Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment

Kim Sung-Bin
POSTECH
sungbin@postech.ac.kr

Arda Senocak
KAIST
arda.senocak@gmail.com

Hyunwoo Ha
POSTECH
hyunwooha@postech.ac.kr

Andrew Owens
University of Michigan
ahowens@umich.edu

Tae-Hyun Oh
POSTECH
taehyun@postech.ac.kr

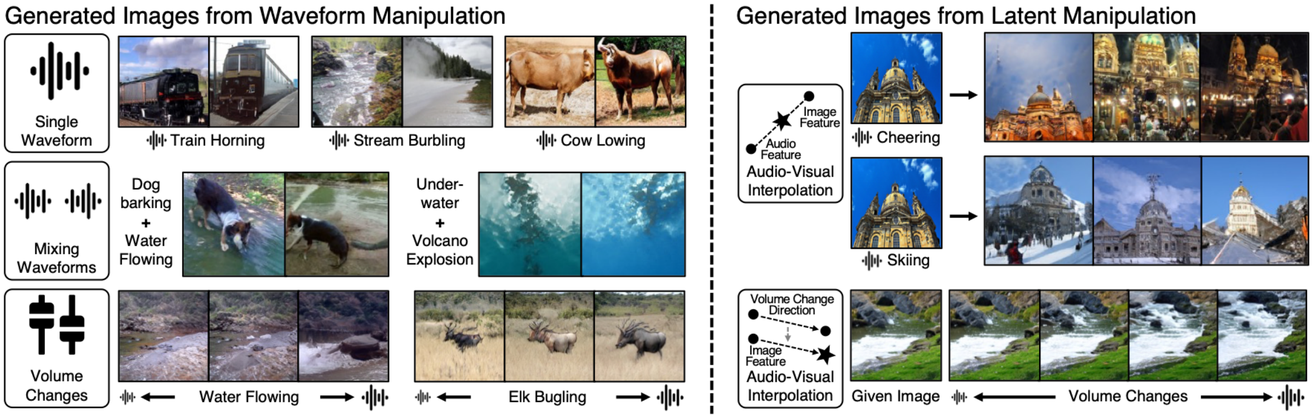


Figure 1: **Sound-to-image generation.** We propose a model that synthesizes images of natural scenes from the sound. Our model is trained solely from paired audio-visual data, without labels or language supervision. Our model's predictions can be controlled by applying simple manipulations to the input waveforms (left), such as by mixing two sounds together or by adjusting the volume. We can also control our model's outputs in latent space, such as by interpolating in directions specified by sound (right).

Abstract

How does audio describe the world around us? In this paper, we propose a method for generating an image of a scene from sound. Our method addresses the challenges of dealing with the large gaps that often exist between sight and sound. We design a model that works by scheduling the learning procedure of each model component to associate audio-visual modalities despite their information gaps. The key idea is to enrich the audio features with visual information by learning to align audio to visual latent space. We translate the input audio to visual features, then use a pre-trained generator to produce an image. To further improve the quality of our generated images, we use sound source localization to select the audio-visual pairs that have strong cross-modal correlations. We obtain substantially better results on the VEGAS and VGGSound datasets than prior approaches. We also show that we can control our model's predictions by applying simple manipulations to the input waveform, or to the latent space.

1. Introduction

Humans have the remarkable ability to associate sounds with visual scenes, such as how chirping birds and rustling branches bring to mind a lush forest, and the flowing water conjures the image of a river. These cross-modal associations convey important information, such as the distance and size of sound sources, and the presence of out-of-sight objects.

In this work, we propose Sound2Scene, a sound-to-image generative model and training procedure, which can be trained solely from unlabeled videos. First, given an image encoder pre-trained in a self-supervised way, we train a conditional generative adversarial network [4] to generate images from the visual features of the image encoder. We

