Learning Visual Styles from Audio-Visual Associations

Tingle Li[1,3], Yichen Liu[1],
Andrew Owens[2], and Hang Zhao[1,3]

[1]Tsinghua University [2]University of Michigan [3]Shanghai Qi Zhi Institute
https://tinglok.netlify.com/files/avstyle

Abstract. From the patter of rain to the crunch of snow, the sounds we hear often convey the visual textures that appear within a scene. We present a method for learning visual styles from unlabeled audio-visual data. Our model learns to manipulate the texture of a scene to match a sound, a problem we term *audio-driven image stylization*. Given a dataset of paired audio-visual data, we learn to modify input images such that, after manipulation, they are more likely to co-occur with a given input sound. In quantitative and qualitative evaluations, our sound-based model outperforms label-based approaches. We also show that audio can be an intuitive representation for manipulating images, as mixing or adjusting a sound's volume leads to predictable changes to visual style.

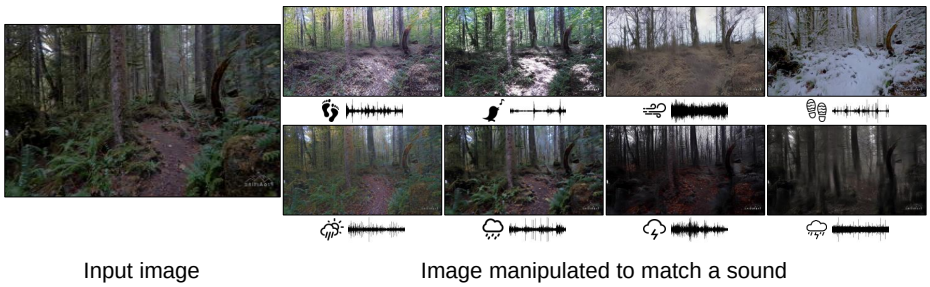


Fig. 1: **Audio-driven image stylization.** We manipulate the style of an image to match a sound. After training with an unlabeled dataset of egocentric hiking videos, our model learns visual styles for a variety of ambient sounds, such as light and heavy rain, as well as physical interactions, such as footsteps.

1 Introduction

Recent work has proposed a variety of methods for manipulating the style [62,19] of an input image. In these methods, the desired style is specified using other example images [27,19,33,30] and, more recently, through human language, such

as through semantic labels, text, or scene graphs [58,34,4,57]. While this approach has been effective, it requires human-provided annotations and hence implicitly relies on a "human in the loop" for supervision. This supervision is often expensive to collect and may fail to capture important scene properties.

We propose to address these problems by learning stylization from unlabeled *audio-visual* data. Many scene properties, such as weather conditions, produce highly distinctive sights and sounds. Training a model to estimate visual information from audio requires it to identify these scene structures and, through the process, learn distinctive visual textures that are associated with a sound. Inspired by this idea, we introduce a model for performing *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 audio-visual style associations, each of which can be specified by a sound — bird chirps and blue skies, crunching footsteps and snow, rain and dark skies, etc. (Figure 1).

Audio naturally comes paired with visual data, and thus provides a free learning signal, complementing human-provided supervision like labels and text. It also conveys important distinctions between scenes that often may not be evident in pre-existing text or label sets. For example, asking a model to generate images depicting a "rainy" scene can be ambiguous. Providing the sound of rain, on the other hand, specifies whether the rain is light or heavy, as well as whether the image is likely to contain dark, stormy skies. Finally, audio can be used as a natural representation for specifying image styles, as intuitive changes to the audio, such as adjusting the volume or mixing two sounds together, result in predictable visual changes.

Based on this idea, we propose a method that combines conditional generative adversarial networks [22] and contrastive learning [23]. Our model uses 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. It also includes a multi-scale patch-wise structure discriminator [51] that maximizes the mutual information between the source and generated images in order to preserve the structural content of the scene. We train the model 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. 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. Through quantitative and qualitative evaluations, we show:

- Unlabeled audio provides supervision for learning visual styles.
- Our proposed model learns to perform audio-driven stylization from in-the-wild audio-visual data.

- Adjusting the volume of a sound or mixing it with other sounds lead to predictable changes in image style.

