Learning Visual Styles from Audio-Visual Associations

Tingle Li[1,3] Yichen Liu[1] Andrew Owens[2,*] Hang Zhao[1,3,*]
[1]IIIS, Tsinghua University [2]University of Michigan [3]Shanghai Qi Zhi Institute
https://tinglok.netlify.com/files/adis/

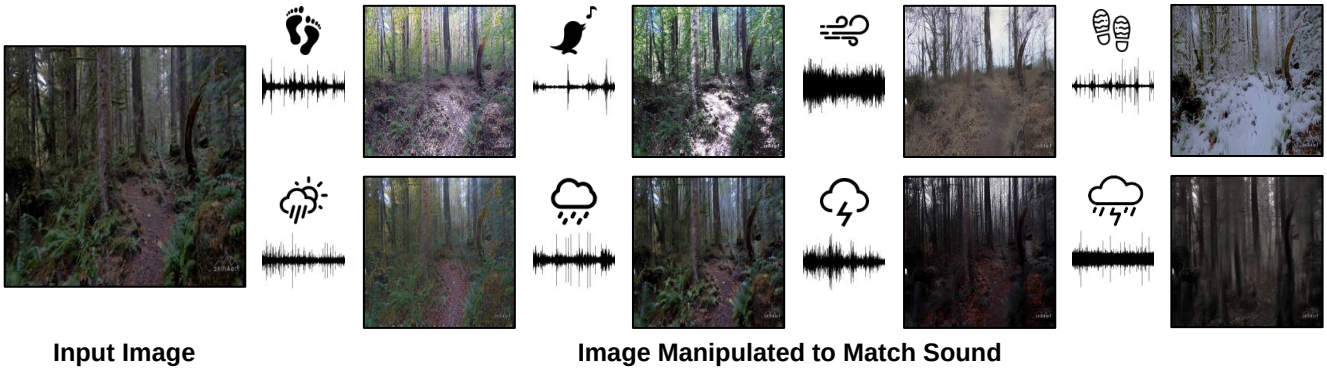


Figure 1. **Audio-driven image stylization.** We train a model to change the style of an input image to match an input sound. After training with an unlabeled dataset of egocentric hiking videos, our model learns visual styles for a variety of ambient sounds, such as sprinkle and heavy rains, as well as physical interactions, such as footsteps.

Abstract

From the patter of rain to the crunch of snow, the sounds we hear often reveal the visual textures within a scene. In this paper, we present a method for manipulating the texture of a scene to match a sound, a problem we term audio-driven image stylization. *Our model learns to restyle images through self-supervised learning. Given a dataset of paired visual and audio data, the model learns to manipulate the image such that, after manipulation, the sound is more likely to co-occur with it. Our model is conditioned on sounds, and it outperforms label-based counterparts in both quantitative and qualitative evaluations. We also empirically find that changing the audio volume and mixture results in predictable visual changes.*

1. Introduction

Recent work has proposed a variety of methods for changing the style of an input image. In these methods, the desired style is specified using other example images [18, 25, 28, 30], such as paintings, and more recently through human language, such as through semantic labels, text, or scene graphs [4,31,52]. These approaches, however, heavily rely on human-provided annotations, such as paired images and text. This limits the possible styles that can be learned by the model, as it can be difficult to precisely capture all possible styles through the medium of words.

Audio, by contrast, is often paired with vision through video recording, making it easy to train a stylization model in a self-supervised manner. Furthermore, audio is a natural signal that contains information about the subtle distinctions between similar scenes. For example, asking a model to generate images depicting "rainy woodland" can be ambiguous. Providing the sound of rain, on the other hand, specifies whether the rain is light or heavy, as well as whether thunder is likely to co-occur with it. Finally, audio can be a useful natural "embedding space" for specifying image styles, as intuitive changes to the audio, such as adjusting the volume or mixing two sounds together, can result in predictable visual changes.

In this paper, we propose to learn *audio-driven* image stylization. Given an input image and a target sound, our model converts the style of the image such that it would better match the sound, while preserving the image's structural content. Through this process, our model learns a variety of style associations, each of which can be specified by a sound. For example, birds chirping to blue skies, crunching footsteps to snow, and different volumes of rains (Figure 1).

*Co-advise on the project.

Based on this idea, we propose a method that combines conditional generative adversarial networks [20] and contrastive learning [21]. The former employs an audio-visual discriminator to determine whether the generated image and target audio are likely to be a pair, with the goal of converting the source image's style. The latter includes a multi-scale patch-wise structure discriminator [46] that maximizes the mutual information between the source and generated images in order to preserve the structural content of the scene after conversion. Then, we train on a dataset of egocentric hiking videos collected from the internet.

After training, our model can manipulate images to match a variety of visual styles, each specified using sound. Since a wide range of sounds are provided during training, our model is capable of one-to-many texture conversion. Through quantitative evaluations and human perceptual studies, we demonstrate the effectiveness of our model's texture conversion capability. We also provide qualitative results showing how that straightforwardly modifying the audio, by mixing it or changing its volume, leads to corresponding changes in image style.

In summary, our contributions are as follows:

- We propose the task of *audio-driven image stylization* (ADIS), which aims to convert the texture of a visual scene conditioned on sound.
- We present a contrastive-based audio-visual GAN model that learns from unlabeled data to manipulate visual scene texture via audio.
- We evaluate our model on a dataset of egocentric hiking videos, which we call the *Into the Wild* dataset. We show through automated evaluation metrics, human perceptual studies, and qualitative examples that our model successfully learns audio-driven stylization.

2. Related Work

Image-to-image translation In past years, image-to-image (I2I) translation has undergone the transition from paired to unpaired manner. Under the paired translation setting [29], the training set only contains paired images, *i.e.*, pixel-wise correspondence, between the source domain and the target domain. Thus, unpaired I2I becomes increasingly popular due to data efficiency. Ever since the emergence of GAN [20], its variants [32, 46, 58, 64] have been widely adopted in unpaired I2I translation. For text-based I2I translation, [35] proposes to manipulate the weather of the source image through pre-defined textual instructions. With the help of GAN and CLIP [51], several methods [4, 13, 15, 42, 52] have been proposed to generate more plausible images using captions. However, text-based methods either requires weak supervision (*e.g.*, paired text and images) [37] or explicit image descriptions (*e.g.*, text describing an image as a line drawing) [55], which can be inefficient for large datasets. Ours, by contrast, uses audio to reach comparable capability without any requirements for supervision.

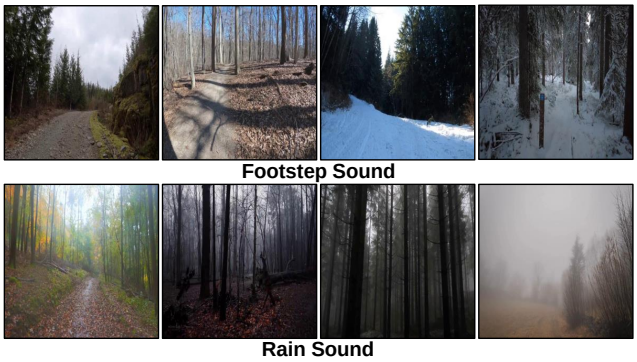


Figure 2. Categories can fail to convey subtle nuances between audio-visual events. We show frames from videos whose corresponding audios are predicted as *footstep* or *rain* sound by a classifier [19, 48].