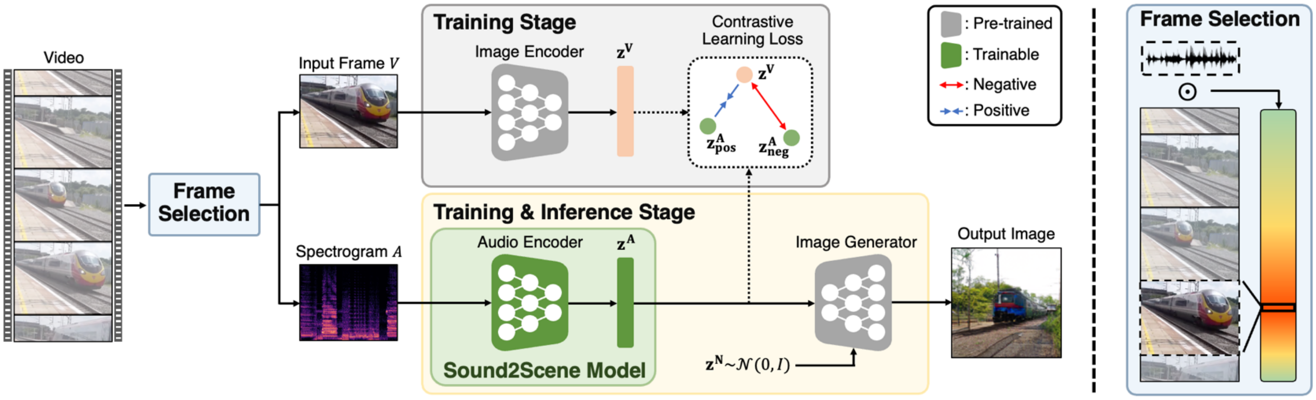

Video
Frame Selection
Input Frame V
Spectrogram A
Training Stage
Image Encoder
$\mathbf{z}^{\mathbf{V}}$
Contrastive Learning Loss
$\mathbf{z}^{\mathbf{V}}$
$\mathbf{z}^{\mathbf{A}}_{\mathbf{pos}}$
$\mathbf{z}^{\mathbf{A}}_{\mathbf{neg}}$
Training & Inference Stage
Audio Encoder
$\mathbf{z}^{\mathbf{A}}$
Sound2Scene Model
$\mathbf{z}^{\mathbf{N}} \sim \mathcal{N}(0, I)$
Image Generator
Output Image
: Pre-trained
: Trainable
: Negative
: Positive
Frame Selection
$\odot$

Figure 2: **Sound2Scene framework.** The frame selection method selects the highly correlated frame-audio segment from a video for training. Then, we train Sound2Scene to produce an audio feature that aligns with the visual feature extracted from the pre-trained image encoder. In the inference stage, the extracted audio feature from input audio is fed to the image generator to produce an image.

then train an audio encoder to translate an input sound to its corresponding visual feature, by aligning the audio to the visual space. Afterwards, we can generate diverse images from sound by translating from audio to visual embeddings and synthesizing an image. Since our model must be capable of learning from challenging in-the-wild videos, we use sound source localization to select moments in time that have strong cross-modal associations.

We evaluate our model on VEGAS [22] and VGGSound [6], as shown in Fig. 1. Our model can synthesize a wide variety of different scenes from sound in high quality, outperforming the prior arts. It also provides an intuitive way to control the image generation process by applying manipulations at both the input and latent space levels, such as by mixing multiple audios together or adjusting the loudness. Our main contributions are summarized as follows:

- Proposing a new sound-to-image generation method that can generate visually rich images from in-the-wild audio in a self-supervised way.
- Generating high-quality images from the unrestricted diverse categories of input sounds for the first time.
- Demonstrating that the samples generated by our model can be controlled by intuitive manipulations in the waveform space in addition to latent space.
- Showing the effectiveness of training sound-to-image generation using highly correlated audio-visual pairs.

2. Method

The goal of our work is to learn to translate sounds into visual scenes. Most of the existing methods [20, 7, 10, 12] train GANs to directly generate images from the raw sound or sound features. However, the large variability of visual scenes make the task of directly predicting images from sound challenging.

In contrast to prior approaches, we sidestep these challenges by breaking down the task into sub-problems. Our proposed Sound2Scene pipeline is illustrated in Fig. 2. It is composed of three parts: an audio encoder, an image encoder, and an image generator. First, we pre-train a powerful image encoder and a generator conditioned by the encoder, separately with a large image dataset alone. Since there is a natural correspondence between sound and visual information, we exploit this natural alignment and transfer the discriminative and expressive visual information from the image encoder into audio representation. In this way, we construct a joint audio-visual embedding space that is trained in a self-supervised manner using only in-the-wild videos. Later, the audio representation from this aligned embedding space is fed into the image generator to produce images corresponding to the input sound.