2 Related Work

Image translation. Paired image translation [32] frames the image prediction problem as a straightforward supervised learning task, which corresponding input and target images. Unpaired image translation [22,74,67,35,51] learns to transform images between two different domains, without ground-truth correspondences. We take inspiration from work [38] that manipulates the global appearance of a scene, such as through labels indicating the desired weather. A variety of methods [58,13,47,4,15,57] have been proposed to generate or manipulate plausible images from text. However, text- and label-based methods either require "weak" supervision from humans (*e.g.*, paired text and images) [42] or explicit image descriptions (*e.g.*, text describing an image as a line drawing) [64], which may not capture the full range of styles. Our approach, by contrast, uses audio to learn styles, without any form of human labeling. It therefore provides a *complementary* learning signal to text and labels.

Audio-visual correspondence. Audio and visual signals naturally co-occur when they are recorded as video. In order to leverage this natural correspondence, researchers have introduced various tasks, such as representation learning [59,48,2,37,49,46], source separation [70,69,14,17], audio source grounding [6,24], audio spatialization [18,45,66], visual speech recognition [1], and scene classification [7,20]. 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. A variety of methods have been proposed for synthesizing images from sound or vice versa. One line of work has generated sounds from video, such as impact sounds [50,68], natural sounds [73,31], or human speech [29,54]. Another line of work has created models that synthesize images from sound, such as by generating talking heads [11,72,55], pose [60,16,41,21], synchronizing rigid body animations with contact sounds [39], estimating depth from ambient sound [9], predicting future video frames [5]. Unlike these works, we concentrate on restyling plausible images using the source image and natural sounds. In concurrent work, Lee et al. [40] used sound to guide a text-based image manipulation method based on CLIP [56]. In contrast, our model learns image styles solely from unlabeled audio-visual data.

3 Audio-driven Image Stylization

We take inspiration from the fact that audio can convey distinctions that may not be obvious from semantic categories. For example, consider the images shown in Figure 2. While these videos have the same category (*e.g.*, rain), their visual

style significantly varies (*e.g.*, heavy or light rain). This distinction, however, is easily captured by the corresponding sound. We propose *audio-driven* image stylization (ADIS) as a novel multi-modal generation task for learning these styles.

We pose this problem as learning a mapping from a source image domain $\mathcal{X}$ to a target domain $\mathcal{Y}$ using an input sound from the audio domain $\mathcal{A}$. To achieve this goal, we propose a self-supervised learning approach that can be trained on unpaired videos. This can be accomplished through two distinct training objectives.

Fig. 2: Categories can fail to convey subtle distinctions between events. We show frames whose corresponding sounds were classified as *footstep* or *rain* [53,20].

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 training scheme, the generated image is encouraged to match the target audio. 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 [22] 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. In our model, D performs 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) [23], which aims to preserve structure information by establishing mutual correspondence between

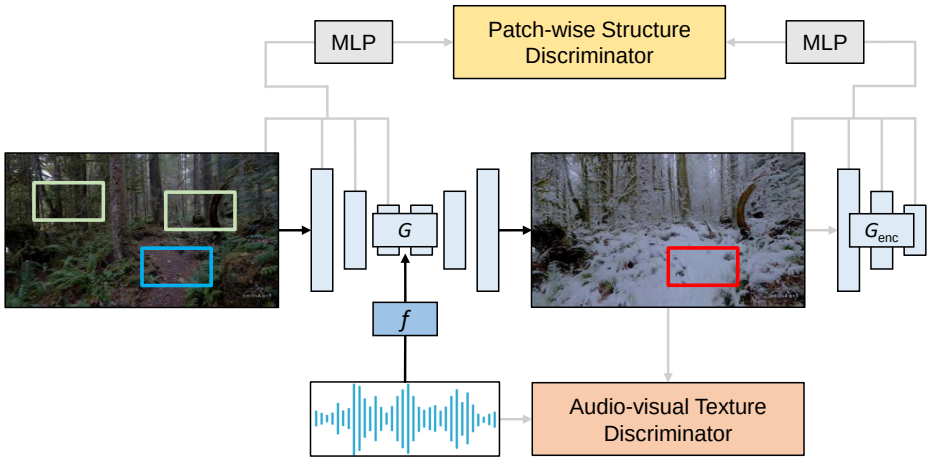


Fig. 3: **Model architecture**. The multi-scale patch-wise structure discriminator [51] is used to preserve the scene structure, while the audio-visual texture discriminator is used to convert the scene texture. This is an example where sunny forest is converted to snowy counterpart. The **generated snow patch** should match its corresponding **input dirt patch**, in comparison to **other random patches**. Note that the MLP component is not used during inference.

