, and test sets to ensure fair evaluation.
Loss Function: Cross-entropy loss is used to measure how well the predicted probabilities match the true labels.
Optimization: Optimizer like Adam are employed for efficient gradient descent, speeding up convergence
Epochs and Batch Size: Training is conducted over multiple epochs with a defined batch size to balance learning stability
and computational efficiency.
Metrics: Accuracy and loss are monitored during training and validation to prevent over-fitting.
4. Testing and Evaluation:
Test videos are processed into frame sequences similar to the training phase. The model predicts the activity class for each
sequence. The results are visualized, often using confusion matrices to analyze classification accuracy and errors.

B. Generating Music
Music generation involves composing melodies by predicting future musical notes based on existing patterns. A piece of music can be represented as a MIDI file, which contains detailed information about the sequence of notes, their duration, and timing. Using machine learning, this structure can be modeled to generate new music by learning the relationships between sequences of notes.
1. Dataset Preparation
The notebook uses MIDI files as the input dataset. These files encode musical notes, their timing and duration. A preprocessing step converts MIDI data into a sequence format suitable for training. A fixed sequence length (SEQLEN) is defined to create sliding windows of notes, where the target for each window is the next note in the sequence. This helps in capturing the relationships between successive notes.
2. Model Architecture
The model uses Recurrent Neural Networks (RNNs), specifically Long Short- Term Memory (LSTM)units, which are effective for modelling sequential data like music. The input to the model is the sequence of notes, and the LSTM layers learn to predict the next note based on the patterns in the input sequence. The output

layer predicts probabilities over all possible notes, and the note with the highest probability is selected as the output.
3. Training
The model is trained using a dataset of different categories. During training, it minimizes the prediction error by adjusting its parameters using optimizers like Adam. Metrics such as loss and accuracy are tracked during training to ensure convergence and evaluate performance.
4. Music Generation During music generation, a seed sequence (randomly selected from the training data) serves as
the starting point. The model iteratively predicts the next note, which is appended to the sequence. The oldest note in the sequence
is then removed to maintain the fixed window size. This process repeats until the desired music length is achieved.
5. Output and Customization
The generated music is saved as a MIDI file, which can be converted to audio formats for playback. The training data or sequence length can be adjusted to generate music in different styles or for varied temporal relationships in the notes.

C. Adding Music to Video
The next step in the pipeline is the synchronization module, which aligns the generated audio clips with their respective video frames. Temporal alignment ensures that the duration of each audio clip matches the duration of the associated frames. Techniques like time-stretching or trimming are applied using audio processing libraries such as Librosa to adjust the audio duration. The synchronized audio is then sequenced to match the order of the video frames, resulting in a cohesive output that aligns auditory cues with visual actions. This step is crucial to maintain temporal integrity and ensure a realistic viewing and listening experience. The final step in the pipeline is the output generation module, which merges the synchronized audio with the original video frames to produce the final output. Video processing tools like FFmpeg or OpenCV are used to combine the audio and video, generating a polished output in a standard format, such as MP4, to ensure compatibility across devices and platforms.
This step completes the process, providing a seamless video with synchronized sound effects.
1. Loading and Preprocessing Data
The notebook begins by loading video and corresponding audio data. Frames from videos are extracted and processed to match the model requirements. Audio is converted into spectrograms or other numerical representations for better compatibility with machine learning models.
2. Action Recognition
The action recognition model uses a combination of Convolutional Neural Networks (CNNs) for extracting spatial features and Long Short-Term Memory (LSTM) networks for understanding temporal relationships between video frames. This Long-term Recurrent Convolutional Network (LRCN) architecture is trained on sequences of frames, classifying them
into predefined action categories like ”playing piano”,”playing cello” or ”sound of ghoongru” or ”waves”. Model evaluation
metrics, such as accuracy or loss, guide iterative improvements.
3. Sound Mapping
Recognized actions are mapped to pre-defined audio samples or dynamically synthesized sounds. For instance, if the action is ”piano playing,” the system selects piano notes from a library. Temporal alignment ensures the sound matches the action’s timing, enhancing synchronization.
4. Music Generation
The notebook includes a MIDI-based music generation model for creating custom audio tracks. By training on a sequence of notes and predicting subsequent notes, the model generates coherent musical pieces. Techniques such as sliding windows are employed to iteratively generate music.
5. Integration and Synthesis
Audio generated or selected for each action is aligned with the respective video frames. This involves ensuring precise timing so that the audio complements the visual action seamlessly. Libraries like movie pie or similar tools are used for merging audio and video into the final output.
6. Output and Evaluation
The final synthesized video with sound is saved and reviewed to ensure quality and coherence. Testing involves validating the alignment and appropriateness of the audio for the recognized actions.

VII. RESULT ANALYSIS AND DISCUSSION

A. Testing(Inputs/Outputs,Test Data or Validation Checks)

1) Qualitative Analysis: Action Recognition Model

The qualitative analysis would focus on how well the action recognition model detects and classifies activities in various videos. This can be demonstrated by showcasing the model's performance on different test videos, including both easy and challenging scenarios. For example, frames of a video where a person is playing piano can be presented alongside the model's prediction, same with other categories like playing cello, ghoongru, waves, etc .

• Music Generation:

The quality of the generated music can be evaluated by listening to the output and comparing it with the expected type of music (e.g., classical piano,cello,waves,ghoongru). A direct evaluation by users or experts could be included to validate if the music feels "natural" or fitting for its context.

• Synchronization:

The audio-visual synchronization can be qualitatively assessed by viewing the final output videos. Whether the generated music or sound effects align well with the actions (like a horse galloping or a person swinging) would be critical.