2.1. Learning the Sound2Scene Model

Using the audio-visual data pairs $\mathcal{D} = \{V_i, A_i\}_{i=1}^{N}$, where V_i is a video frame, and A_i is audio, our objective is to learn the audio encoder to extract informative audio features $\mathbf{z}^{\mathbf{A}}$ that are aligned well with anchored visual features $\mathbf{z}^{\mathbf{V}}$. Specifically, given the unlabeled data pairs $\mathcal{D}$, the audio encoder $f_A(\cdot)$, and the image encoder $f_V(\cdot)$, we extract audio features $\mathbf{z}^{\mathbf{A}}{=}f_A(A)$ and visual features $\mathbf{z}^{\mathbf{V}}{=}f_V(V)$, where $\mathbf{z}^{\mathbf{V}}, \mathbf{z}^{\mathbf{A}} \in R^{2048}$. Since we exploit the well pre-trained image encoder $f_V(\cdot)$, the visual feature $\mathbf{z}^{\mathbf{V}}$ serves as the self-supervision signal for the audio encoder to predict the informative feature $\mathbf{z}^{\mathbf{A}}$ in the way of feature-based knowledge distillation [14, 11]. These aligned features across modalities construct the shared audio-visual embedding space on which the image generator $G(\cdot)$ is separately trained compatibly.

To align the embedding spaces defined by the heterogeneous modalities, a metric learning approach can be used. Representations are aligned if they are close to each other under some distance metric. A simple approach to align the features of $\mathbf{z^A}$ and $\mathbf{z^V}$ is to minimize the L_2 distance, $\|\mathbf{z^V} - \mathbf{z^A}\|_2$. However, we discover that solely using L_2 loss can only teach the relationship between two different modalities within the pair without considering the other unpaired samples. This results in unstable training and leads to poor image quality. Therefore, we use InfoNCE [16] as a specific type of contrastive learning, which has been successfully applied to audio-visual representation learnings [1, 5, 18, 8, 21, 15]:

$$\texttt{InfoNCE}(\mathbf{a}_j, \{\mathbf{b}\}_{k=1}^{N}) = -\log \frac{\exp(-d(\mathbf{a}_j,\mathbf{b}_j))}{\sum_{k=1}^{N}\exp(-d(\mathbf{a}_j,\mathbf{b}_k))}, \quad (1)$$

where $\mathbf{a}$ and $\mathbf{b}$ denotes arbitrary vectors with the same dimension, and $d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2$. With this loss, we maximize the feature similarity between an image and its true audio segment (positive) while minimizing the similarity with the randomly selected unrelated audios (negatives). Given the j-th visual and audio feature pair, we first define our audio feature-centric loss as $L_j^A = \texttt{InfoNCE}(\mathbf{\hat{z}_j^A}, \{\mathbf{\hat{z}^V}\})$, where $\mathbf{\hat{z}^A}$ and $\mathbf{\hat{z}^V}$ are representations with unit-norm. To make our objective symmetric, we compute the visual feature-centric loss term as $L_j^V = \texttt{InfoNCE}(\mathbf{\hat{z}_j^V}, \{\mathbf{\hat{z}^A}\})$. Then, our final learning objective is to minimize the sum of each loss term for all the audio and visual pairs in the mini-batch B:

$$L_{total} = \frac{1}{2B}\sum_{j=1}^{B}\left(L_j^A + L_j^V\right). \quad (2)$$

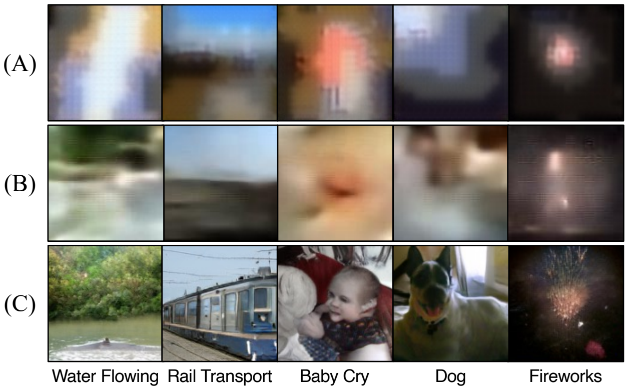
After training the audio encoder with Eq. (2), our model learns visually enriched audio features that are aligned with the visual features. Thus, we can directly feed the learned audio feature $\mathbf{z^A}$ with noise vector $\mathbf{z^N} \sim \mathcal{N}(0, I)$ to the frozen image generator as $G(\mathbf{z^N}, \mathbf{z^A})$ to generate a visual scene at the inference stage.

