Sound-Guided Semantic Image Manipulation

Seung Hyun Lee[1], Wonseok Roh[1], Wonmin Byeon[2], Sang Ho Yoon[3], Chanyoung Kim[1],
Jinkyu Kim[4*] and Sangpil Kim[1*]

[1]Department of Artificial Intelligence and [4]CSE, Korea University
[2]NVIDIA Research, NVIDIA Corporation
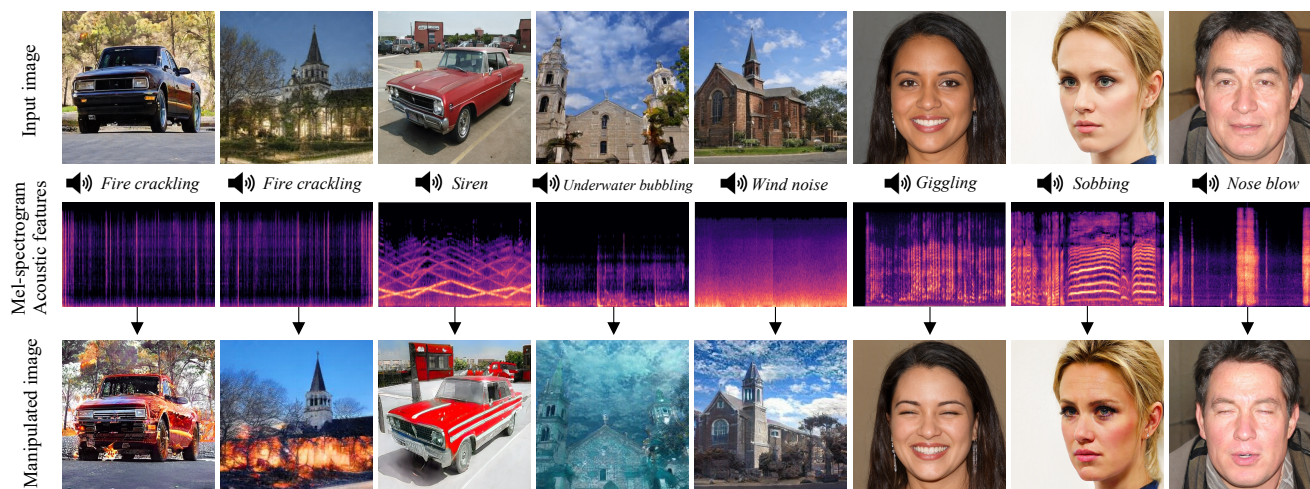[3]Graduate School of Culture Technology, KAIST

Figure 1: Modified images with sound-guided semantic image manipulation. Our method manipulates source images (top row) given user-provided sound (middle row) into semantic images (last row).

Abstract

The recent success of the generative model shows that leveraging the multi-modal embedding space can manipulate an image using text information. However, manipulating an image with other sources rather than text, such as sound, is not easy due to the dynamic characteristics of the sources. Especially, sound can convey vivid emotions and dynamic expressions of the real world. Here, we propose a framework that directly encodes sound into the multi-modal (image-text) embedding space and manipulates an image from the space. Our audio encoder is trained to produce a latent representation from an audio input, which is forced to be aligned with image and text representations in the multi-modal embedding space. We use a direct latent optimization method based on aligned embeddings for sound-guided image manipulation. We also show that our method can mix different modalities, i.e., text and audio, which enrich the variety of the image modification. The experiments on zero-shot audio classification and semantic-level image classification show that our proposed model outperforms other text and sound-guided state-of-the-art methods.

1. Introduction

Image manipulation has been widely studied in the field of computer vision due to its usefulness in photo-realistic manipulation applications, social media image sharing, and image-based advertisement. An image can be used to transfer its style into the target image [14, 13]. Also, modifying specific parts in the human face image, such as hairstyle or color, is useful in image manipulation applications [49, 34]. The purpose of semantic image manipulation is to generate a novel image that contains both source image identification and semantic information of user intention. In this pa-

per, we tackle the semantic image manipulation task, which is the task of modifying an image with user-provided semantic cues. To apply the user intention into the image, a mixture of sketches and text is used to perform image manipulation and synthesis [33, 49]. User intention can be applied by drawing a paint [33] or writing text with semantic meanings [49, 13].

Text-based image manipulation methods are proposed to edit the image conditionally [11, 21, 26, 29, 49]. These works modify target contents in the image based on the text information. Among the text-based image manipulation methods, StyleCLIP [34] considered leveraging the representational power of Contrastive Language-Image Pre-training (CLIP) [37] models to produce text-relevant manipulations with given text input. StyleCLIP maintains high quality image generation ability using StyleGAN [20] while allowing insertion of semantic text into the image.

