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

Kim Sung-Bin[1] Arda Senocak[2] Hyunwoo Ha[1] Andrew Owens[3] Tae-Hyun Oh[1,4,5]

[1]Dept. of Electrical Engineering and [4]Grad. School of Artificial Intelligence, POSTECH
[2]Dept. of Electrical Engineering, KAIST [3]University of Michigan
[5]Institute for Convergence Research and Education in Advanced Technology, Yonsei University
https://sound2scene.github.io/

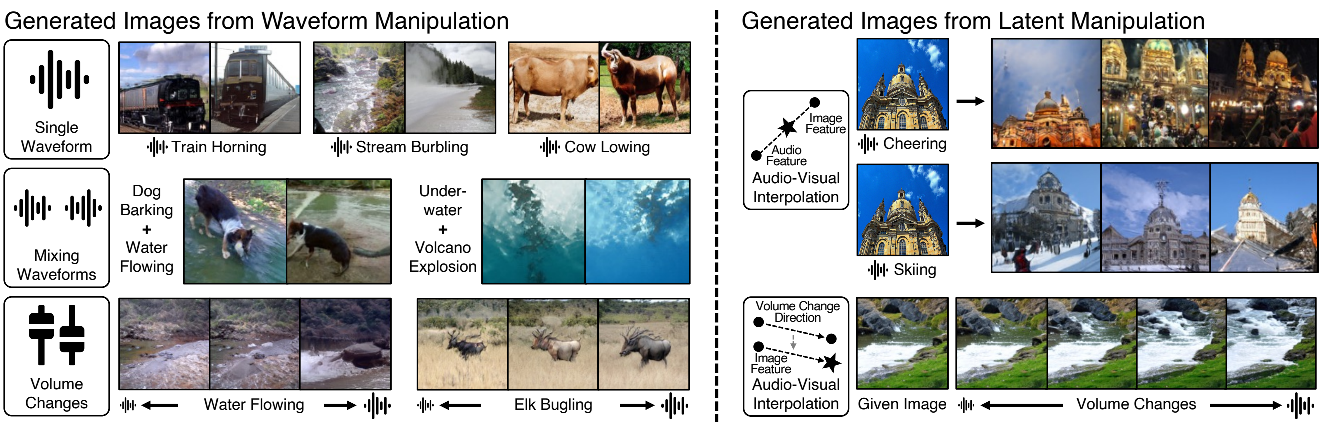


Figure 1. **Sound-to-image generation.** We propose a model that synthesizes images of natural scenes from the sound. Our model is trained solely from paired audio-visual data, without labels or language supervision. Our model's predictions can be controlled by applying simple manipulations to the input 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.

An emerging line of work has sought to create multi-modal learning systems that have these cross-modal prediction capabilities, by synthesizing visual imagery from sound [15, 20, 26, 36, 37, 63, 69]. However, these existing methods come with significant limitations, such as being limited to simple datasets in which images and sounds are closely correlated [63, 69], relying on vision-and-language supervision [36], and being capable only of manipulating

Acknowledgment. This work was supported by IITP grant funded by Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities). The GPU resource was supported by the HPC Support Project, MSIT and NIPA.

the style of existing images [37] but not synthesis.

Addressing these limitations requires handling several challenges. First, there is a significant modality gap between sight and sound, as sound often lacks information that is important for image synthesis, *e.g*., the shape, color, or spatial location of on-screen objects. Second, the correlation between modalities is often incongruent, *e.g*., highly contingent or off-sync on timing. Cows, for example, only rarely moo, so associating images of cows with "moo" sounds requires capturing training examples with the rare moments when on-screen cows vocalize.

In this work, we propose Sound2Scene, a sound-to-image generative model and training procedure that addresses these limitations, and 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 [11] to generate images from the visual features of the image encoder. We 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 [73] and VGGSound [14], 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. Related Work

Audio-visual generation. The audio-visual cross-modal generation field is explored in two directions: vision-to-sound and sound-to-vision generation. The vision-to-sound task has been actively researched in instrument/music [15,26, 65] and open-domain generic audio generation [16,33,46,73] perspectives. In the opposite direction, early work on sound-to-image investigated only restricted and specialized audio domains, such as instruments [12, 15, 26, 42], birds [63] or speech [43]. Later, Wan *et al*. [69] and Fanzeres *et al*. [20] attempt to alleviate the restrictions on the data domain and generate images by conditioning on sounds from nine categories of SoundNet [8] and five categories of VEGAS [73], respectively. Although we have a similar goal of generating images from unrestricted sounds, our approach is capable of handling much more diverse audio-visual generation problems. For example, it is capable of generating images from sounds that come from a variety of categories in the VGGSound [14] and the VEGAS dataset. Unlike the low-quality results of the previous methods, our model generates visually plausible images that are related to the given audio.