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 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 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:

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

where $\boldsymbol{v}_n^-$ denotes the n-th negative sample and τ is a temperature parameter, as suggested in SimCLR [8], 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 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 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 $\boldsymbol{v}^{(N+1)\backslash n}_l \in \mathbb{R}^{N\times M}$ 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:

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

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:

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

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 former provides impact sounds from different materials, while the latter is a new dataset of egocentric hiking videos.

- ***Into the Wild* dataset**: We collect a new dataset to study the audio-visual associations that one would encounter on a hike (Figure 4). These include sounds that are related to seasonal variations, rainfall, animal vocalizations, and footsteps. We collect 94 untrimmed egocentric videos from YouTube, ranging from 1.5 to 130 minutes long (50 hours in total). We chose videos that only contain sounds naturally present in the scene (*e.g.*, no background music). See Appendix A.1 for more dataset details.

- **The *Greatest Hits* dataset** [50]: 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 ADIS. 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 dirt, water, gravel and grass.

Fig. 4: **Selected frames from the *Into the Wild* dataset.** We show example images corresponding to the top-1 categorical sounds deduced by a classifier [53,20].

Network architecture. The encoder and decoder of the GAN generator are 2D fully convolutional networks, with 9 layers of ResNet-based CNN bottlenecks [33] in between. Except for the first CNN layer with a kernel size of 7×7, the others are 3×3, and the stride size is determined by whether downsampling is required. We used the PatchGAN architecture [32] for the discriminator. A ResNet18 backbone [25] 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 512×512 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 [10] 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 hyperparameters, both λ and μ in Eq.(4) are set to 0.5. We also employ random crop and horizontal flip as data augmentation. Our model is trained using the Adam optimizer [36] with a batch size of 16 and an initial learning rate of 2×10^{-4} over 50 epochs. Other training strategies are described 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 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 [12], a variant of L3-Net [2] pre-trained on AudioSet [20], 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)** [28]: 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.
- **Amazon Mechanical Turk (AMT)**: We use human participants to evaluate the audio-visual correlations (*i.e.*, via a subjective evaluation). Each participant is asked to rank the quality of the correlation between a sound and the images generated by various methods. The scores range from 1 (indicating low correlation) to 4 (high correlation).
- **Contrastive Language-Image Pretraining (CLIP)** [56]: CLIP is a network trained using contrastive learning to associate corresponding image and text pairs. In order to provide an additional evaluation metric that captures semantics, we use the keywords from the title of each video as text inputs to CLIP, then measure the text-image similarity. A higher CLIP score indicates a better correlation between a given text and image.

Baselines. We adopt two label-based methods for comparison. For both of them, Word2Vec [44] is used for generating the class embeddings, which is incorporated with the input image and serves as a textual condition. In addition, we create an image-conditioned baseline.

- **Class Pred.** [53]: we use YAMNet, a state-of-the-art audio classification network [26] trained on AudioSet [20], 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 title, thereby conveying the information provided in the video metadata.
- **AdaIN** [30]: AdaIN is an image-conditioned arbitrary stylization method that incorporates the adaptive instance normalization to fuse the content image and the style one. It takes two images as input and restyles one to match the other. Note that the style image is picked at random from the video frames corresponding to the selected audio.

4.2 Comparison to Baselines

Quantitative results. Since the diverse hitting and scratching sounds are not well-modeled by AudioSet [20], which L3-Net [2] is trained on, we cannot meaningfully evaluate the *Greatest Hits* with the AVC metric. As a result, we only provide quantitative results yielded from the *Into the Wild* dataset. Table 1 shows the quantitative comparisons between our model and label/image-conditioned baselines. For objective evaluation, our model outperforms three baselines across the AVC, FID, and CLIP metrics, suggesting that our model can generate more

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 (↓)	AMT (↑)	CLIP (↑)
Target	0.842	/	/	0.247
Class Pred. [53]	0.801	91.417	1.833 ± 0.042	0.228
AdaIN [30]	0.812	62.851	2.269 ± 0.044	0.232
Keyword	0.809	38.066	2.626 ± 0.045	0.236
Ours	**0.820**	**34.139**	**3.273 ± 0.046**	**0.238**

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-to-Rainy	Snowy-to-Sunny	Sunny-to-Snowy
Class Pred. [53]	0.819	0.796	0.793
Keyword	0.827	0.802	0.808
Ours	**0.831**	**0.820**	**0.816**

realistic images. In particular, our method outperforms AdaIN [30], despite the fact that AdaIN has already been pre-trained using ImageNet while ours is trained from scratch. We find that Keyword outperforms Class Pred., perhaps due to errors introduced by automatic labeling. Notably, Class Pred. contains 132 label classes from AudioSet, whereas Keyword only has 3 classes (sunny, snowy and rainy), which are all closely related to the scenes in *Into the Wild*. We also observe that the CLIP metric for our model is on par with Keyword, which also indicates the benefit of using audio over labels. For human evaluation, we randomly select 1000 images from the test set, and ask participants to assess the level of the audio-visual correlation. It turns out that they consistently preferred our model's results, as shown in the penultimate column of Table 1, which is consistent with the objective evaluation results.