Audio-visual correspondence Audio and visual signals are naturally co-occurred when they are recorded as video. In order to leverage this natural correspondence, researchers have introduced various tasks, such as representation learning [2, 12, 34, 41, 43, 44], audio source separation [14, 16, 59, 60], audio source grounding [5, 22], audio spatialization [17, 40, 57], visual speech recognition [1], and scene classification [6, 19]. Inspired by these works that use audio-visual correspondence, we propose a novel task termed audio-driven image stylization, aiming to conduct image translation using sounds like birds chirping, rain and footsteps.

Audio-visual synthesis Previous work has looked into the idea of synthesizing sound or images using the relationship between them. One line of work has developed models for generating sound from silent videos, such as impact sound [45], natural sound [63] and human speech [27, 49]. Another line of work has devised models to synthesize images or videos from sound. For example, talking head generation [10, 50, 62] aims to generate a talking face from a set of images of a person conditioned on speech. Inverse-Foley animation [36] seeks to synchronize rigid-body motions with contact sound. Other work [8] estimates depth from quiet indoor sounds. Unlike the above works, we concentrate on restyling plausible images using the source image and various sorts of sound, such as ambient and impact sounds.

3. Audio-driven Image Stylization

Audio is an abundant signal that conveys useful information, which co-occurs with visual signals in video. This allows us to establish audio-visual style associations without the need for human-provided supervision. Audio may also convey subtle distinctions that may not be obvious from simple categories. Figure 2 shows examples where videos

from the same semantic category could appear differently, such as different amounts of rainfall (sprinkle *vs*. heavy rain), but this subtle difference can be perceived by listening to the sound. Thus, we propose *audio-driven* image stylization (ADIS) as a novel multi-modal generation task. To the best of our knowledge, this is the first work that performs image stylization conditioned on audio.

3.1. Proposed Method

In ADIS, the goal is to learn a feature mapping from a source domain $\mathcal{X}$ to a target domain $\mathcal{Y}$, where $\mathcal{Y}$ is determined by the audio condition, denoted as the audio domain $\mathcal{A}$. To achieve this goal, we propose a self-supervised training method, which can be trained on unpaired videos, *i.e.*, the video pairs taken from $\mathcal{X}$ and $\mathcal{Y}$ are not sharing content. This can be accomplished through two distinct training objectives.

Texture conversion via adversarial training We introduce an audio-visual adversarial objective that discriminates whether an image is co-occurred with a given audio. Under this novel training scheme, the generated image is encouraged to match the target audio which is particularly suitable for ADIS. Specifically, the generator G consists of two components, an encoder G_{enc} followed by a decoder G_{dec}. For a given dataset of unpaired image instances $X = \{\boldsymbol{x} \in \mathcal{X}\}$, $Y = \{\boldsymbol{y} \in \mathcal{Y}\}$, and the audios $A_Y = \{\boldsymbol{a}_Y \in \mathcal{A}\}$ corresponding to Y, G_{enc} and G_{dec} are applied sequentially to generate the output image $\hat{\boldsymbol{y}} = G_{\text{dec}}(\text{concat}(G_{\text{enc}}(\boldsymbol{x}), f(\boldsymbol{a}_Y)))$, where f is a audio feature extractor. The audio-visual adversarial loss [20] is then applied to increase the association between $\hat{\boldsymbol{y}}$ and $\boldsymbol{a}_Y$:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G_{X\to Y}, D_Y) = \ & \mathbb{E}_{\boldsymbol{y}\sim Y} \log D(\boldsymbol{y}, \boldsymbol{a}_Y) + \\ & \mathbb{E}_{\boldsymbol{x}\sim X} \log\left(1 - D(G(\boldsymbol{x}, f(\boldsymbol{a}_Y)), \boldsymbol{a}_Y)\right)\end{aligned} \tag{1}$$

where D is the discriminator. Please note that the fusion of two modalities in D is an early fusion, where the spectrogram of $\boldsymbol{a}_Y$ is directly concatenated to $\hat{\boldsymbol{y}} = G(\boldsymbol{x}, \boldsymbol{a}_Y)$ before feeding into D. We empirically found that this fusion strategy yields better results in terms of visual quality.

Structure preservation via contrastive learning In this task, a successfully restyled image should be equipped with the texture that can be interpreted by the target audio, while fully preserving the structure of the source image. However, both information, *i.e.*, texture and structure information, are inherently entangled within the learned feature, and adversarial training can only convert texture. One trivial solution could be that we get the same image for any inputs. Therefore, as shown in Figure 3, we introduce the second training objective based on noise contrastive estimation (NCE) [21], which aims to preserve structure information by establishing mutual correspondence between the source and generated images, $\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$ respectively. Note that this training objective is only employed to the encoder network G_{enc}, which is a multi-layer convolutional network that transforms the source image into feature stacks at each layer. In this way, we encourage G_{enc} to abandon the texture

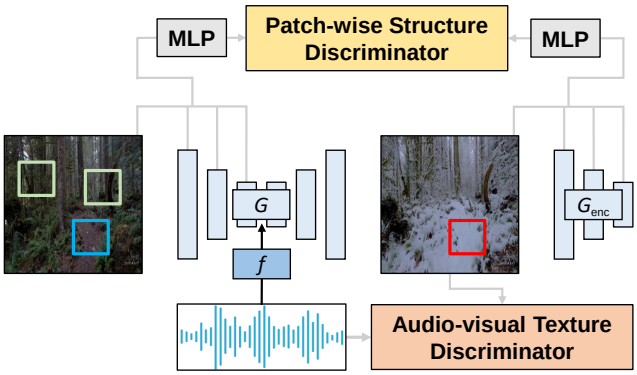


Figure 3. **Model architecture**. The multi-scale patch-wise structure discriminator [46] is used to preserve the scene structure, while the audio-visual texture discriminator is used to convert the scene texture. This is an example where sunny forest is converted to snowy counterpart. The **generated snow patch** should match its corresponding **input dirt patch**, in comparison to **other random patches**. Note that the MLP component will be waived during inference.

of the source image while preserving the structure, and then the job of the decoder network G_{dec} is to integrate the target texture to the source image.

Given a "query" vector $\boldsymbol{q}$, the fundamental objective in contrastive learning is to optimize the probability of selecting the corresponding "positive" sample $\boldsymbol{v}^+$ among N "negative" samples $\boldsymbol{v}^-$. The query, positive and N negatives are mapped to M-dimensional vectors by a MLP, *i.e.*, $\boldsymbol{q}, \boldsymbol{v}^+ \in \mathbb{R}^M$ and $\boldsymbol{v}^- \in \mathbb{R}^{N\times M}$. This problem setting can be expressed as a multi-classification task with $N+1$ classes:

$$\ell(\boldsymbol{q}, \boldsymbol{v}^+, \boldsymbol{v}^-) = -\log\left(\frac{\exp(\frac{\boldsymbol{q}\cdot\boldsymbol{v}^+}{\tau})}{\exp(\frac{\boldsymbol{q}\cdot\boldsymbol{v}^+}{\tau}) + \Sigma_{n=1}^{N}\exp(\frac{\boldsymbol{q}\cdot\boldsymbol{v}_n^-}{\tau})}\right) \tag{2}$$

where $\boldsymbol{v}_n^-$ denotes the n-th negative sample and τ is a temperature parameter, as suggested in SimCLR [7], that scales the similarity distance between $\boldsymbol{q}$ and other samples. The cross-entropy term in Eq.(2) represents the probability of matching $\boldsymbol{q}$ with the corresponding positive sample $\boldsymbol{v}^+$. Thus, iteratively minimizing the negative log-cross-entropy is equivalent to establishing mutual correspondence between the query and sample spaces.