Audio-driven image manipulation. Instead of directly generating images from audio, a recent line of work has proposed to edit existing images using sound-based input. Lee *et al*. [36] used the text-based image manipulation model [51] and extend its embedding space to that of audio-visual modality with the text modality. Similarly, Li *et al*. [37] used conditional generative adversarial networks (GANs) [24] to edit the visual style of an image to match a sound, and showed that the manipulations could be controlled by adjusting a sound's volume or by mixing together multiple sounds. Our work differs from them in two ways. First, our model is capable of *generating* images conditioned on sound, rather than only *editing* them. Second, unlike Lee *et al*., we do not require a text-based visual-language embedding space. Instead, our model is trained entirely on unlabeled audio-visual pairs.

Cross-modal generation. Learning to translate one modality to another, *i.e*., cross-modal generation, is an interesting yet open research problem. Various tasks have been tackled in diverse domains, such as text-to-image/video [19, 30, 51, 55, 56, 64, 68], speech-to-face/gesture [23, 43], scene graph/layout-to-image [34, 72], image/audio-to-caption [4, 35, 39], *etc*. For bridging the heterogeneous modalities in cross-modal generations, several works [43, 53] leverage existing pre-trained models or extend pre-trained CLIP [54] embedding space anchored with text-visual modality to suit their purpose [36, 51, 55, 71]. In this trend, we tackle the task of generating images from sound by leveraging freely acquired audio-visual signals from the video.

Audio-visual learning. The natural co-occurrence of audio-visual cues is often leveraged as a self-supervision signal to learn the associations between two modalities and assist each other for learning better representations. The learned representations in such a way are exploited for diverse applications including, cross-modal retrieval [7, 48], video recognition [17, 41], and sound source localization [13, 50, 60–62]. A line of work for constructing an audio-visual embedding space is to jointly train two different neural networks for each modality by judging if the frame and audio correspond to each other [6, 45]. Recent works use clustering [5, 31] or contrastive learning [3, 40, 41] for better learning of the joint

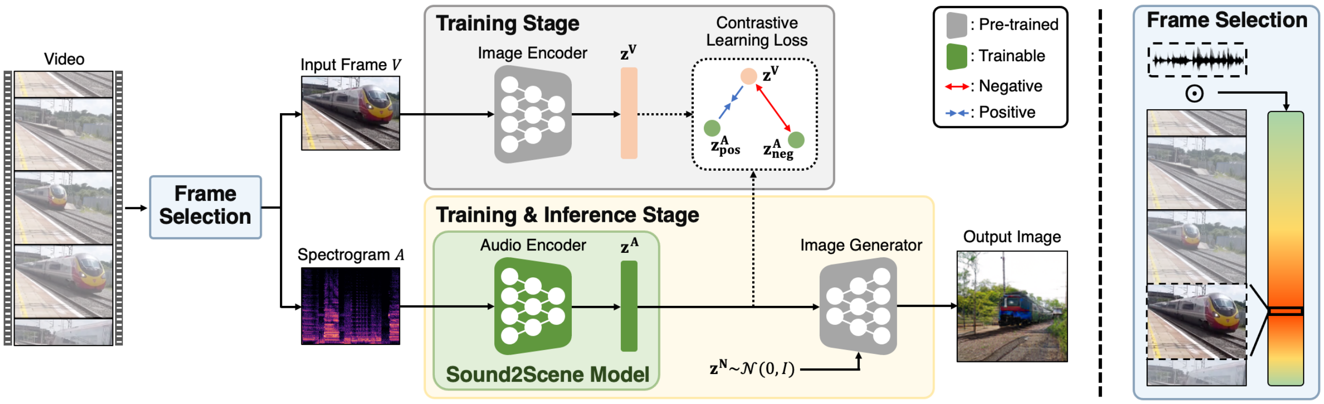


Figure 2. **Sound2Scene framework.** The frame selection method selects the highly correlated frame-audio segment from a video for training. Then, we train Sound2Scene to produce an audio feature that aligns with the visual feature extracted from the pre-trained image encoder. In the inference stage,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

audio-visual embedding space. While the above-mentioned approaches learn the audio-visual representations jointly from scratch, another line of work creates a joint embedding space by exploiting existing knowledge from the expert models. The knowledge can be transferred from the audio to visual representation [47], visual representation to audio [8,21], or distilled from both audio-visual representations to that of video [17]. Our work takes the latter line, assuming a visual expert model exists. We use the image feature extractor to distill rich visual information from large-scale internet videos to the audio modality.