To gain a better understanding of our model's performance, we divide the entire test set into three categories: sunny, rainy, and snowy and report results on each subset. In this experiment, as shown in Table 2, our model still holds the best performance compared to label-based baselines. Furthermore, we observe that when the target scene is sunny, the disparity between our model and Keyword (0.018) is larger than that of other scenes (0.004 & 0.008). This may be because the ambient sounds in sunny forests are highly varied (*e.g.*, crunching gravel/leave, birds chirping, *etc.*).

Qualitative results. We show qualitative results in Figure 5 and provide additional results in the Appendix A.4. We note that all of the results are produced

Fig. 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.

by a single model, *i.e.*, through "one-to-many" conversion. We observe that the AdaIN model sometimes cannot reliably preserve the input image's content (the first row of first input image). The Keyword model can generate plausible images that match the class of the target audio, but with apparent flaws when converting between the same scene categories (the second row of the second input image). For the YAMNet model, the generated images occasionally match the target images, but this does not happen in all cases. This may be because the success of a stylization is strongly dependent on whether the labels inferred by YAMNet are correct. Our model, by comparison, can capture the subtle distinctions within the same scene class. For example, our model can adjust the hue of the snow, when given a wind-and-footstep sound (which is not successfully captured by other models).

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 [74] instead of NCE loss, termed as CycleGAN; ii) late fusion discriminator [63] 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

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

Ablation	**Objective Evaluation**		
	AVC (↑)	FID (↓)	CLIP (↑)
CycleGAN [74]	0.812	35.244	0.232
Late Fusion [63]	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**

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 [51]. As a baseline, cycle-consistency loss [74] 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 [51].

Late fusion discriminators are more likely to collapse. In audio-visual learning, the late fusion architecture [63] is commonly used, in which two unimodal 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 [74] 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. We find that the variation of the model without identity loss tends to has 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.

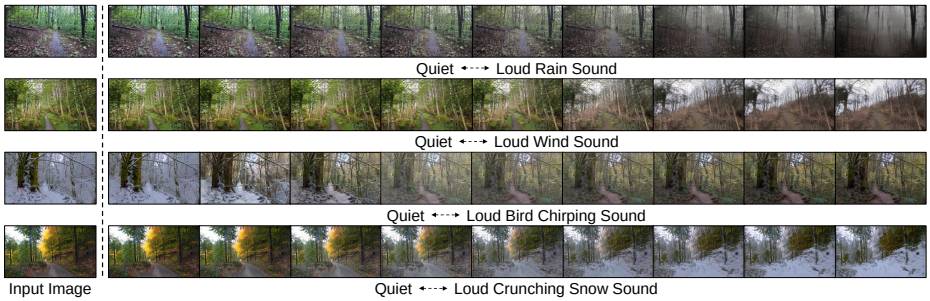


Fig. 6: Qualitative results on image manipulation with increasing sound volumes.

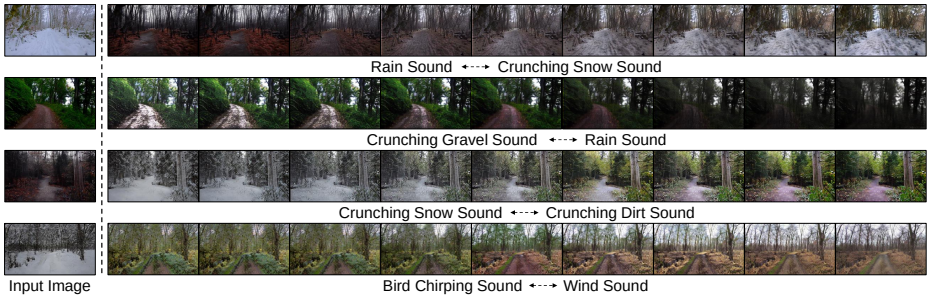


Fig. 7: Qualitative results on image manipulation with different mixture sounds.