In our task, we draw the $N+1$ positive/negative samples from the the source image $\boldsymbol{x} \in X$, and the query $\boldsymbol{q}$ is selected from the generated image $\hat{\boldsymbol{y}}$. From Figure 3, it can be seen that the selected samples are "patches" that capture local information among the image features. This setup is motivated by the logical assumption that the global correspondence between $\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$ is determined by the local, *i.e.*, patch-wise, correspondences.

Since the encoder G_{enc} is a multi-layer convolutional network that maps $\boldsymbol{x}$ into feature stacks after each layer, we choose L layers and pass their feature stacks through a

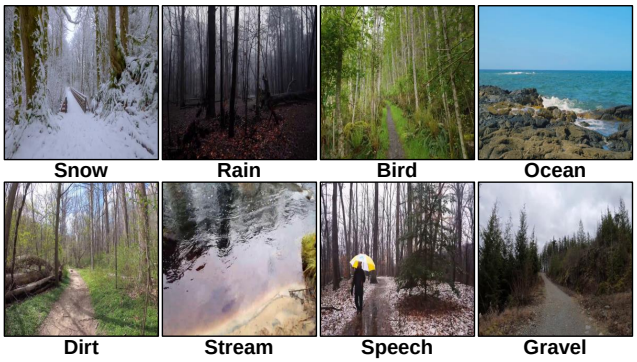


Figure 4. **Selected frames from the *Into the Wild* dataset.** We show example images corresponding to the top-1 categorical sounds deduced by a classifier [19, 48].

small MLP network P. The output of P is $P(G^l_{\text{enc}}(\boldsymbol{x})) = \{\boldsymbol{v}^1_l, ..., \boldsymbol{v}^N_l, \boldsymbol{v}^{N+1}_l\}$, where $l \in \{1, 2..., L\}$ denotes the index of the chosen encoder layers and $G^l_{\text{enc}}(\boldsymbol{x})$ is the output feature stack of the l-th layer. Similarly, we can obtain the query set by encoding the generated spectrogram $\hat{\boldsymbol{y}}$ into $\{\boldsymbol{q}^1_l, ..., \boldsymbol{q}^N_l, \boldsymbol{q}^{N+1}_l\} = P(G^l_{\text{enc}}(\hat{\boldsymbol{y}}))$. Now we let $\boldsymbol{v}^n_l \in \mathbb{R}^M$ and $\boldsymbol{v}^{(N+1)\setminus n}_l \in \mathbb{R}^{N\times M}$ denote the corresponding positive sample and the N negative samples, respectively, where n is the sample index and M is the channel size of P. By referring to Eq.(2), our second training objective can be expressed as:

$$\mathcal{L}_{\text{NCE}}(G_{\text{enc}}, P, X) = \mathbb{E}_{\boldsymbol{x}\sim X} \sum_{l=1}^{L} \sum_{n=1}^{N+1} \ell(\boldsymbol{q}^n_l, \boldsymbol{v}^n_l, \boldsymbol{v}^{(N+1)\setminus n}_l) \quad (3)$$

which is the average NCE loss from all L encoder layers.

Overall objective In addition to the two objectives discussed above, we have also employed an identity loss $\mathcal{L}_{\text{identity}} = \mathcal{L}_{\text{NCE}}(G_{\text{enc}}, P, Y)$ which also leverages the NCE expression in Eq.(3). By taking the NCE loss on the identity generation process, *i.e.*, generating $\hat{\boldsymbol{y}}$ from $\boldsymbol{y}$, we are likely to prevent the generator from making unexpected changes. Now we can define our final training objective as:

$$\begin{aligned}\mathcal{L}_{\text{final}} = \; & \mathcal{L}_{\text{GAN}}(G_{X\to Y}, D_Y) + \lambda\mathcal{L}_{\text{NCE}}(G_{\text{enc}}, P, X) + \\ & \mu\mathcal{L}_{\text{NCE}}(G_{\text{enc}}, P, Y)\end{aligned} \quad (4)$$

where λ and μ are two parameters for adjusting the strengths of the NCE and identity loss.

4. Experiments

4.1. Experimental Setup

Dataset We perform ADIS with two different datasets: *Greatest Hits* and *Into the Wild*. The Greatest Hits offers the sounds of different materials, and the hiking videos in *Into the Wild* introduce various sounds in nature.

- ***Into the Wild* dataset**: We collect *Into the Wild* to study the audio-visual association in nature (Figure 4), such as seasonal variation, rainfall, and *etc*. In particular, we collect 94 untrimmed videos from YouTube, ranging from 1.5 to 130 minutes long, with 50 hours in total. It is worth noting that all the videos are not containing any noisy sound, *e.g.*, background music, such that the sounds are only related to hiking in nature. See Appendix A.1 for more dataset details.
- **The *Greatest Hits* dataset** [45]: The *Greatest Hits* dataset contains a drumstick hitting, scratching, and poking different objects in both indoor and outdoor scenes. There are 977 videos in total, including both indoor (64%) and outdoor scenes (36%). However, since this dataset was originally gathered for sound generation, each video more or less contains visual noise, making it challenging to perform AVTC. For example, ceramic bowls have different colors but the hitting sounds are similar across all bowls. It can be sometimes difficult for the model to determine the texture of a material with different colors. To alleviate this issue, we manually select some outdoor scene videos with less diverse backgrounds, such as hitting dirt, water, gravel and grass.

Network architecture The encoder and decoder of the GAN generator are fully 2D convolutional networks, with 9 layers of ResNet-based CNN bottlenecks [30] in between. Except for the first CNN layer with a kernel size of 7 $\times$ 7, the others are 3 $\times$ 3, and the stride size is determined by whether downsampling is required. We employed the PatchGAN architecture [29] for the discriminator. A ResNet18 backbone [23] is also used for extracting audio features before feeding them into the decoder of the GAN generator. Furthermore, before computing the NCE loss, we extract intermediate features from the encoder of the generator with five different scales, and then apply a 2-layer MLP with 256 units to map each feature.

Training details For training efficiency, we devise the following pre-processing paradigm: i) before saving as images, each video is interpolated to 512$\times$512 scale and uniformly sampled 8 frames from it; ii) each audio is randomly truncated or tiled to a fixed duration of 3 seconds, then converted to 16 kHz and 32-bit precision in floating-point PCM format; iii) nnAudio [9] is used for conducting a 512-point discrete Fourier transform with a frame length of 25 ms and a frame-shift of 10 ms. For the hyper-parameters, both λ and μ in Eq.(4) are set to 0.5. We also employ random crop and horizontal flip as the image data augmentation. Our model is then trained using the Adam optimizer [33] with a batch size of 16 and an initial learning rate of 2×10^{-4} over 50 epochs. Other training strategies are available in Appendix A.2.

Evaluation metrics To get a better understanding of why audio is important, we quantitatively compare our model to several label-based baselines, using both objective and

Table 1. Evaluation results on the *Into the Wild* dataset. The subjective AMT metric is presented with 95% confidence intervals.

Method	Evaluation Metrics			
	AVC (↑)	FID (↓)	CLIP (↑)	AMT (↑)
Target	0.842	/	0.247	/
Class Pred. [48]	0.801	91.417	0.228	1.547 ± 0.044
Keyword	0.809	38.066	0.236	1.982 ± 0.045
Ours	**0.820**	**34.139**	**0.238**	$\mathbf{2.471 \pm 0.046}$

Table 2. AVC metric of specific scenes under our model and label-based baselines on the *Into the Wild* dataset.