2.2. Architecture

All the following modules are separately trained according to the proposed steps.

Image encoder $f_V(\cdot)$**.** We use ResNet-50 [13]. To cope with general visual contents, we train the image encoder in a self-supervised way [3] with ImageNet [9] without labels.

Image generator $G(\cdot)$**.** We use the BigGAN [2] architecture to deal with high-quality generation and a large variability of scene contents. To make the BigGAN a conditional generator, we follow the modification of the input condition structure of ICGAN [4]. We train the generator to generate photo-realistic 128×128 resolution images from the conditional visual embeddings $\mathbf{z^V}$ obtained from the image encoder. To train the generator, we use ImageNet without labels in a self-supervised way. While training the image generator, the image encoder is pre-trained and fixed.



	Method	VEGAS (5 classes)		
		R@1	FID (↓)	IS (↑)
(A)	Pedersoli *et al.* [17]	23.10	118.68	1.19
(B)	S2I [10]	39.19	114.84	1.45
(C)	Ours	**77.58**	**34.68**	**4.01**

Figure 3: **Comparison to the baseline** [17] **and existing sound-to-image method** [10]. Our method outperforms the others both qualitatively and quantitatively in the VEGAS dataset.

Audio encoder $f_A(\cdot)$**.** We use ResNet-18, which takes the spectrogram as input. After the last convolutional layer, adaptive average pooling aggregates temporal-frequency information into a single vector. The pooled feature is fed into a single linear layer to obtain an audio embedding $\mathbf{z^A}$. The audio network is trained on either VGGSound or VEGAS with the loss in Eq. (2) according to target benchmarks.

2.3. Audio-Visual Pair Selection Module

Learning the relationship between the images and sounds accurately requires highly correlated data pairs of two modalities. Knowing which frame/segment in the video is informative for audio-visual correspondence is not an easy task. One straightforward way to collect data pairs for training, $\mathcal{D}$, is to extract a mid-frame of the video with the corresponding audio segment [15, 5]. However, the mid-frame cannot guarantee to contain informative corresponding audio-visual signals [19]. To this end, we leverage a pre-trained sound source localization model [19] and extract highly correlated audio and visual pairs. The backbone networks of [19] enable us to have fine-grained temporal time steps of audio-visual features, $\mathbf{q^A}$ and $\mathbf{q^V}$, respectively. Correlation scores are computed by $\mathbb{C}_{av}[t] = \mathbf{q_t^V} \cdot \mathbf{q_t^A}$ at each time step. After computing the correlation scores, $\mathbb{C}_{av}$ are sorted by $\texttt{top-k}(\mathbb{C}_{av}[t])$. With this correlated pair selection method, we annotate `top-1` moment frames for each video in the training splits and use them for training.

3. Experiment

We validate our method on the VGGSound and VEGAS datasets both qualitatively and quantitatively. Here, we in-

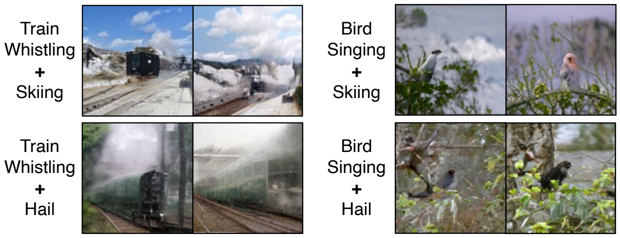


Figure 4: **Generated images by mixing two different audios in the *waveform space*.**

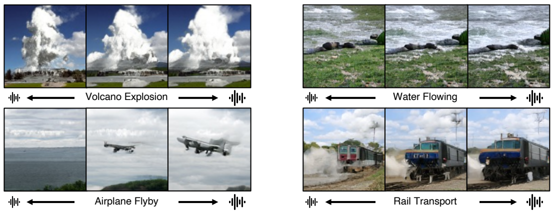
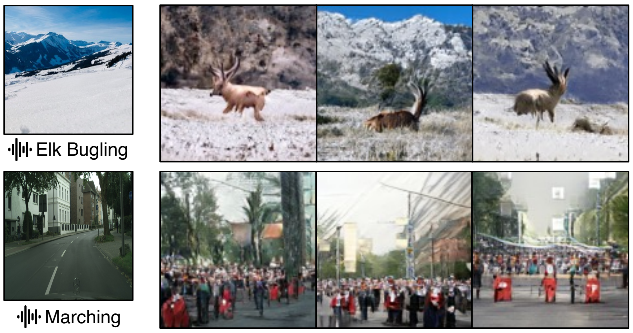


Figure 5: **Generated images by changing the volumes of the input audio in the *waveform space*.** As the volume increases, the objects of the sound source become larger or more dynamic.