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 sounds. 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

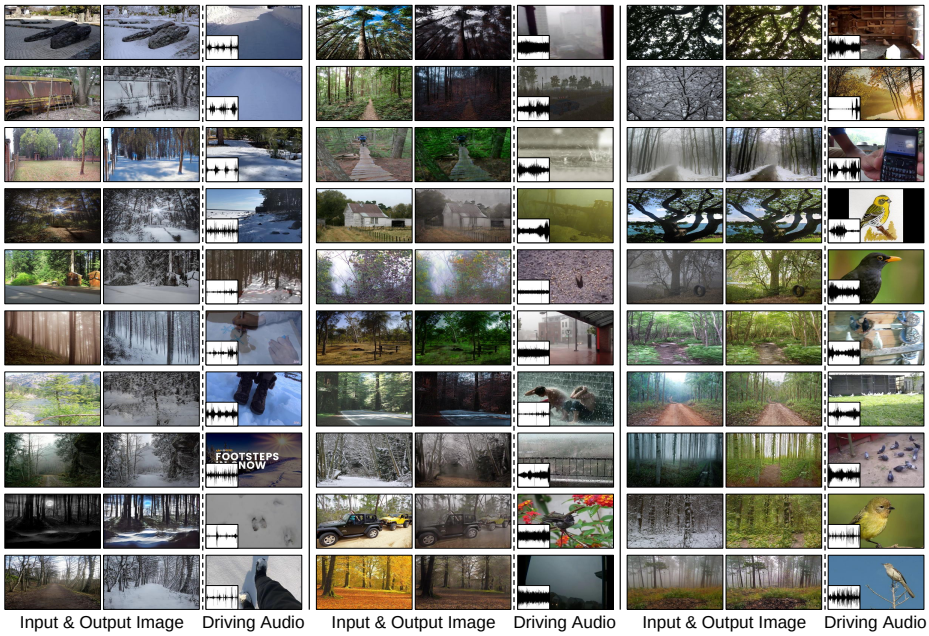


Fig. 8: **Qualitative generalization results.** We restyle images from Places [71] using crunching snow and rain sounds taken from VGG-Sound [7].

additivity finding shows that audio cues have a prospective advantage over label ones for image translation.

Generalization to other datasets. We ask whether our model can generalize to out-of-distribution data. We consider restyling images from the Places dataset [71] and audio from the VGG-Sound dataset [7] to examine our model's generalization performance. In Figure 8, we use crunching snow, rain and birds chirping sounds with a high probability of a class deduced by YAMNet [53]. Our model generates plausible images that match the content of in-the-wild audio.

Adjusting an image's style through its sound. We apply our method to a task inspired by video editing: adjusting an image's appearance by manipulating its *existing* sound. We take a video frame, manipulate its corresponding sound, and then resynthesize its video frames to match. This allows a user to make *consistent* changes to the two modal-

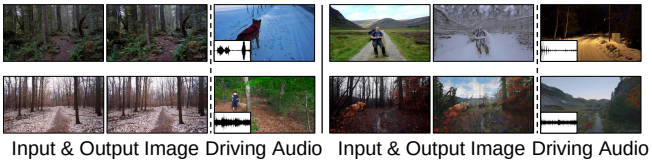


Fig. 9: **Failure cases.** Our model fails to manipulate the style of the scene, perhaps due to the presence of speech in the sound (left). It also fails to learn how to style certain objects in a scene (right).

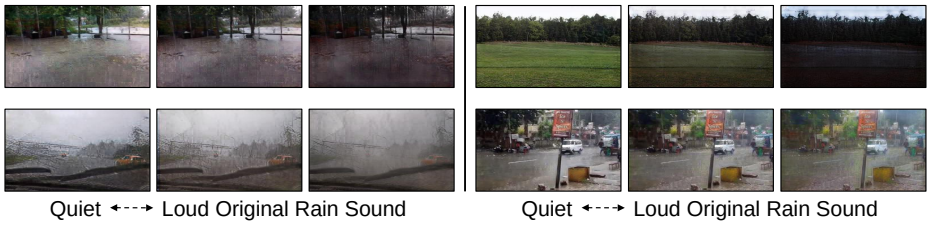


Fig. 10: **Restyling with a video's existing sound.** We adjust the appearance of a video by increasing the volume of its soundtrack, and restyling the corresponding video frame.

ities: *e.g.*, an editor can adjust the volume of rain through intuitive volume-based controls, while automatically propagating these changes to images.

We restyle videos from VGG-Sound [7] by adjusting the volume of their already-existing soundtracks. Figure 10 shows qualitative examples obtained by increasing the volume of videos recorded during light rain. As expected, the resulting images contains significantly more rain.