However, text-based image manipulation has an inherent limitation when applying sound semantics into the image. It is impossible to fully express vivid sound with discrete text since sound has continuous and dynamic properties. For example, every "thunder" generates different loudness and characteristic of "*sound of thunder*". The infinite range of sound variations is not interchangeable with discrete text expressions. Therefore, the text-based image manipulation methods suffer from the discreteness of the text expression, limiting the transfer of specific and vivid sound semantics into the source image in image manipulation with sound semantics.

Several studies [7, 18, 30, 36, 47, 52] have attempted to visualize the meaning of sound, but it is still challenging to reflect sound events in high-resolution images due to two reasons. The first reason is the lack of a suitable high-resolution audio-visual dataset. Audio-visual benchmark video datasets [5, 24, 42] for GAN training has generally lower resolution than high-resolution image datasets including Flickr-Faces-HQ (FFHQ) [22] and The Large-scale Scene Understanding Challenge (LSUN) [50]. There is no dataset with as many audio-visual pairs as the number of image-text pairs used for CLIP training. CLIP uses 400 million image-text pair data to learn the relationship between very large and diverse image and text modalities, whereas audio-visual pair data is still insufficient. Secondly, it is difficult to discover potential correlations between auditory and visual modalities [52]. Extracting appropriate temporal context, tone, and theme from the sound is difficult.

To overcome these challenges of manipulating images with sound semantics, we introduce a novel image manipulation method driven by sound semantics (see Fig. 2). As shown in Fig. 1, an image of an old car is manipulated into an old car with a fire truck-like exterior appearance when adding a siren sound. Our model consists of two main stages: (i) the CLIP-based Multi-modal Representa-

tion Learning, where an audio encoder is trained to produce a latent representation aligned with textual and visual semantics by leveraging the representation power of pre-trained CLIP models. (ii) the Sound-Guided Image Manipulation, where we use the direct latent code optimization to produce a semantically meaningful image in response to a user-provided sound.

Our experimental results show that the proposed method supports a variety of sound sources with a better reflection of given audio information when transferring image styles. The sound-based approach supports more diverse and detailed information related to scenes compared to text-based image manipulation methods. We illustrate diverse examples in the supplemental material and project website.

Our main contributions are listed as follows:
- We propose multi-modal contrastive losses to expand the CLIP-based embedding space. Moreover, we introduce contrastive learning on augmented audio data, which helps to learn a more robust representation.
- We propose semantic-level image manipulation solely based on the given audio features, including temporal context, tone, and volume.
- We propose the sound-guided code optimization steps with adaptive layer masking for putting sound meaning into images, enhancing the realism of the output.

2. Related Work

Text-guided Image Manipulation. Text-guided image manipulation is the most widely studied among guidance based tasks. Several studies [10, 26, 29] employed the GAN-based encoder-decoder structure to preserve the features of the image while presenting image manipulations corresponding to the text description. StyleCLIP [34] and TediGAN [49] utilize the latent space of the pre-trained StyleGAN and the prior knowledge from CLIP [37]. StyleCLIP performed image manipulation using a user-provided text prompt. TediGAN enabled image generation and manipulation using GAN inversion technique using multi-modal mapping. Beyond text and images, a sound can express a complex context appearing in a scene, and there is correspondence between a sound and an event occurring in the scene.