3. Method

The goal of our work is to learn to translate sounds into visual scenes. Most of the existing methods [15, 20, 26, 69] train GANs to directly generate images from the raw sound or sound features. However, the aforementioned challenges and 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.

3.1. 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 $\mathbf{z^A}$ that are aligned well with anchored visual features $\mathbf{z^V}$. 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 $\mathbf{z^A}{=}f_A(A)$ and visual features $\mathbf{z^V}{=}f_V(V)$, where $\mathbf{z^V}, \mathbf{z^A} \in R^{2048}$. Since we exploit the well pre-trained image encoder $f_V(\cdot)$, the visual feature $\mathbf{z^V}$ serves as the self-supervision signal for the audio encoder to predict the informative feature $\mathbf{z^A}$ in the way of feature-based knowledge distillation [25, 29]. 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 [44] as a specific type of contrastive learning, which has been successfully applied to audio-visual representation learnings [2, 13, 17, 36, 59, 70]:

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

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

Figure 3. **Examples of selected `top-1` frame vs. mid-frame.**

j-th visual and audio feature pair, we first define our audio feature-centric loss as $L_j^A = \texttt{InfoNCE}(\hat{\mathbf{z}}_\mathbf{j}^\mathbf{A}, \{\hat{\mathbf{z}}^\mathbf{V}\})$, where $\hat{\mathbf{z}}^\mathbf{A}$ and $\hat{\mathbf{z}}^\mathbf{V}$ are representations with unit-norm. To make our objective symmetric, we compute the visual feature-centric loss term as $L_j^V = \texttt{InfoNCE}(\hat{\mathbf{z}}_\mathbf{j}^\mathbf{V}, \{\hat{\mathbf{z}}^\mathbf{A}\})$. 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:

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

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 $\mathbf{z}^\mathbf{A}$ with noise vector $\mathbf{z}^\mathbf{N} \sim \mathcal{N}(0, I)$ to the frozen image generator as $G(\mathbf{z}^\mathbf{N}, \mathbf{z}^\mathbf{A})$ to generate a visual scene at the inference stage.

3.2. Architecture

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

Image encoder $f_V(\cdot)$. We use ResNet-50 [27]. To cope with general visual contents, we train the image encoder in a self-supervised way [10] with ImageNet [18] without labels.

Image generator $G(\cdot)$. We use the BigGAN [9] 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 [11]. We train the generator to generate photo-realistic 128×128 resolution images from the conditional visual embeddings $\mathbf{z}^\mathbf{V}$ 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.

Audio encoder $f_A(\cdot)$. We use ResNet-18, which takes the audio 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 $\mathbf{z}^\mathbf{A}$. The audio network is trained on either VGGSound or VEGAS with the loss in Eq. (2) according to target benchmarks.

3.3. 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 [13, 36]. However, the mid-frame cannot guarantee to contain informative corresponding audio-visual signals [62]. To this end, we leverage a pre-trained sound source localization model [62] and extract highly correlated audio and visual pairs. The backbone networks of [62] enable us to have fine-grained temporal time steps of audio-visual features, $\mathbf{q}^\mathbf{A}$ and $\mathbf{q}^\mathbf{V}$, respectively. Correlation scores are computed by $\mathbb{C}_{av}[t] = \mathbf{q}_\mathbf{t}^\mathbf{V} \cdot \mathbf{q}_\mathbf{t}^\mathbf{A}$ 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. Fig. 3 shows the comparison between selected frames and mid-frames. Even though frames are selected automatically, they contain the corresponding distinctive objects to the audio accurately.

4. Experiments

We validate our proposed sound-to-image generation method with experiments on VGGSound [14] and VEGAS [73]. First, we visualize samples of generated images on diverse categories of sounds. Then, we quantitatively examine the generation quality, diversity, and correspondence between the audio and generated images. Note that we do not use any class information during training and inference.

4.1. Experiment Setup

Datasets. We train and test our method on VGGSound [14] and VEGAS [73]. VGGSound is an audio-visual dataset containing around 200K videos. We select 50 classes among this dataset and follow the train and test splits provided. VEGAS contains about 2.8K videos with 10 classes. For the data statistic balance, we select 800 videos for training and 50 videos for testing per class. Test splits in both datasets are used for the following qualitative and quantitative analysis.

Evaluation metrics. We demonstrate the objective and subjective metrics to evaluate our method quantitatively.