5 Discussion and Limitations

Despite the fact that our model can yield promising results in various cases, the results are far from uniformly positive. Because 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.

6 Conclusion

In this paper, we introduce a novel task called *audio-driven image stylization*, which aims to learn the visual styles from paired audio-visual data. 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 and image conditioned baselines 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. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence (2018) 3
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017) 3, 8, 21
3. Asano, Y.M., Patrick, M., Rupprecht, C., Vedaldi, A.: Labelling unlabelled videos from scratch with multi-modal self-supervision. In: Advances in Neural Information Processing Systems (2020) 10, 11, 22
4. Bau, D., Andonian, A., Cui, A., Park, Y., Jahanian, A., Oliva, A., Torralba, A.: Paint by word. In: arXiv:2103.10951 (2021) 2, 3
5. Chatterjee, M., Cherian, A.: Sound2sight: Generating visual dynamics from sound and context. In: European Conference on Computer Vision. pp. 701–719. Springer (2020) 3
6. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3
7. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020) 3, 13, 14, 22
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607 (2020) 5
9. Chen, Z., Hu, X., Owens, A.: Structure from silence: Learning scene structure from ambient sound. In: 5th Annual Conference on Robot Learning (2021) 3
10. Cheuk, K.W., Anderson, H., Agres, K., Herremans, D.: nnAudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks. IEEE Access **8**, 161981–162003 (2020) 7
11. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? In: British Machine Vision Conference (2017) 3
12. Cramer, J., Wu, H.H., Salamon, J., Bello, J.P.: 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). pp. 3852–3856. IEEE (2019) 8, 21
13. Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic image synthesis via adversarial learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5706–5714 (2017) 3
14. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics (TOG) **37**(4) (2016) 3
15. Fu, T.J., Wang, X.E., Wang, W.Y.: Language-driven image style transfer. arXiv preprint arXiv:2106.00178 (2021) 3
16. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10478–10487 (2020) 3
17. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 35–53 (2018) 3

18. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 324–333 (2019) 3
19. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015) 1
20. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017) 3, 4, 7, 8
21. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3497–3506 (2019) 3
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014) 2, 3, 4
23. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 297–304 (2010) 2, 4
24. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J.: Jointly discovering visual objects and spoken words from raw sensory input. In: Proceedings of the European conference on computer vision (ECCV). pp. 649–665 (2018) 3
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 7, 20
26. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp). pp. 131–135. IEEE (2017) 8
27. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 327–340 (2001) 1
28. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (2017) 8, 21
29. Hu, C., Tian, Q., Li, T., Wang, Y., Wang, Y., Zhao, H.: Neural dubber: Dubbing for silent videos according to scripts. In: Advances in neural information processing systems (2021) 3
30. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017) 1, 8, 9
31. Iashin, V., Rahtu, E.: Taming visually guided sound generation. arXiv preprint arXiv:2110.08791 (2021) 3
32. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 3, 7, 20
33. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 1, 7, 20

34. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1219–1228 (2018) 2
35. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning. pp. 1857–1865. PMLR (2017) 3
36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference for Learning Representations (2015) 7, 20
37. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Proceedings of the Advances in Neural Information Processing Systems (2018) 3
38. Laffont, P.Y., Ren, Z., Tao, X., Qian, C., Hays, J.: Transient attributes for high-level understanding and editing of outdoor scenes. ACM Transactions on graphics (TOG) **33**(4), 1–11 (2014) 3
39. Langlois, T.R., James, D.L.: Inverse-foley animation: Synchronizing rigid-body motions to sound. ACM Transactions on Graphics (TOG) **33**(4), 1–11 (2014) 3
40. Lee, S.H., Roh, W., Byeon, W., Yoon, S.H., Kim, C.Y., Kim, J., Kim, S.: Sound-guided semantic image manipulation. arXiv preprint arXiv:2112.00007 (2021) 3
41. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. In: ACM SIGGRAPH 2010 papers, pp. 1–11 (2010) 3
42. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196 (2018) 3
43. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017) 20
44. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013) 8
45. Morgado, P., Vasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: Advances in Neural Information Processing Systems (2018) 3
46. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12475–12486 (2021) 3
47. Nam, S., Kim, Y., Kim, S.J.: Text-adaptive generative adversarial networks: Manipulating images with natural language. In: Advances in neural information processing systems (2018) 3
48. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011) 3
49. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (2018) 3
50. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2405–2413 (2016) 3, 7
51. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision. pp. 319–345 (2020) 2, 3, 5, 11, 20
52. Parmar, G., Zhang, R., Zhu, J.Y.: On buggy resizing libraries and surprising subtleties in fid calculation. arXiv preprint arXiv:2104.11222 (2021) 21