Sound-guided Image Manipulation. Sound contains temporal dynamic information of a scene, which can be used as an imagery source for image manipulation. Some approaches have been introduced for the sound-guided image manipulation task. However, the previous works mainly focus on music (instead of using sound semantics), which includes music-to-visual style transfer with cross-modal learning strategy [25] and a neural music visualizer by mapping music embeddings to visual embeddings from StyleGAN [20]. To manipulate the image according to the sound, *TräumerAI* [20] visually expresses music by latent transfer mapping of music to StyleGAN's style embedding.
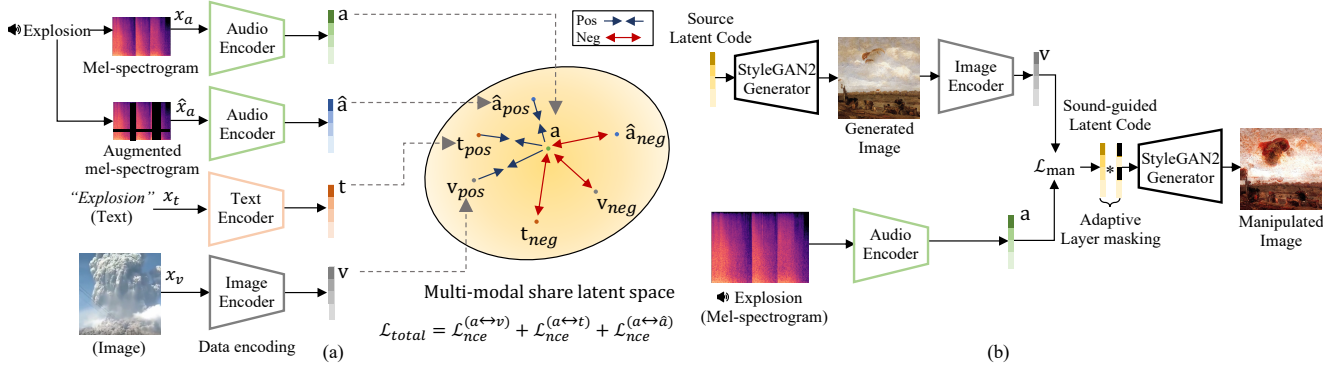
Figure 2: Our model consists of two main steps: (a) the *CLIP-based Contrastive Latent Representation Learning* step and (b) the *Sound-Guided Image Manipulation* step. In (a), we train a set of encoders with three different modalities (audio, text, and image) to produce the matched latent representations. The latent representations for a positive triplet pair (e.g., audio input: "Explosion", text: "explosion", and corresponding image) are mapped close together, while that of negative pair samples further away in the (CLIP-based) embedding space (left). In (b), we use a direct code optimization approach where a source latent code is modified in response to user-provided audio, producing a sound-guided image manipulation result (right).

However, the above studies have limitations in focusing only on the reaction, not the semantic of the sound, in the direction of navigation in the latent space of StyleGAN. *Crossing you in style* [25] uses the period to define the semantic relationship between sound and visual domain, but there is still a limitation which only can transfer the image style. Our proposed method can isolate the modification area in the source image, such as modifying the emotion of the face while preserving the color of the hair.

Interpreting Latent Space in StyleGAN. The intermediate latent space in pre-trained StyleGAN [22] solves the disentanglement issue and allows the generated images to be manipulated meaningfully according to changes in the latent space. Extended latent space $\mathcal{W}+$ allows image manipulation with interpretable controls from a pre-trained GAN generator [1, 22, 23]. For latent space analysis in audio sequences, *Audio-reactive StyleGAN* [4] generates an image every time step by calculating the magnitude of the audio signal and moving it in the latent space of StyleGAN. However, the method cannot control the meaning of sound in the latent space. StyleGAN's motion in the latent space is only mapped to the magnitude of the sound. There is a novelty in that we manipulate images with the properties of sound.

Audio-visual Representation Learning. Cross-modal representation learning obtains relationships between different modalities in audio-visual tasks such as video retrieval and text-image cross-modal tasks such as image captioning and visual question answering. Audio-visual representation learning studies [3, 28, 44] aim to map both modalities to the same embedding space. The correlation between modalities is learned by contrastive learning between composite audio-visual pairs [8, 27, 43]. However, audio-visual representation learning is still challenging because there is no adequate data as much as CLIP [37] for learning the corre-

lation between different modalities. CLIP learned the relationship between image and text embedding by multi-modal self-supervised learning of 400 million image-text pairs and showed zero-shot inference performance comparable to supervised learning in most image-text benchmark datasets. In this paper, the audio encoder not only exploits the representation ability of CLIP but also learns supervisory signals from the audio data itself through self-supervised manners. As a result, our method obtains an audio-specific representation for sound-guided image manipulation.

3. Method

We follow the existing text-guided image manipulation model, StyleCLIP [34]. Our model and StyleCLIP manipulate the latent code of StyleGAN using joint embedding space between modalities. However, our model extends the CLIP [37] embedding space to the audio embedding space, which was not embedded before. We also introduce novel contrastive losses and adaptive masking for sound-guided image manipulation. Our model consists of two main steps: (i) the CLIP-based Multi-modal Latent Representation Learning and (ii) the Sound-guided Image Manipulation. First, we train audio, text, and image encoders to generate new latent representations. To do so, we train the audio encoder using the InfoNCE loss [31, 2, 51] to produce a latent representation that is aligned with the representations from the pre-trained CLIP's text and image encoders. Such aligned representations can be used for image manipulation with the provided audio input. After the pre-training step, we use encoders to manipulate images according to a target sound input (e.g., images with different facial expressions can be manipulated with different sound inputs).