- **CLIP** [54] **retrieval** : Inspired by the CLIP R-Precision metric [49], we quantify the generated images by measuring image-to-text retrieval performance with recall at K ($R@K$). We feed the generated images and the texts from the name of the audio category to CLIP. Then, we measure the similarities between the image and text features and rank the candidate text descriptions for the query image.
- **Fréchet Inception Distance (FID)** [28] **and Inception Score (IS)** [58] : FID measures the Fréchet distance between the features obtained from real and synthesized images using a pre-trained Inception-V3 [66]. This same model can also be used for measuring inception score (IS), which computes the KL-divergence between the conditional class distribution and the marginal class distribution.

Figure 4. **Qualitative results by feeding single waveform from VGGSound test set.** Sound2Scene generates diverse images in a wide variety of categories from generic sounds as input.

- **Human evaluations** : We recruit 70 participants to analyze the performance of our method from a human perception perspective. We first compare our method with the image-only model [11] and then evaluate whether our model generates proper images corresponding to input sound. More details can be found in Sec. 4.3.

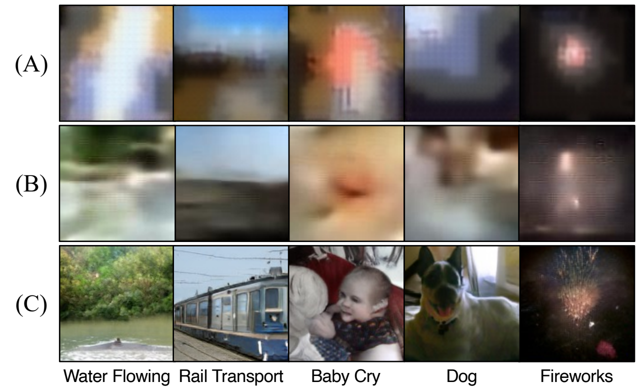
Implementation details. The input of the audio encoder is 1004×257-dimensional log-spectrogram converted from 10 seconds of audio. The extracted frame from the video is resized to 224×224 and fed as an input of the image encoder. We train our model on a single GeForce RTX 3090 for 50 epochs with early stopping. We use the Adam optimizer and set the batch size to 64, the learning rate to 10^{-3}, and the weight decay to 10^{-5}.

4.2. Qualitative Results

Sound2Scene generates visually plausible images compatible with a single input waveform, as shown in Fig.1 and 4. It is not limited to a handful number of categories but handles diverse categories, from animals, and vehicles, to sceneries, *etc*. We highlight that the proposed model can even distinguish subtle differences in similar sound categories, such as "engine" (Fig. 4 col. 1-2) and "water" (Fig. 4 col. 3) related sounds, and produces accurate and distinct images. See the supplement for more results.

4.3. Quantitative Analysis

Comparison with other methods. We compare our model with the prior arts of which codes are publicly available,



	Method	VEGAS (5 classes)		
		R@1	FID (↓)	IS (↑)
(A)	Pedersoli *et al.* [52]	23.10	118.68	1.19
(B)	S2I [20]	39.19	114.84	1.45
(C)	Ours	**77.58**	**34.68**	**4.01**

Figure 5. **Comparison to the baseline** [52] **and existing sound-to-image method** [20]. Our method outperforms the others both qualitatively and quantitatively in the VEGAS dataset.

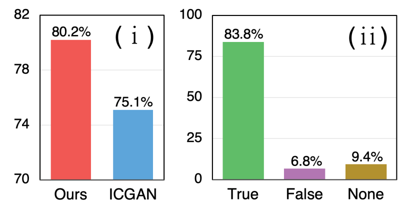
S2I[2] [20] and Pedersoli[3] *et al.* [52]. Note that Pedersoli *et al.* is not targeted for sound-to-image but uses VQVAE-based model [67] for sound-to-depth or segmentation generations. Though our model can handle more diverse categories of in-the-wild audio, we follow the training setup as in S2I by training our model and Pedersoli *et al.* with five categories in VEGAS for a fair comparison. As shown in Table 5, our model outperforms all the other methods. Additionally, it generates visually plausible images, while previous methods fail to generate recognizable images. We postulate that learning visually enriched audio embeddings combined with a powerful image generator leads to superior results.[4]

Comparison with strong baselines. We further compare our proposed method with closely-related baselines in Table 1 (a). First, we compare with an image-to-image generation model identical to ICGAN [11] (A). Our model (B) shares the same image generator with (A), but differs in the encoder type and the input modality. As shown, (B) outperforms or gives comparable results to (A) in all metrics. We presume that the noisy characteristic of the video datasets causes (A) to fail to extract a good visual feature for image generation, while audio is relatively robust and less sensitive to those limitations, resulting in generating plausible images. We also compare our model (B) with a retrieval system (C) that can be regarded as a strong baseline. Given an input audio embedding, the retrieval system finds the closest image from the database. (C) shares the same audio encoder with (B), but the image generator G is replaced with the same

[2] https://github.com/leofanzeres/s2i
[3] https://github.com/ubc-vision/audio_manifold
[4] More qualitative comparisons can be found in the supp. material.