Method	Audio-visual Correspondence (↑)		
	Sunny-Rainy	Snowy-Sunny	Sunny-Snowy
Class Pred. [48]	0.819	0.796	0.793
Keyword	0.827	0.802	0.808
Ours	**0.831**	**0.820**	**0.816**

subjective metrics (See Appendix A.3 for more evaluation details):

- **Audio-visual Correspondence (AVC)** [2]: AVC measures the correlation between audio and image. In our case, we extract audio and visual features using OpenL3 [11], a variant of L3-Net [2] pre-trained on AudioSet [19], and then use those features to compute the average cosine similarity. A higher correlation is associated with a higher AVC score.
- **Fréchet Inception Distance (FID)** [26]: FID estimates the distribution of real and generated image activations using trained network and measures the divergence between them. A lower FID score indicates that real and generated images are more relevant.
- **Contrastive Language-Image Pre-Training (CLIP)** [51]: CLIP is a neural network trained on a variety of image-text pairs. It can be instructed to predict the relationship between text snippet and image. A higher CLIP score indicates a better correlation between text and image.
- **Amazon Mechanical Turk (AMT)**: AMT is applied to examine audio-visual correlation from the human perspective, *i.e.*, a subjective evaluation. MTurker is required to rank such correlations based on the audios and images generated by various approaches, where the best one receives the highest score of 3, and the lowest score is 1.

Baselines We adopt two label-based methods for comparison. For both of them, Word2Vec [39] is used for generating the class embeddings, which is incorporated with the input image and serves as a textual condition.

- **Class Pred.** [48]: Class Pred. uses YAMNet, a state-of-the-art audio classification network [24] trained on AudioSet [19], to calculate the class logits. It is employed as an auto-labeling method to yield the semantic labels for all the audio clips.
- **Keyword**: Keyword is a human-labeling method in which each audio class is manually labeled with keywords from the video captions. It may be considered as the quasi-upper bound of label-based methods.

4.2. Comparison to Baselines

Quantitative results Since the sounds of hitting or scratching objects in the *Greatest Hits* are not applicable for quantitative evaluations, we only provide quantitative results yielded from the *Into the Wild* dataset. Table 1 shows the quantitative comparisons between our model and the label-based baselines. In objective evaluations, our model outperforms the baselines across the AVC, FID, and CLIP metrics, suggesting that our model can generate more realistic and congruent images. It is reasonable that Keyword is better than Class Pred., since auto-labeling methods can suffer from unsatisfactory accuracy. Specifically, Class Pred. involves 132 label classes inherited from AudioSet, whereas Keyword only has 3 classes (sunny, snowy and rainy) that are closely related to the scenes in *Into the Wild*. Moreover, the CLIP result of our model is on par with Keyword, which also indicates the benefit of using audio over label. For human evaluation, we randomly select 1000 images from the test set, and ask one MTurker to rank the level of the audio-visual correlation. It appears that the MTurker prefers our model over the baseline results in terms of relevance to the target audios, as shown in the last column of Table 1, which is consistent with the objective evaluation results.

To gain a better understanding of what makes our model work better, we divide the entire test set into three scenarios: sunny, rainy, and snowy. In this experiment, as shown in Table 2, our model still holds the best performance compared to the baselines. Furthermore, we observe that when the target scene is sunny, the disparity between our model and Keyword (0.018) is larger than the other scenes (0.004 & 0.008). We suppose that this is because the ambient sound in a sunny forest could be varied (*e.g.*, crunching gravel/leave, birds chirping, *etc*.), whereas in snowy and rainy forests, the ambient sound is scarce. In other words, if the sound is diverse, the gap between audio-based and label-based approaches will be large.

Qualitative results We show qualitative results in Figure 5 and additional results in Appendix A.4. It is worth mentioning that all of the results are yielded from one model, *i.e.*, by one-to-many conversion. For Keyword, it can generate plausible images that match the target audios, but with apparent flaws when converting between the same scene categories (the second and fourth row). For YAMNet, the generated images occasionally match the target im-

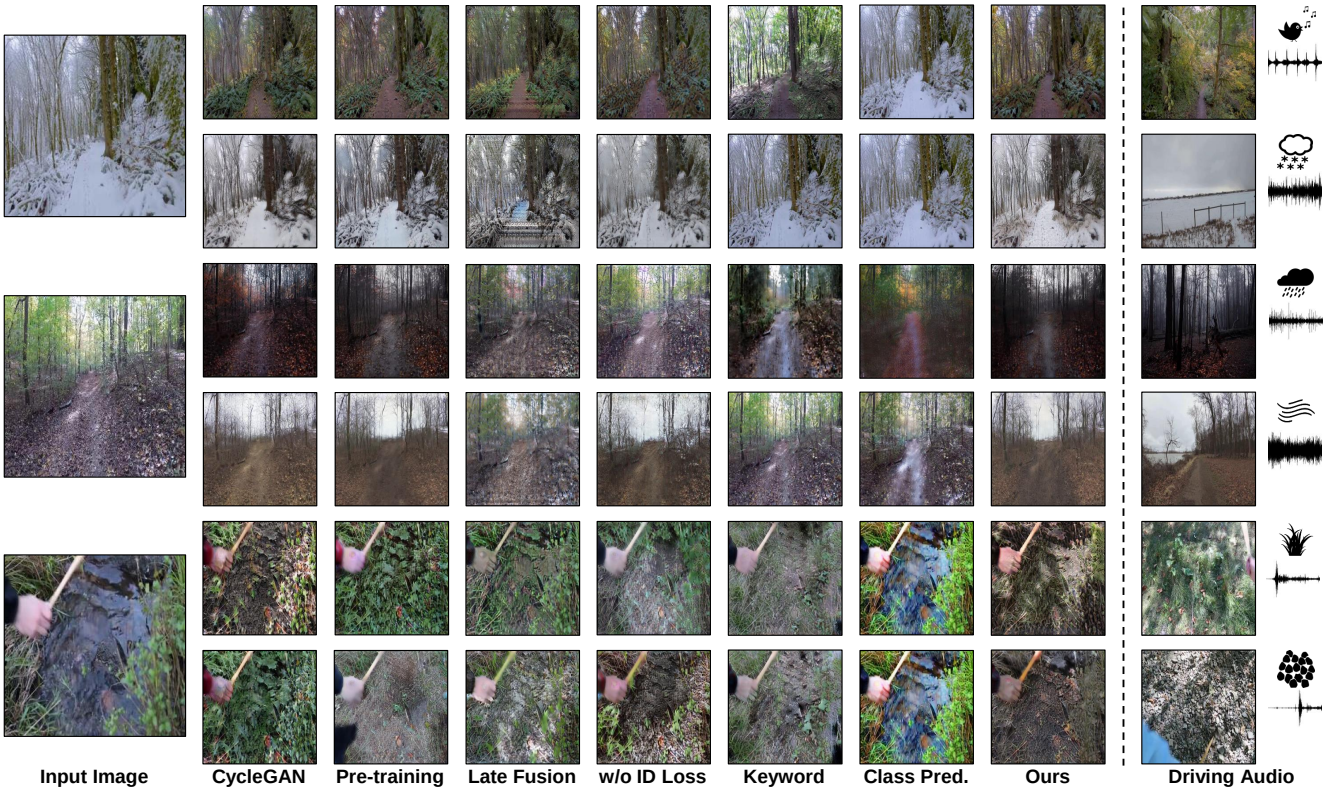


Figure 5. Qualitative comparison of baselines, ablations, and our model on audio-visual texture conversion. For reference, we also show driving audios as well as their corresponding images in the last column.

ages, but this does not happen in most cases. It's logical to see this because whether or not a transfer is successful is strongly dependent on whether or not the labels inferred by YAMNet are correct. Our model, by comparison, can capture the subtle distinction of the same scene class without any labels. When our model receives a wind+footstep sound, for example, in the second row of the first input image, it can change the hue of the snow to match the target image. Label-based baselines, by contrast, cannot achieve this merely with "snow" as text input.