3.1. Multi-modal Latent Representation Learning

As shown in Fig. 2 (a), we train a set of encoders with three different modalities {audio, text, and image} to pro-

duce the matched representations in the embedding space. Specifically, given audio, text, and image inputs, i.e. x_a, x_t, and x_v, we use three different encoders to obtain a set of d dimensional latent representations, i.e. $\mathbf{a}$, $\mathbf{t}$, and $\mathbf{v} \in \mathbb{R}^d$, respectively. These latent representations are learned via a typical contrastive learning approach following the work by Radford *et al.* [37] – the latent representations for a positive triplet pair are mapped close together in the embedding space, while that of negative pair samples further away. Learning such a joint representation from scratch is, however, generally challenging due to the lack of multi-modal datasets, which can provide positive and negative pairs. Thus, we instead leverage the pre-trained CLIP model, which optimized a visual-textual joint representation by contrastive learning. Then, we train an audio encoder to produce an aligned representation by using contrastive learning. Details are explained in the next section. Note that we obtain a latent representation $\hat{\mathbf{a}} \in \mathbb{R}^d$ from an augmented audio input $\hat{x}_a$, which is shown useful to improve the quality of the latent representation as this is a common practice in the self-supervised representation learning.

Matching Multi-modal Representations via Contrastive Loss. We use the InfoNCE loss [2] to map positive audio-text pairs close together in the CLIP-based joint embedding space, while negative pairs further away. Formally, given a minibatch of N audio-image representation pairs $\{\mathbf{a}_i, \mathbf{t}_j\}$ for $i \in \{1, 2, \ldots, N\}$, we first compute the following audio-to-text loss function for the i-th pair:

$$l_i^{(a \to t)} = -\log \frac{\exp(\langle \mathbf{a}_i, \mathbf{t}_i \rangle / \tau)}{\sum_{j=1}^N \exp(\langle \mathbf{a}_i, \mathbf{t}_j \rangle / \tau)}, \quad (1)$$

where $\langle \mathbf{a}_i, \mathbf{t}_j \rangle$ represents the cosine similarity, i.e. $\langle \mathbf{a}_i, \mathbf{t}_j \rangle = \mathbf{a}_i^\mathsf{T} \mathbf{t}_j / \|\mathbf{a}_i\| \|\mathbf{t}_j\|$ and τ is a temperature parameter. This loss function is the log loss of an N-way classifier that wants to predict $\{\mathbf{a}_i, \mathbf{t}_j\}$ as the true representation pair. As the loss function is asymmetric, we define the following similar text-to-audio contrastive loss:

$$l_i^{(t \to a)} = -\log \frac{\exp(\langle \mathbf{t}_i, \mathbf{a}_i \rangle / \tau)}{\sum_{j=1}^N \exp(\langle \mathbf{t}_i, \mathbf{a}_j \rangle / \tau)}. \quad (2)$$

Concretely, we minimize the following loss function $\mathcal{L}_{nce}$ as a sum of the two losses $l_i^{(a \to t)}$ and $l_i^{(t \to a)}$ for all positive audio-text representation pairs in each minibatch of size N:

$$\mathcal{L}_{nce}^{(a \leftrightarrow t)} = \frac{1}{N} \sum_{i=1}^N (l_i^{(a \to t)} + l_i^{(t \to a)}). \quad (3)$$

Applying Self-supervised Representation Learning for Audio Inputs. Self-supervised learning approaches rely on a contrastive loss that encourages representations of the same-class different views to be close in the embedding space, while that of different-class views to be pushed away
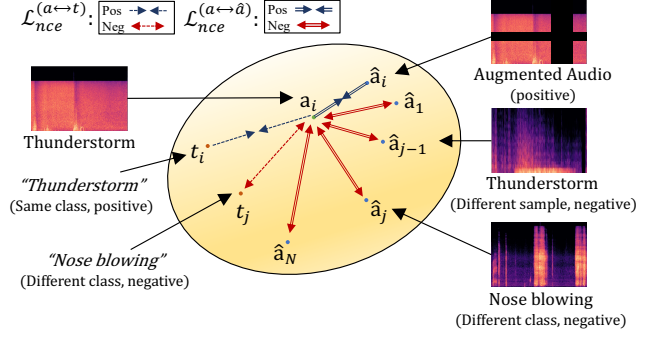

Figure 3: Multi-modal contrastive learning with audio self-supervised loss.