	Method	Encoder (V/A)	Generator (G/R)	VGGSound (50 classes)				VEGAS	
				R@1	R@5	FID (↓)	IS (↑)	R@1	R@5
(A)	ICGAN [11]	V	G	30.06	62.59	**16.11**	12.61	46.60	82.48
(B)	Ours	A	G	**40.71**	**77.36**	17.97	**19.46**	**57.44**	**84.08**
(C)	Retrieval	A	R	51.28	80.37	-	-	67.20	85.00
(D)	Upper bound	-	-	57.82	85.79	-	-	73.60	88.2

(a) Comparison to baselines



(b) User study

Table 1. **Quantitative evaluations.** We compare our method with different baselines (different settings for the encoder and the generator) on CLIP retrieval (R@k), FID, and IS in (a). For user study, we first compare our method with ICGAN by measuring recall probability between generated images of ICGAN and our method from the same audio-visual pair. Second, we validate our method's output for the given audio. Results are in (b) respectively. *Abbr.* V: image encoder, A: audio encoder, G: image generator, R: retrieval system.

	Loss	F	Duration	VGGSound (50 classes)			
				R@1	R@5	FID (↓)	IS (↑)
(A)	L_2	✓	10 sec.	18.21	46.69	24.05	9.97
(B)	L_{nce}	✓	10 sec.	31.63	66.04	27.05	12.92
(C)	L_{total}		10 sec.	37.20	73.13	21.20	17.51
(D)	L_{total}	✓	1 sec.	35.85	72.02	19.05	17.87
(E)			5 sec.	38.24	75.76	20.43	18.81
(F)	L_{total}	✓	10 sec.	**40.71**	**77.36**	**17.97**	**19.46**

Table 2. **Ablation studies of our proposed method.** We compare the different configurations of our method by changing the loss functions, frame selection method, and duration of the audio. F denotes the frame selection method.

memory-sized database of the images from the training data. (D) is an upper bound in which the extracted video frames are directly used for the evaluations. The performance gap between (C) and (D) is dramatically lower than that of (B) and (D), which justifies that our audio encoder properly maps the input audio to the joint embedding space. (C) outperforms (B) on $R@1$ for both datasets, while (B) performs comparably to (C) in $R@5$. Though the image generators have room to improve, these results show that our method can reach to the proximity of the strong baseline.

User study. We summarize the user study in Table 1 (b) with two experiments: (i) comparison to ICGAN and (ii) validation of the proper image generation for given audio. Each experiment has 20 questions. In (i), audio and five images are given to the participants. Among the five, two are generated by ours and ICGAN, respectively, and the rest are randomly generated from either method. Participants choose all the images that illustrate the given sound, and we check the preference by comparing the recall probability of ICGAN and ours. In (ii), audio and four images are provided to the participants. All four images are generated by ours, but only one is from the given sound. Participants choose only one image that best illustrates the given sound. As in (i), our model is more preferred. Moreover, (ii) shows that the precision of our method is 83.8%, which supports our model generating highly-correlated images to the given sounds.

Ablation studies. We conduct a series of experiments in order to verify our design choices in Table 2. We compare the performance of applying different distillation losses: a simple L_2 loss between the image and audio feature, and InfoNCE loss [44] with a cosine similarity measurement, L_{nce}, rather than using L_2 distance as in Eq. (2). As the results of (A), (B), and (F) reveal, our loss choice (F) leads to producing more diverse and higher quality results. We also observe that the frame selection method, discussed in Sec. 3.3, brings extra performance improvement; see the performance difference between (C) and (F). Finally, we test the effect of audio duration. We train models with 1, 5, and 10 seconds of sounds with other experimental settings being fixed. By comparing (D), (E), and (F), we observe that feeding longer sounds consistently improves performance. We presume that the longer audios capture more descriptive audio semantics, while shorter ones are vulnerable to missing them.

5. Controllability of Sound2Scene

Our model learns the natural correspondence between audio and visual signals via aligned audio-visual embedding space. Thus, intuitively, we ask if manipulations on input can result in corresponding changes in the generated images. We observe that even without an explicit objective, our model allows controllable outputs by applying simple manipulations on inputs in the *waveform space* or learned *latent space*. This opens up interesting experiments that we explore below.