53. Plakal, M., Ellis, D.: YAMNet. Jan 2020 [Online], available: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet 4, 7, 8, 9, 13, 20
54. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: Learning individual speaking styles for accurate lip to speech synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13796–13805 (2020) 3
55. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: 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. pp. 484–492 (2020) 3
56. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (2021) 3, 8, 21, 22
57. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092 (2021) 2, 3
58. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning. pp. 1060–1069 (2016) 2, 3
59. de Sa, V.R.: Learning classification with unlabeled data. In: Advances in neural information processing systems. pp. 112–119. Citeseer (1994) 3
60. Shlizerman, E., Dery, L., Schoen, H., Kemelmacher-Shlizerman, I.: Audio to body dynamics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7574–7583 (2018) 3
61. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 21
62. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural computation **12**(6), 1247–1283 (2000) 1
63. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12695–12705 (2020) 10, 11
64. Wu, C., Timm, M., Maji, S.: Describing textures using natural language. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 52–70. Springer (2020) 3
65. Wu, H.H., Seetharaman, P., Kumar, K., Bello, J.P.: Wav2clip: Learning robust audio representations from clip. arXiv preprint arXiv:2110.11499 (2021) 22
66. Yang, K., Russell, B., Salamon, J.: Telling left from right: Learning spatial correspondence of sight and sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9932–9941 (2020) 3
67. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE international conference on computer vision. pp. 2849–2857 (2017) 3
68. Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J.H., Tenenbaum, J.B., Freeman, W.T.: Generative modeling of audible shapes for object perception. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1251–1260 (2017) 3
69. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1735–1744 (2019) 3

70. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018) 3
71. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 13, 22
72. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9299–9306 (2019) 3
73. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3550–3558 (2018) 3
74. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2223–2232 (2017) 3, 10, 11, 20

A Appendix

A.1 *Into the Wild* dataset

We introduce the *Into the Wild* dataset, a set of egocentric hiking videos for our proposed audio-driven image stylization (ADIS), because hiking is featured with a strong audio-visual association of nature.

We collected these videos on YouTube by searching for the keywords like hike+POV, hike+footsteps, hike+ASMR, and hike+binaural. We employ YAMNet [53] to tag each associated soundtrack to ensure that they play the actual sound and are not replaced by any other sounds, such as background music.

The duration statistics of the *Into the Wild* dataset are shown in Figure 11a. Specifically, it 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 [53] 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 11b.

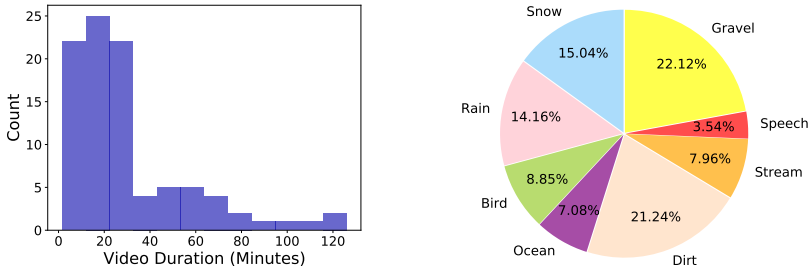
A.2 Training Details

Training Setting Except for the batch size and audio network, we intentionally match the architecture and hyperparameter settings with CycleGAN [74] and CUT [51]. We employ ResNet-based generator [33] with 9 residual blocks, PatchGAN discriminator [32], Least Square GAN loss [43], ResNet18-based audio encoder [25], with the batch size of 16, and the Adam optimizer [36] 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 [74]. 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) Distribution of video duration (b) Distribution of video categories

Fig. 11: Statistical analysis of the *Into the Wild* Dataset.

Fig. 12: A screenshot of AMT for rating the audio-visual correspondence.

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 [12] 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 [28] 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 [61]. We adopt Clean-FID [52] 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) [56] 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 [56].

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 [65]	**0.831**	**30.334**	**0.246**

Amazon Mechanical Turk (AMT) In addition to the objective evaluations mentioned above, we employ AMT to 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 12. 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 4 points and the worst earning 1 point. Thus, the scores range from 1 to 4. Notably, twenty Mturkers were asked to rank a total of 1000 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 13. 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 14. These are accomplished by using images from the Places dataset [71] and the audios from the VGG-Sound dataset [7]. Our model is able to generate plausible images that match the content of the out-of-distribution audio.