4.3. Ablation Study and Analysis

We conduct an ablation study to test various settings and ablations of our model, summarized in Table 3. By default, we use the architecture and loss function above. We also try to use: i) the forward cycle-consistency loss [64] instead of NCE loss, termed as CycleGAN; ii) late fusion discriminator [54] to incorporate audio and visual features rather than early fusion one; iii) without the identity loss; iv) a pre-trained audio-visual self-supervised method, *i.e.*, SeLaVi [3], as the initial weight for the audio network in addition to training from scratch. Besides, we show qualitative examples and additional pre-training comparisons in Figure 5 and Appendix A.4 respectively.

NCE loss is a strong substitute for cycle-consistency loss Our model employs NCE loss following CUT [46]. As a baseline, cycle-consistency loss [64] can also preserve the image structure. As shown in Table 3, our model achieves comparable results to its counterpart, CycleGAN, implying that it can generate realistic images like CycleGAN. Figure 5 also shows some qualitative results that support this. Besides, CycleGAN involves the joint learning of two generators, while our model only requires one, which can reduce training time [46].

Late fusion discriminator is more likely to collapse In audio-visual learning, the late fusion architecture [54] is commonly used, in which two uni-modal encoders are employed to extract features, followed by a classifier (discriminator). We also take into account this architecture in ablations, with the results shown in Table 3 and Figure 5. We find that leveraging this type of discriminator induces the model to collapse, which means the generator would eventually become too weak to sustain the image structure, resulting in unsatisfactory results.

Identity loss helps to capture nuances Given an image from the output domain, the identity loss [64] pushes the generator to leave the image unchanged with our patch-based contrastive loss. We also test a variant without this loss, as depicted in Table 3. It comes out that the one with-

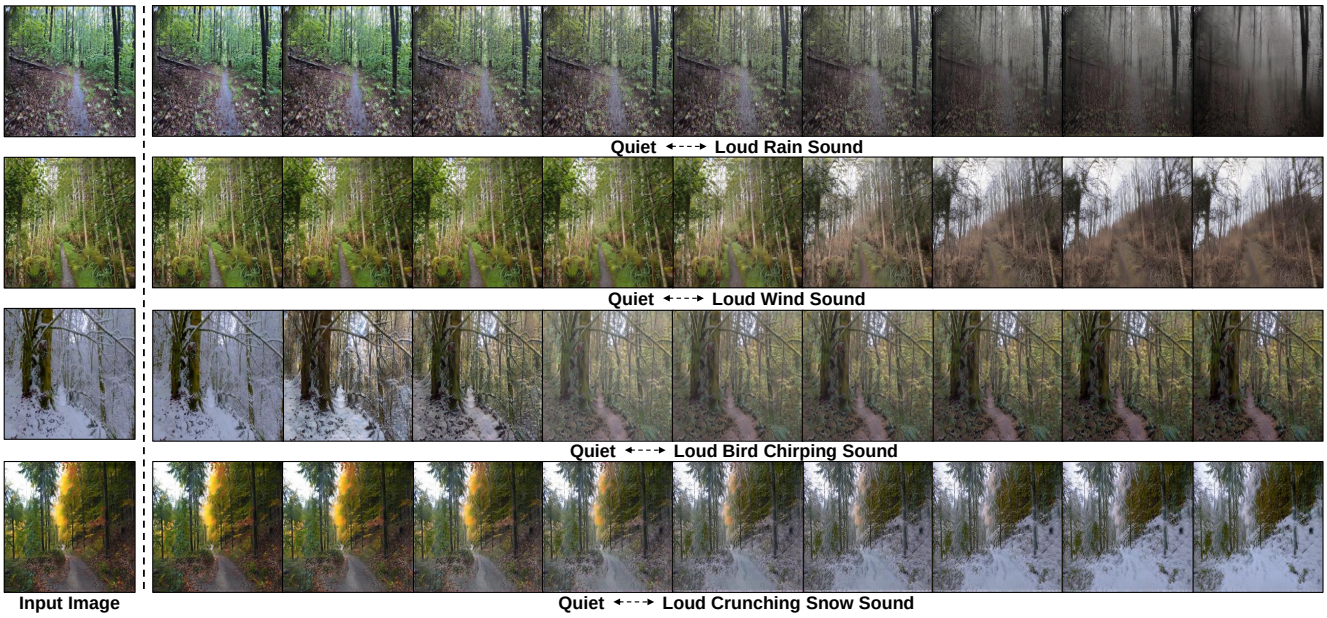


Figure 6. Qualitative results on image manipulation with increasing sound volumes.

Table 3. Quantitative results for ablations on *Into the Wild* dataset.

Ablation	Objective Evaluation		
	AVC (↑)	FID (↓)	CLIP (↑)
CycleGAN [64]	0.812	35.244	0.232
Late Fusion [54]	0.811	54.025	0.230
w/o ID Loss	0.810	41.019	0.236
Ours	0.820	34.139	0.238
+ Pre-training [3]	**0.822**	**32.882**	**0.242**

out identity loss tends to have worse performance. We further investigate by presenting qualitative results in Figure 5. In the first row of the second example, in particular, when the conversion is from sunny to rainy forest, it is unsuccessful for the one without identity loss, whilst the one with succeeds. As a result, we propose that employing such a loss as a regularizer might be beneficial in capturing nuances, particularly when converting between similar landscapes, such as forest-to-forest and snow-to-snow conversions.

Self-supervised pre-training improves stylization We ask whether models pre-trained to solve audio-visual self-supervised learning tasks will result in performance gains. Table 3 shows that fine-tuning our task using a pre-trained SeLaVi model [3] yields a small improvement.

4.4. Audio Manipulation for Image Manipulation

Sound provides a natural "embedding space" for image manipulation, since intuitively manipulating the audio leads to corresponding changes in the images. We ask whether changing the volume of the sound or mixing two sounds together will result in corresponding visual changes. We also evaluate out-of-distribution images and audio.

Changing sound volumes A qualitative comparison using a sound at various volumes is shown in Figure 6. This is accomplished by simply rescaling the input waveform. Regardless of whether the input image is snowy or sunny forest, we observe that the texture in the image becomes more prominent as the sound gradually increases, indicating that our model implicitly learns to predict the prominence of the texture according to the volume.

Mixing audios We create sound mixtures by taking convex combinations of input sounds. The qualitative results are presented in Figure 7. In the third row, for example, we can see that the snowy texture will be gradually erased while mixing a crunching snow sound with a muddy footstep sound from small to large. Furthermore, it appears to be a balanced state with both snowy and sunny features in the middle, *i.e.*, white and green hues coexist. Surprisingly, such mixed audio is not available when our model is being trained. This linear additivity finding shows that audio cues have a prospective advantage over label ones for image translation.