5.1. Waveform Manipulation for Image Generation

Changing the volume. Humans can roughly predict the distance or the size of an instance by the volume of the sound. To check if our model can also understand the volume differences, we reduce and increase the volume of the reference audio. Each audio with a different volume is fed into our model with the same noise vector. As shown in Fig. 6, the instances in the synthesized images get larger as the volumes increase. Interestingly, the volume changes in "Water Flowing" illustrate the different flows of water, while "Rail Transport" shows a train approaching in the scene. These results highlight that our model has not only class-specific understanding but also the relation between the volume of

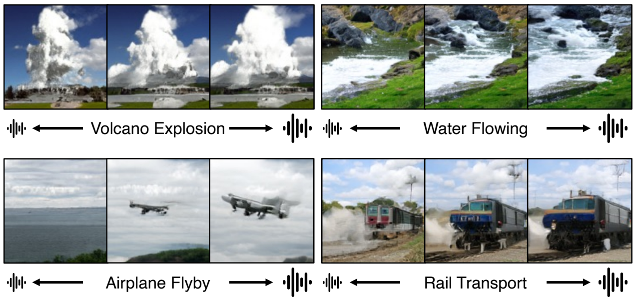


Figure 6. **Generated images by changing the volumes of the input audio in the *waveform space*.** As the volume increases, the objects of the sound source become larger or more dynamic.

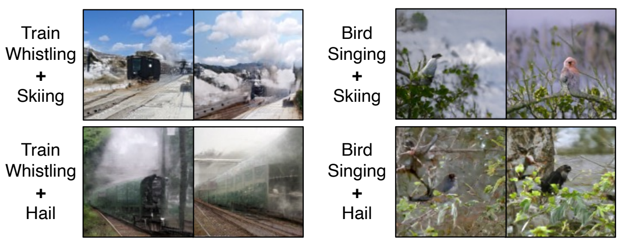


Figure 7. **Generated images by mixing two different audios in the *waveform space*.**

the audio and visual changes. We assume that the supervision from the visual modality enables our model to capture such strong and expressive audio-visual relationships.

Mixing waveforms. We investigate if our model can capture the existence of multiple sounds in the generated images. To this end, we mix two waveforms into a single one and feed it to our model. As shown in Fig. 1 and Fig. 7, our model synthesizes images by reflecting those multiple audio semantics. For example, the railroad or a bird pops up across the snowy scene when mixing with the "Skiing" sound, and the train and bird appear in the misty scene when mixing with the "Hail" sound. Also, as in Fig. 1, mixing the "Dog Barking" and "Water Flowing" sounds outputs a scene with a dog playing in the water. Sensing multiple separate sounds from a single mixed audio input [32], *i.e.*, audio source separation [22], and generating their visual appearance in the proper context is not trivial. However, our results show that the proposed model can handle this to a certain extent.

Mixing waveforms and changing the volume. Here, we manipulate the input waveform by combining the multiple waveforms and changing their volumes at the same time. In Fig. 8, we mix the "Wind" sound with each of the "Bird" and "Dog" sound with volume changes. As the "Wind" sound gets larger while the "Bird" sound decreases, the bird gets smaller and is finally covered with the bushes. In the same experiment setting, a close-up shot of the dog indoors starts zooming out and gets a wide shot in the outdoor environment. These results show that our model can capture subtle changes in the audio and reflect them to generate images.

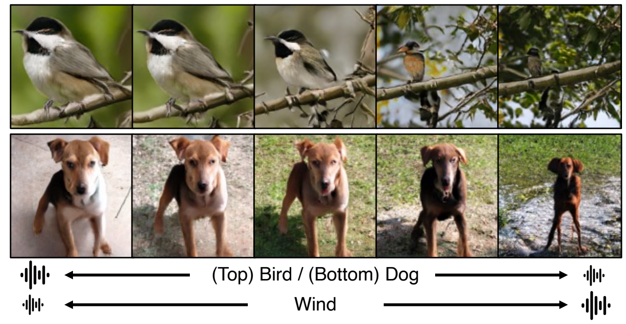
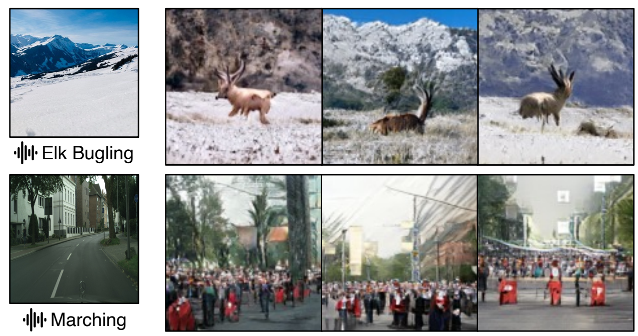


Figure 8. **Generated images by mixing multiple audios with volume changes in the *waveform space*.** We observe that Sound2Scene mimics the camera movement by placing the object further as the wind sound gets larger.