(a) Inputs (b) Generated images

Figure 6: **Generated images conditioned on image and audio.** We interpolate between a given visual feature and an audio feature in the *latent space*. This interpolated feature is then fed to the image generator to get a novel image.

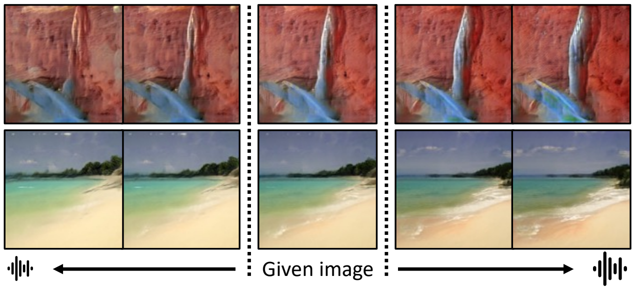


Figure 7: **Image editing by volume changes in *latent space*.** We extract an image feature and noise vector by GAN inversion, and two audio features with different volumes. Then, we move the image feature in the direction of the audio feature differences.

troduce only partial results of our extensive experiments. As shown in Fig. 3, our method outperforms the previous method on Frechet Inception Distance (FID), Inception score (IS), and image-to-text retrieval, denoted as CLIP retrieval (R@k). Furthermore, our method generates more clear and realistic images compared to the prior arts.

We qualitatively observe that Sound2Scene generates visually plausible images compatible with a single input waveform, as shown in Fig. 1. Furthermore, we observe that the model can perform similarly to human perception. We vary the input types to investigate how accurately Sound2Scene can capture the relationships between vision and different characteristics in the sound. Our model can capture a single instance that makes the sound, multiple sound characteristics (Fig. 4), and even the volume changes (Fig. 5) of the same sound by making the dog bigger or by making strong streams while increasing the volume. As for further application, our model can take both image and audio together and generate images conditioned on the composition of multiple modalities (Fig. 6), or edit the given image by changing the volume of the corresponding sound (Fig. 7).

4. Conclusion

In this paper, we propose Sound2Scene, a model for generating images that are relevant to the given audio. This task inherently has challenges: a significant modality gap between audio and visual signals, such that audio lacks visual information, and audio-visual pairs are not always correspondent. Existing approaches have limitations due to these difficulties. We show that our proposed method overcomes these challenges in that it can successfully enrich the audio features with visual knowledge, selects audio-visually correlated pairs for learning, and generates rich images with various characteristics. Furthermore, we demonstrate our model allows controllability in inputs to get more creative results, unlike the prior arts. We would like to note that our proposed learning approach and the audio-visual pair selection method are independent of the specific design choice of the model. We hope that our work encourages further research on multi-modal image generation.

Acknowledgment

This work was supported by IITP grant funded by Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No.RS-2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation).

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision (ECCV)*, 2020.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[5] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

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

[8] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] Leonardo A Fanzeres and Climent Nadeu. Sound-to-imagination: Unsupervised crossmodal translation using deep dense network architecture. *arXiv preprint arXiv:2106.01266*, 2021.

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision (IJCV)*, 2021.

[12] Wangli Hao, Zhaoxiang Zhang, and He Guan. Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[15] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[17] Fabrizio Pedersoli, Dryden Wiebe, Amin Banitalebi, Yong Zhang, and Kwang Moo Yi. Estimating visual information from audio through manifold learning. *arXiv preprint arXiv:2208.02337*, 2022.

[18] Arda Senocak, Junsik Kim, Tae-Hyun Oh, Hyeonggon Ryu, Dingzeyu Li, and In So Kweon. Event-specific audio-visual fusion layers: A simple and new perspective on video understanding. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[19] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[20] Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee. Towards audio to scene image synthesis using generative adversarial network. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[21] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

[22] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.