Generalization experiment We ask whether our model can generalize to out-of-distribution data. We consider restyling images from the Places dataset [61] and audio from the VGG-Sound dataset [6] to examine our model's generalization performance. In Figure 8, we use crunching snow and rain sounds with a high probability of a class deduced by YAMNet [48]. Our model generates plausible images that match the content of the audio.

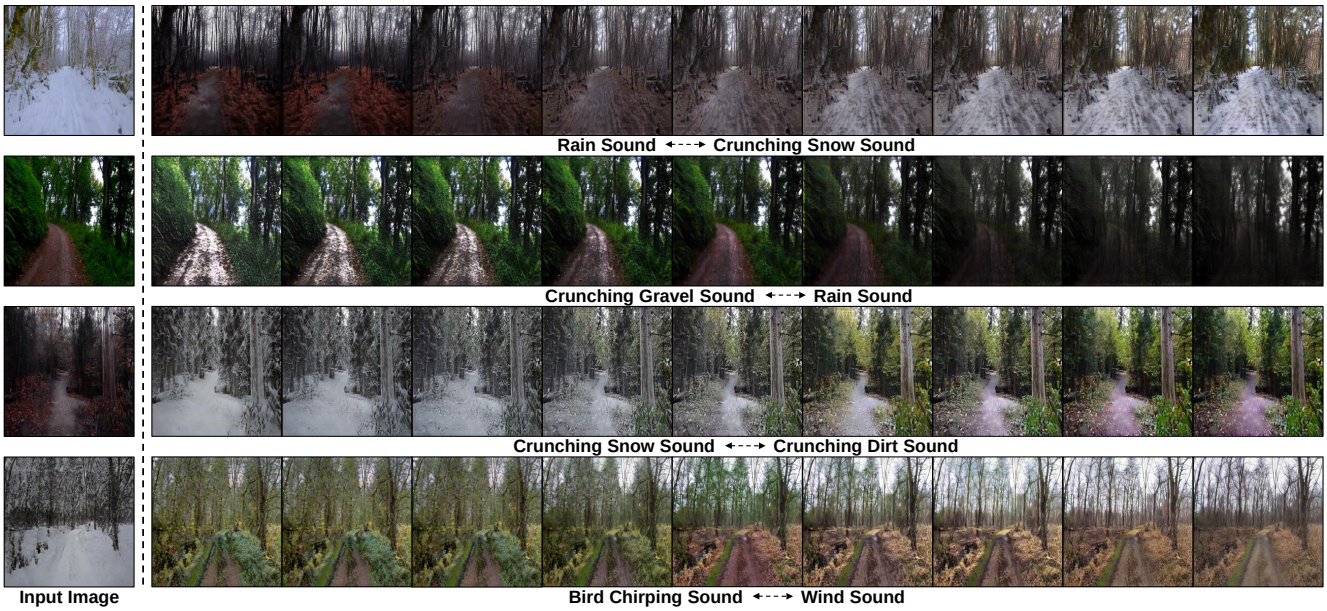


Figure 7. Qualitative results on image manipulation with different mixture sounds.

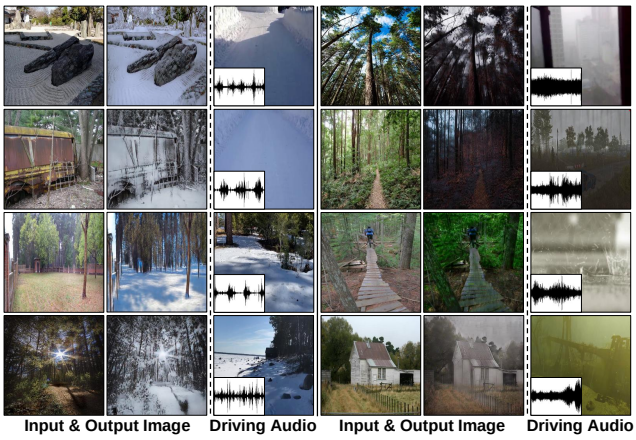


Figure 8. **Qualitative generalization results.** We restyle images from Places [61] using crunching snow and rain sounds taken from VGG-Sound [6].

5. Discussion and Limitations

Despite the fact that our model can yield promising results in many cases, the results are far from uniformly positive. Since ambient sounds in real life are diverse, our model can be easily upset with unexpected sounds. Figure 9 shows some typical failure cases. Specifically, if the sound is interfered by human speech, the learned translation will devolve to making minor adjustments to the input. As a result, handling a greater spectrum of mixture sound, particularly urban sound, will become increasingly important in the future. Another potential concern is that our model's performance will be suffered if the proportion of the scene to be converted is too small. In the lower right of Figure 9, for example, the trees and sky each account for half of the input image, resulting in an odd conversion. This is because the model is unable to detect the region of the scene that needs conversion, but instead converts the entire scene. Nevertheless, as paired audio-visual data is ubiquitous in our daily life, this paper paves the way for image translation under the audio-visual context.

Figure 9. **Some failure cases of our model.** Left: fail to learn visual scenes due to the acoustic interference of human speech. Right: fail to detect or recognize objects in some visual scenes.

6. Conclusion

In this paper, we introduce a novel task called *audio-driven image stylization*, which aims to convert the texture of a visual scene conditioned on sound. To study this task, we propose a contrastive-based audio-visual GAN model, together with an unlabeled egocentric hiking dataset named *Into the Wild*. Experimental results show that our model outperforms label-based counterparts in both quantitative and qualitative evaluations. We also empirically find that changing the audio volume and mixture results in predictable visual changes. We hope our work will shed new light on cross-modal image synthesis.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 2, 5, 12

[3] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Advances in Neural Information Processing Systems*, 2020. 6, 7, 13

[4] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. In *arXiv:2103.10951*, 2021. 1, 2

[5] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2, 7, 8, 13

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Everest Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020. 3

[8] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. In *5th Annual Conference on Robot Learning*, 2021. 2

[9] Kin Wai Cheuk, Hans Anderson, Kat Agres, and Dorien Herremans. nnAudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks. *IEEE Access*, 8:161981–162003, 2020. 4

[10] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *British Machine Vision Conference*, 2017. 2

[11] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE, 2019. 5, 12

[12] Virginia R de Sa. Learning classification with unlabeled data. In *Advances in neural information processing systems*, pages 112–119. Citeseer, 1994. 2

[13] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017. 2

[14] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4), 2016. 2

[15] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven image style transfer. *arXiv preprint arXiv:2106.00178*, 2021. 2

[16] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 2

[17] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 1

[19] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 2, 4, 5

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2, 3

[21] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010. 2, 3

[22] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665, 2018. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 12

[24] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 5

[25] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 1

[26] 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*, 2017. 5, 12

[27] Chenxu Hu, Qiao Tian, Tingle Li, Yuping Wang, Yuxuan Wang, and Hang Zhao. Neural dubber: Dubbing for silent

videos according to scripts. In *Advances in neural information processing systems*, 2021. 2

[28] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 4, 12

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 4, 12

[31] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 1

[32] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017. 2

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015. 4, 12

[34] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of the Advances in Neural Information Processing Systems*, 2018. 2

[35] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on graphics (TOG)*, 33(4):1–11, 2014. 2

[36] Timothy R Langlois and Doug L James. Inverse-foley animation: Synchronizing rigid-body motions to sound. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 2

[37] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 2

[38] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 12

[39] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 5

[40] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*, 2018. 2

[41] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 2

[42] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in neural information processing systems*, 2018. 2

[43] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 2

[44] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[45] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 2, 4