(a) Inputs (b) Generated images

Figure 9. **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.

5.2. Latent Manipulation for Image Generation

As we construct an aligned audio-visual embedding space, our model can also take an image and audio together as input and generate images. We introduce two different approaches (Fig. 1) for audio-visual conditioned image generation where both use the features of the inputs in the latent space.

Image and audio conditioned image generation. Given an image and audio, we extract a visual feature $\mathbf{z^V}$ and an audio feature $\mathbf{z^A}$. Then, we interpolate two different features in the latent space and obtain a novel feature: $\mathbf{z^{new}} = \lambda \mathbf{z^V} + (1-\lambda)\mathbf{z^A}$, where λ differs throughout the examples. This feature is fed to the image generator for generating an image. As shown in Fig. 9, this simple approach can produce an image by putting the sound context into the scene, such as a marching sound bringing parade-looking people or a loud elk sound making an elk appears in a snowy scene.

We further use this approach to compare our method with the recent sound-guided image manipulation approach[5] [36] in Fig. 10. Note that this task is not targeted explicitly by our model but appears as a natural outcome of our design.

[5] https://github.com/kuai-lab/sound-guided-semantic-image-manipulation

Method	Target task
[36]	Sound-guided image manipulation
Ours	Sound-to-image generation

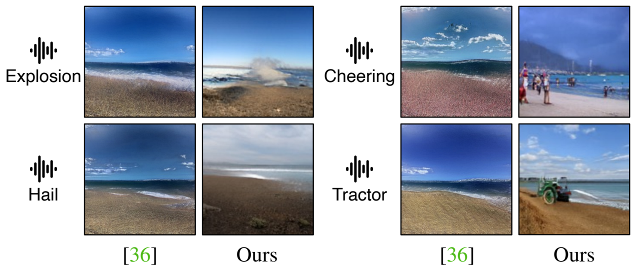


Figure 10. **Qualitative comparison of our method and Lee *et al*.** [36]. Lee *et al*. fail to insert an object while maintaining the contents of the given image. Our method, by contrast, successfully inserts objects that sound in the scene by generating a new image. Note that both works target different tasks.

While Lee *et al*. [36] preserves the overall content of the given image, it fails to insert an object corresponding to the sound. In contrast, our method creates an image (nearly the same as the given one) by conditioning on both modalities, for example, inserting an explosion and tractor in the scene or making the ocean view look cloudy due to the hail.

Image editing with paired sound. We approach sound-guided image editing from a different perspective by manipulating the inputs in the latent space. By using GAN inversion [1, 57], we extract a visual feature $\mathbf{z}^{\mathbf{V}}_{\mathbf{inv}}$ and corresponding noise vector $\mathbf{z}^{\mathbf{N}}_{\mathbf{inv}}$ for the given image. Additionally, we change the volume of the corresponding audio and extract two different audio features, $\mathbf{z}^{\mathbf{A}}_{\mathbf{1}}$ and $\mathbf{z}^{\mathbf{A}}_{\mathbf{2}}$, respectively. We move the visual feature toward the direction of the difference between the two audio features and obtain a novel feature: $\mathbf{z}^{\mathbf{new}} = \mathbf{z}^{\mathbf{V}}_{\mathbf{inv}} + \lambda(\mathbf{z}^{\mathbf{A}}_{\mathbf{1}} - \mathbf{z}^{\mathbf{A}}_{\mathbf{2}})$, *i.e*., manipulating the visual feature with the audio guidance. By using this new feature with image generator, $G(\mathbf{z}^{\mathbf{N}}_{\mathbf{inv}}, \mathbf{z}^{\mathbf{new}})$, the original image is edited. As shown in Fig. 11, by simply changing the volume – moving through the latent space – we can change the flow of the waterfall or make the ocean wave stronger or calmer.

6. Discussion

Generalization. We show generalization of the model to *some extent* in two settings: 1) Generating images from unseen categories that are semantically similar to the training set as sound often carries overlapping information (Fig. 12).