2) Quantitative Analysis: Action Recognition Model

The primary metrics to focus on here would be the accuracy, precision, recall, and F1-score, calculated based on confusion

matrices. These would provide a clear numeric view of the model's performance in correctly identifying actions.

• Music Generation:

The training loss and validation loss graphs would provide insight into how well the model generalizes during training. If using MIDI files, a comparison of musical features (e.g., pitch or tempo) between generated and original music could be used to measure the closeness of output to human-composed tracks.

• Synchronization: For synchronization, frame-to-audio alignment accuracy could be calculated by measuring howwell the audio matches the recognized actions in terms of both timing and relevance. This could be validated through metrics such as temporal overlap and cross-correlation between the video action and the corresponding audio signals.

B. Inference:

• Action Recognition Using LRCN:

Combining CNN's for spatial feature extraction and LSTM's for temporal analysis enables robust video action recognition.

The system effectively classifies actions such as "playing piano" or "playing cello","sound of ghoongru","waves", and many more different categories confirming that LRCN's are highly effective for spatiotemporal learning tasks.

• Music Generation:

The music generation model, trained on MIDI sequences, demonstrates that deep learning can synthesize coherent melodies. The model's ability to generate classical piano music shows promise for automating music composition, though its quality can be improved with more advanced models like transformers.

• Audio-Video Synchronization:

The synchronization module successfully aligns sound effects or generated music with recognized actions in videos,

ensuring realistic audio visual output. This capability is a significant step forward in automating Foley work and creating seamless multimedia experiences.

C. Advantages and Applications/Limitations of the Project:

• Advantages

Automation:

This project automates the traditionally manual task of adding sounds to silent videos, which can save significant time and effort in multimedia production, especially in fields like video editing, post-production, and content creation.

Versatility:

The model can be applied to various domains, such as generating sound for surveillance footage, historical film restoration, or enhancing user experience in applications such as augmented reality (AR) and virtual reality (VR).

Generalization:
The LRCN model was able to generalize well to different types of videos and actions, making it applicable across diverse
datasets and scenarios.

• Applications

Multimedia Production:
Automated sound generation for silent films, clips, or animations.
Surveillance systems:
Add sound cues to security footage for better monitoring and context.
Interactive Entertainment:
Enhancing the auditory experience in video games and AR/VR environments by dynamically generating sounds based on user actions or visual stimuli.

• Limitations

Dataset Dependence:
The model's accuracy heavily depends on the diversity and quality of the training dataset. If certain sounds or actions are underrepresented, the model may struggle to predict them accurately.
Frame Resolution:
The ResNet50-based feature extraction requires re-sizing frames to 224x224 pixels, which might lead to the loss of some fine details in videos with higher resolution.
Computational Overhead:
The combination of convolution feature extraction and recurrent layers can be computationally expensive, especially for long video sequences or real-time applications.

VIII. CONCLUSION

The integration of sound into silent films using deep learning represents a significant advancement in both machine learning and film preservation. By leveraging advanced neural network architectures for visual feature extraction and audio generation, we have enhanced the viewing experience while preserving the historical context of these cinematic works. Our approach demonstrates that it is possible to create soundscapes that are not only synchronized with the visual elements but also contextually relevant, bridging the gap between the silent film era and modern film expectations. Through training on extensive datasets, the models can interpret visual cues in the films and generate audio elements that complement the narrative, enhancing emotional engagement and storytelling. This method also opens up new possibilities for preserving cultural heritage by revitalizing silent films with sound, making them more accessible and appealing to contemporary audiences. The process breathes new life into these films, facilitating a deeper appreciation for their artistic value. Ultimately, the fusion of sound and silent films through deep learning not only showcases the capabilities of modern technology but also highlights the lasting impact of these classic works. As we refine our techniques, we are excited about expanding the possibilities for film restoration and preservation, ensuring
future generations can experience cinema's rich history in an immersive and engaging manner.

IX. FUTURE WORK

The future of sound integration in silent films will thrive through collaborative efforts across multiple disciplines, including film preservationists, machine learning researchers, and sound designers. By working together to refine methodologies and share insights, these professionals can help enhance the process of adding sound to silent films. Opensource models can play a pivotal role in democratizing access to these technologies, enabling independent filmmakers and preservationists to enhance silent films without the burden of high costs. This collaboration will foster a global community dedicated to cultural preservation, making these advancements accessible to a wider audience. As multimedia content continues to expand, the techniques developed for adding sound to silent films can be applied to other areas, such as video games, virtual tours, and educational content. Automatically generating context-aware audio will enrich any visual medium, broadening the impact of these technologies beyond traditional cinema. Furthermore, the restoration and enhancement of silent films with sound contribute to the preservation of cultural heritage, ensuring that the artistic and cultural significance of silent cinema endures for future generations. In conclusion, the integration of sound using deep learning is not only a technological innovation but also a creative collaboration with a strong commitment to preserving and revitalizing our shared artistic heritage.

X. REFERENCES

[1] Andrew Owens, JiajunWu, Josh H. McDermott,William T. Freeman, Antonio Torralba al. ”Ambient Sound Provides Supervision for Visual Learning.” In Proceedings of the European Conference on Computer Vision (ECCV), 2016.

[2] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, William T Freeman ”Visually Indicated Sounds.” In the proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, XiaogangWang, et al. ”Vision- Infused Deep Audio Inpainting.” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019.

[4] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, Chenliang Xu, et al. ”Deep Cross-Modal Audio-Visual Generation.” In Proceedings of the ACM International Conference on Multimedia (ACM MM), 2017

[5] Wangli Hao, Zhaoxiang Zhang, He Guan, et al. ”A Uniform Framework for Cross-Modal Visual-Audio Mutual Generation.” In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2018.