[46] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345, 2020. 2, 3, 6, 12

[47] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021. 12

[48] Manoj Plakal and Daniel Ellis. YAMNet. Jan 2020 [Online]. Available: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet. 2, 4, 5, 7, 12

[49] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805, 2020. 2

[50] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 2

[51] 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*, 2021. 2, 5, 12, 13

[52] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069, 2016. 1, 2

[53] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 12

[54] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12695–12705, 2020. 6, 7

[55] Chenyun Wu, Mikayla Timm, and Subhransu Maji. Describing textures using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 52–70. Springer, 2020. 2

[56] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. *arXiv preprint arXiv:2110.11499*, 2021. 13

[57] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 2

[58] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[59] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 2

[60] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2

[61] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 7, 8, 13

[62] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9299–9306, 2019. 2

[63] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3558, 2018. 2

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2223–2232, 2017. 2, 6, 7, 12

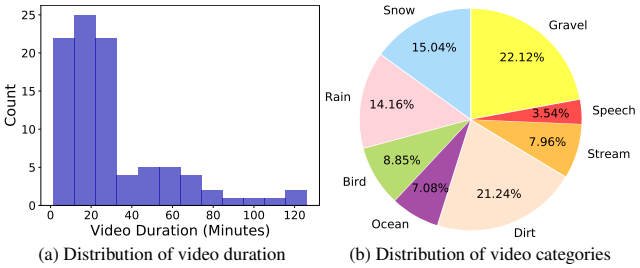
A. Appendix

A.1. *Into the Wild* dataset

Considering hiking videos are likely to contain a strong audio-visual association of nature, we introduce the *Into the Wild* dataset, an egocentric hiking dataset for our proposed audio-driven image stylization (ADIS).

We collected these videos on YouTube by searching for keywords like hike+binaural, hike+footsteps and hike+ASMR, hike+POV. We employ YAMNet [48] to tag each associated audio after downloading them to ensure that they play the actual sound and are not disrupted by any background music.

The duration statistics are shown in Figure 10a. Specifically, *Into the Wild* contains 94 untrimmed videos, some of which are already presented in Figure 4 of the main paper. Please note that the category labels of these videos are not labeled by humans, but acquired from the YAMNet [48] predictions, which roughly consist of 8 categories: crunching snow, gravel, and dirt; rain; birds chirping; ocean; stream and human speech. The detailed categorical distribution is illustrated in Figure 10b.



(a) Distribution of video duration (b) Distribution of video categories

Figure 10. Statistical analysis of the *Into the Wild* Dataset.

Figure 11. A screenshot of AMT for rating the audio-visual correspondence.

A.2. Training Details

Except for the batch size and audio network, we intentionally match the architecture and hyperparameter settings with CycleGAN [64] and CUT [46]. We employ ResNet-based generator [30] with 9 residual blocks, PatchGAN discriminator [29], Least Square GAN loss [38], ResNet18-based audio encoder [23], with the batch size of 16, and the Adam optimizer [33] with 0.002 learning rate. Both λ and μ in Eq.(4) of the main paper are set to 0.5.

Our model is trained for 50 epochs, with the learning rate remaining constant for the first 30 epochs and linearly decaying to zero over the last 20 epochs. The encoder G_{enc} follows the first half of the CycleGAN generator [64]. We also extract features from 5 different scales to calculate the patch-based structure discriminator loss: the input RGB pixels, the first and second downsampling convolution features, and the first and fifth residual block features. We sample 256 random locations for each layer's features and apply a 2-layer MLP to obtain 256-dimension features as the final output for computing the multi-scale patch-wise contrastive loss.

***Into the Wild* dataset** We divide all of the videos into 3-seconds video clips, then uniformly sample 8 frames from each video clip to save as images, yielding a total of 454560 images and 56820 audios. We then randomly sample 20% audios as the test set.

The *Greatest Hits* dataset We first identify the videos by the type of object being hit on, and then only the outdoor videos are used for training: dirt, grass, gravel, leaf, and water, resulting in a total of 32172 images and 8043 audios. We then select 15% audios at random as the test set.

A.3. Evaluation Details

Audio-visual Correspondence (AVC) A two-stream network is utilized to compute AVC [2], with one stream extracting audio feature and the other extracting visual feature. Specifically, we apply OpenL3 [11] to obtain these features, and then compute the average cosine similarity for each image-audio pair. To be more explicit, we employ an "env" content type pre-trained model with 512-dimensional linear spectrogram representation.

Fréchet Inception Distance (FID) FID [26] is calculated by scaling the images to 299-by-299 using the PyTorch framework's bi-linear sampling, and then take the activation of the last average pooling layer of a pre-trained Inception V3 [53]. We adopt Clean-FID [47] to circumvent the issue that FID computation requires complicated and error-prone steps, such as the resizing functions in different libraries often produce inaccurate results.

Contrastive Language-Image Pre-Training (CLIP) [51] is computed by performing contrastive pre-training on a variety of image-text pairs. It's widely known for zero-shot prediction, but we use it as a feature extractor to compute the cosine similarity between images and labels in order to assess conversion quality. To calculate it, we leverage an off-the-shelf "ViT-B/32" CLIP model [51].

Amazon Mechanical Turk (AMT) In addition to the objective evaluations mentioned above, we employ AMT to

Table 4. Quantitative comparison for different pre-training methods on the *Into the Wild* dataset.

Pre-training Method	Objective Evaluation		
	AVC (↑)	FID (↓)	CLIP (↑)
Ours (from scratch)	0.820	34.139	0.238
+ SeLaVi [3]	0.822	32.882	0.242
+ Wav2CLIP [56]	**0.831**	**30.334**	**0.246**

study the relationship between audio and visual from a subjective standpoint, *i.e.*, human perspective. A screenshot of the demo page is shown in Figure 11. The MTurker is required to rank such correlations based on audios and images generated by our method and the baseline methods, with the best earning 3 points and the worst earning 1 point. Thus, the scores range from 1 to 3. Notably, only one Mturker was asked to rank 2000 random samples from the test set in our case. The final scores are reported on average.

A.4. Additional Results

Additional qualitative comparisons Additional qualitative comparisons on our method to the baselines and ablations are shown in Figure 12. It turns out that our model produces better or competitive results, exhibiting its versatility compared to label-based baselines.

Additional generalization results Additional qualitative results of the generalization experiment are shown in Figure 13. These are accomplished by using images from the Places dataset [61] and the audios from the VGG-Sound dataset [6]. our model generates plausible images that match the content of the audio, even though they have never been seen before.

Additional pre-training comparisons We also use Wav2CLIP [56], an audio representation learning method derived on CLIP [51], to fine-tune ADIS. It employs a frozen image model to bridge the gap between a sophisticated language model and an audio model, in order to impart information. Wav2CLIP may be a better audio representation for ADIS than SeLaVi [3] since it is implicitly exposed to numerous well-annotated image-text pairs. Table 4 shows the quantitative comparison results. It appears that Wav2CLIP surpasses both training from scratch and SeLaVi pre-training methods with respect to the AVC, FID, and CLIP metrics, indicating that it has a stronger representation ability than the others.