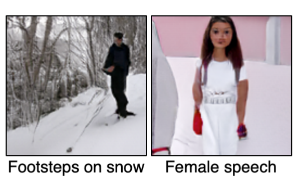


Figure 12. **Generalization to unseen classes.**

2) Compositionality (dog barking+water flowing in Fig. 1). However, our method may not be generalized for every unseen category as similar limitation is also common in other X-to-Vision tasks.

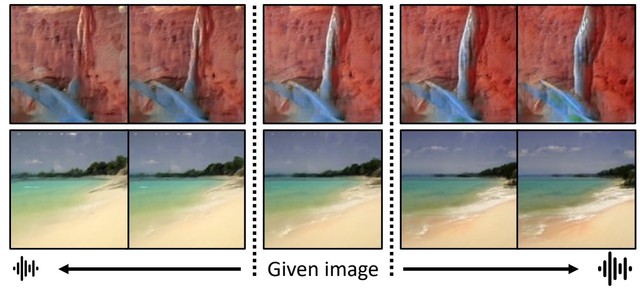


Figure 11. **Image editing by volume changes in *latent space*.** We extract an image feature and noise vector by GAN inversion, and two audio features with different volumes. Then, we move the image feature in the direction of the audio feature differences.

(a) Mixing objects (b) Incomplete human forms

Figure 13. **Examples of failure results.** In cases where both source audio specify certain objects, our model mixes two objects into one (a). Our model also tends to produce incomplete human forms (b).

Failure cases. Although our method shows favorable results for given audios, single or mixture, there are some failure cases we have observed. The first phenomenon is that our model often generates images with a single or a blended object unintentionally when the audios used in the mixture specify two distinct but similar-sounding objects (see examples in Fig. 13 (a)). Another one we have observed is that the quality of the outputs is lower in human-related categories (see examples in Fig. 13 (b)). This phenomenon is shared across typical GAN-based generative models, *e.g*., [38], when training the generator with generic object dataset, *i.e*., ImageNet.

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.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 8

[2] 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. 3

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[5] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[6] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2, 3

[9] 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. 4

[10] 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

[11] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 4, 5, 6

[12] Moitreya Chatterjee and Anoop Cherian. Sound2sight: Generating visual dynamics from sound and context. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[13] 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. 2, 3, 4

[14] 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. 2, 4

[15] 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. 1, 2, 3

[16] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing (TIP)*, 2020. 2

[17] 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. 2, 3

[18] 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. 4

[19] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[20] Leonardo A Fanzeres and Climent Nadeu. Sound-to-imagination: Unsupervised crossmodal translation using deep dense network architecture. *arXiv preprint arXiv:2106.01266*, 2021. 1, 2, 3, 5

[21] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[22] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[23] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2

[25] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision (IJCV)*, 2021. 3

[26] 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. 1, 2, 3

[27] 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. 4

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 4

[29] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[30] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[31] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[32] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 7

[33] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *British Machine Vision Conference (BMVC)*, 2021. 2

[34] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[35] Minkyu Kim, Kim Sung-Bin, and Tae-Hyun Oh. Prefix tuning for automated audio captioning. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023. 2

[36] 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. 1, 2, 3, 4, 7, 8

[37] Tingle Li, Yichen Liu, Andrew Owens, and Hang Zhao. Learning visual styles from audio-visual associations. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[38] Yuheng Li, Yijun Li, Jingwan Lu, Eli Shechtman, Yong Jae Lee, and Krishna Kumar Singh. Collaging class-specific gans for semantic image synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 8

[39] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[40] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[41] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[42] Medhini Narasimhan, Shiry Ginosar, Andrew Owens, Alexei A Efros, and Trevor Darrell. Strumming to the beat: Audio-conditioned contrastive video textures. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022. 2

[43] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 6

[45] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[46] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[47] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[48] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *International Journal of Computer Vision (IJCV)*, 2018. 2

[49] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 4

[50] Sooyoung Park, Arda Senocak, and Joon Son Chung. Margin-nce: Robust sound localization with a negative margin. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023. 2

[51] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[52] 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. 5

[53] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2, 4

[55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2021. 2

[57] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 8

[58] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 4

[59] 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. 3

[60] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[61] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 2

[62] 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. 2, 4

[63] Joo Yong Shim, Joongheon Kim, and Jong-Kook Kim. S2i-bird: Sound-to-image generation of bird species using generative adversarial networks. In *International Conference on Pattern Recognition (ICPR)*, 2021. 1, 2

[64] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[65] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: Audio generation for a silent performance video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[66] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[67] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[68] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[69] 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. 1, 2, 3

[70] 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. 3

[71] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[72] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[73] 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. 2, 4