Additional pre-training comparisons We also use Wav2CLIP [65], an audio representation learning method derived on CLIP [56], to fine-tune ADIS. To transfer knowledge, it employs a frozen image model to bridge the gap between a sophisticated language model and a scratch audio model. Wav2CLIP could be a better pre-training method 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.

Fig. 13: Randomly selected qualitative results of our model, baselines and ablations. This is an extension of Figure 5 in the main paper.

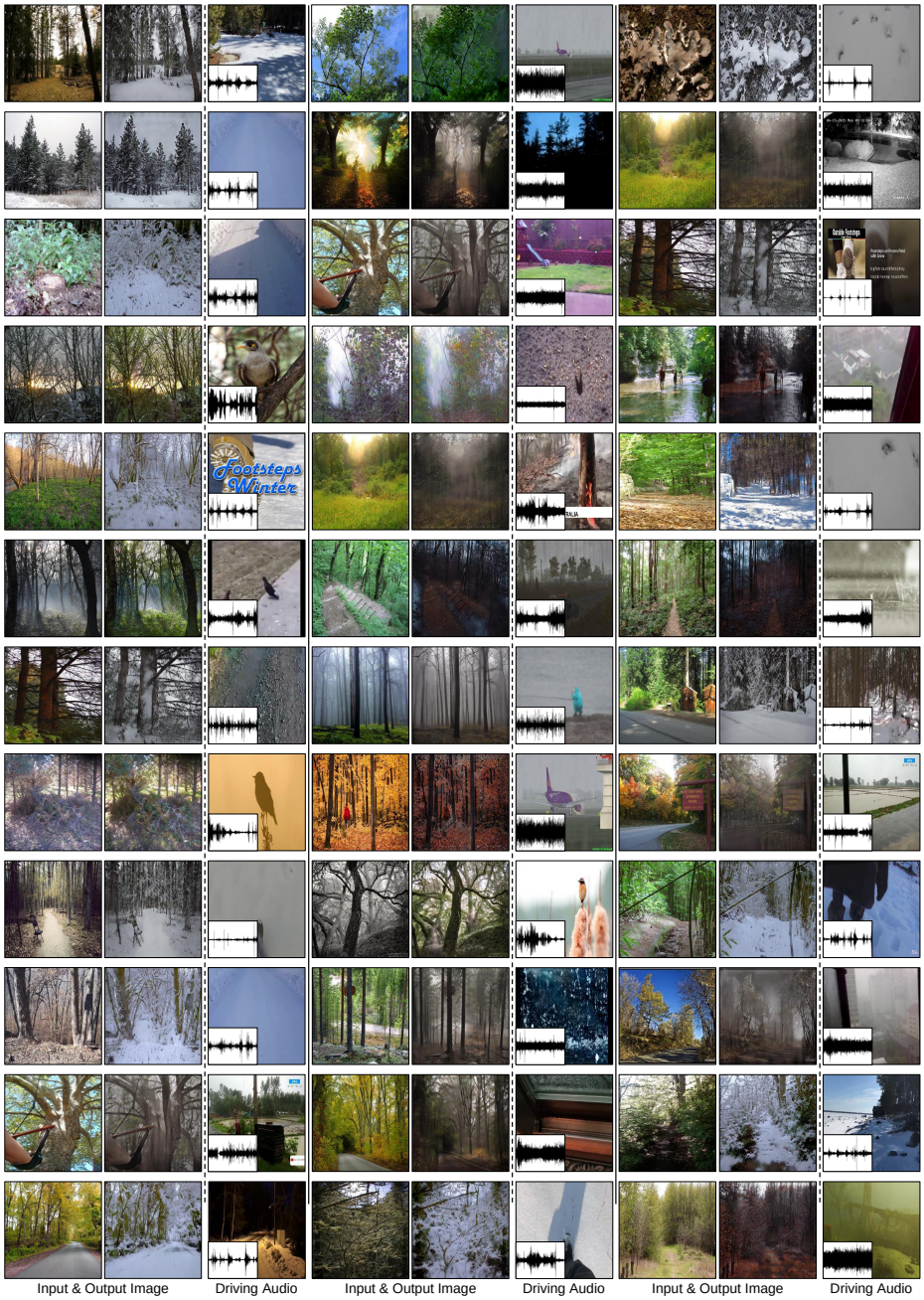


Fig. 14: Randomly selected qualitative results of generalization experiment. This is an extension of Figure 8 in the main paper.