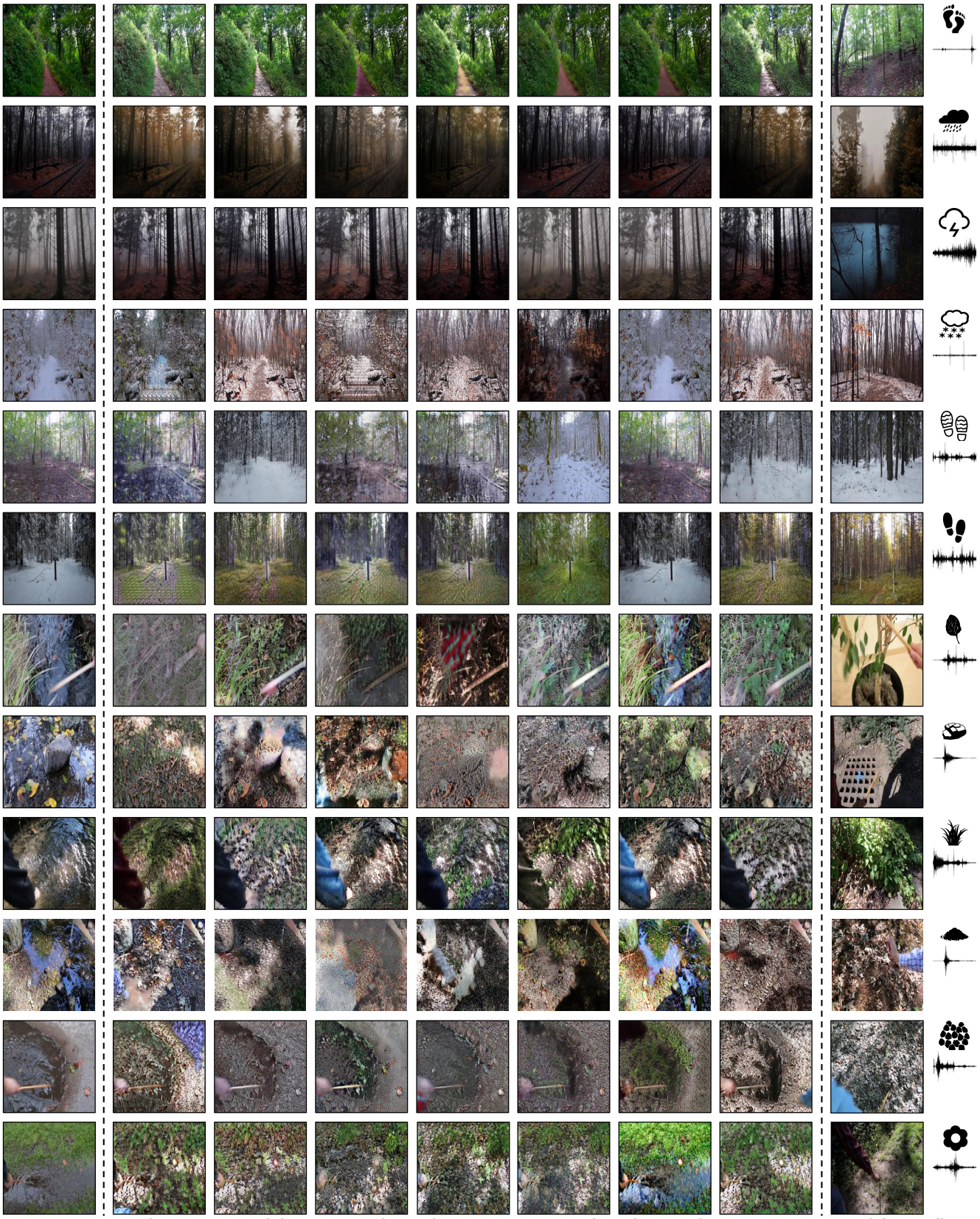


Figure 12. Randomly selected qualitative results of our model, baselines and ablations. This is an extension of Figure 5 in the main paper.

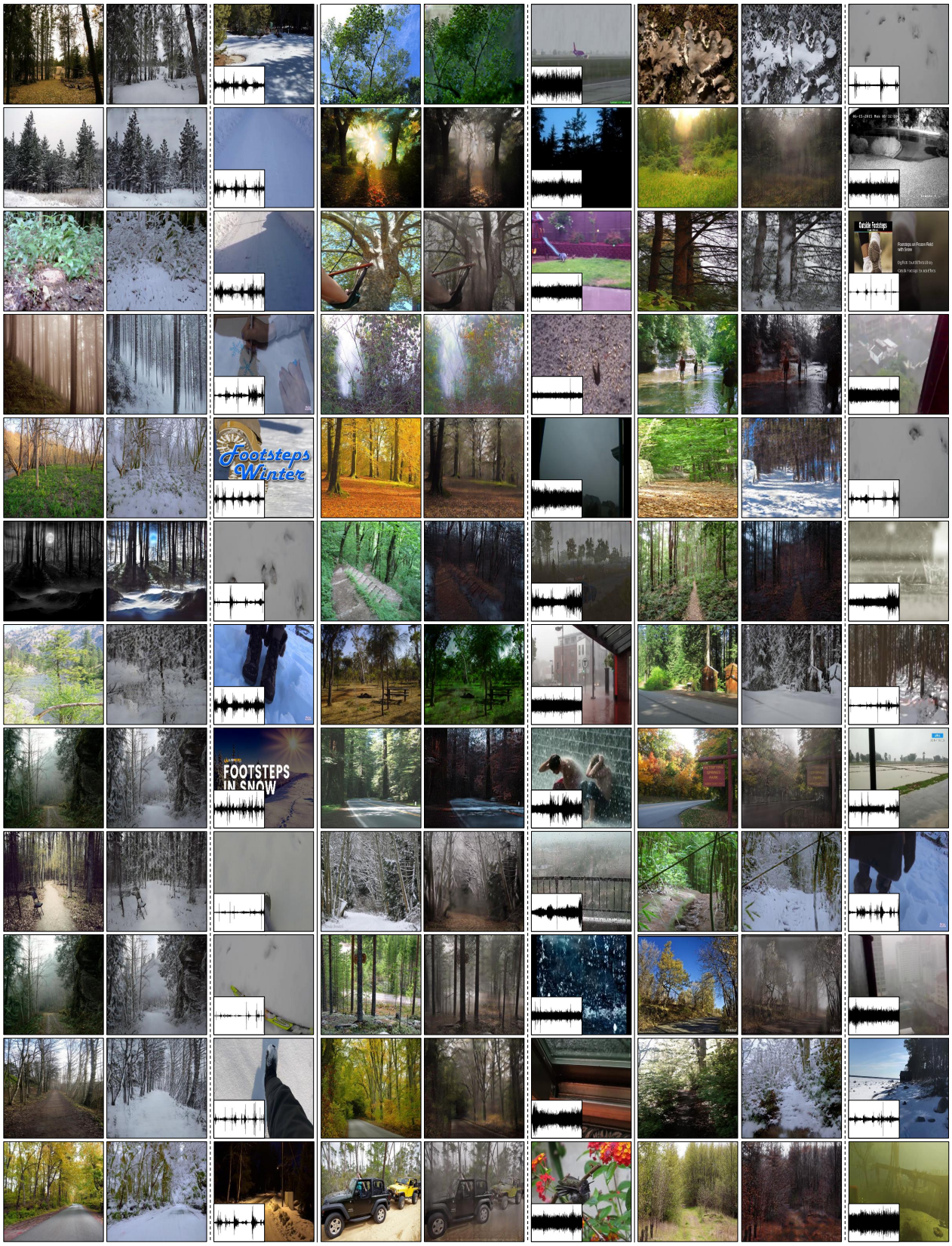


Figure 13. Randomly selected qualitative results of generalization experiment. This is an extension of Figure 8 in the main paper.