from each other. We apply this technique to improve the quality of audio representations by minimizing the following $\mathcal{L}_{self}^{(a \leftrightarrow \hat{a})}$:

$$\mathcal{L}_{self}^{(a \leftrightarrow \hat{a})} = \frac{1}{N} \sum_{i=1}^N (l_i^{(a \to \hat{a})} + l_i^{(\hat{a} \to a)}), \quad (4)$$

where $l_i^{(a \to \hat{a})}$ and $l_i^{(\hat{a} \to a)}$ are defined in a similar way as in Eq. 1 and 2. This loss function is useful to learn subtle differences over sound inputs as it needs to maximize the mutual information between two different views of the same inputs but to minimize the mutual information between two views of the different inputs. For example, as shown in Fig. 3, an audio sample a_i forms a negative pair with $\hat{a}_j$ for $i \neq j$, which induces a diffusive effect in the embedding space.

Data Augmentation. We further apply the data augmentation strategy to improve the quality of representations and to overcome the lack of large-scale audio-text multimodal datasets. For audio inputs, we apply the SpecAugment [32], which visually augments Mel-spectrogram acoustic features by warping the features and masking blocks of frequency channels. For text inputs, we augment text data by (i) replacing words with synonyms, (ii) applying a random permutation of words, and (iii) inserting random words. Note that, for (i) we find synonyms of the given word from WordNet [12] and insert the synonym anywhere randomly in the given text input. For example, we augment original texts *"rowboat, canoe, kayak rowing"* to produce new text *"row canoe, kayak quarrel rowboat."*

Loss Function. To summarize, we minimize the following loss function $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \mathcal{L}_{nce}^{(a \leftrightarrow v)} + \mathcal{L}_{nce}^{(a \leftrightarrow t)} + \mathcal{L}_{self}^{(a \leftrightarrow \hat{a})}. \quad (5)$$

3.2. Sound-guided Image Manipulation

After learning the multi-modal joint embedding space by minimizing Eq. 5, we use a direct latent code optimization method to manipulate the given image similar to StyleCLIP [34]. As shown in Fig. 2 (b), our model minimizes the

distance between a given source latent code and an audio-driven latent code in the learned joint embedding space to produce sound-guided manipulated images. Moreover, we propose a *Adaptive Layer Masking* technique, which adaptively manipulates the latent code.

Direct Latent Code Optimization. We employ the direct latent code optimization for sound-guided image manipulation by solving the following optimization problem:

$$\mathcal{L}_{man} = \underset{w_a \in \mathcal{W}+}{\mathrm{argmin}}\ d_{\text{cosine}}(G(w_a), a) + \lambda_{\text{ID}}\mathcal{L}_{ID}(w_a)$$
$$+ \lambda_{sim}||g \cdot (w_a - w_s)||_2,$$
$$(6)$$

where a given source latent code $w_s \in \mathcal{W}$ (the intermediate latent space in StyleGAN), audio-driven latent code $w_a \in \mathcal{W}+$. g is a trainable vector to mask the specific style layer adaptively. $\mathcal{L}_{\text{ID}}$ and G are the identity loss and StyleGAN-based generator, respectively. λ_{sim} and λ_{ID} are hyperparameters to control the strengths of the similarity loss term and the identity loss function $\mathcal{L}_{\text{ID}}$. High values of λ_{sim} and λ_{ID} lead to maintaining the content of the source image, while low values do not. The source latent code w_s means the randomly generated latent code from G or the latent code obtained from the existing input image through GAN inversion [38, 45]. With such an optimization scheme, we minimize the cosine distance $d_{\text{cosine}}(G(w_a), a)$ between the embedding vectors of the manipulated image $G(w_a)$ and the audio input a.

Identity Loss. The similarity to the input image is also controlled by the identity loss function $\mathcal{L}_{\text{ID}}$, which is defined:

$$\mathcal{L}_{\text{ID}}(w_a) = 1 - \langle R(G(w_s), R(G(w_a)))\rangle. \qquad (7)$$

R is the pre-trained ArcFace [9] model for face recognition, thus this loss function minimizes the cosine distance $\langle R(G(w_s), R(G(w_a)))\rangle$ between its arguments in the latent space of the ArcFace network. This allows manipulating facial expressions without changing the personal identity. Note that we disable the identity loss by setting $\lambda_{\text{ID}} = 0$ for all other image manipulations.

Adaptive Layer Masking. We control style changes with adaptive layer masking. L_2 regularization is effective in keeping the image generated from the moved latent code from being different from the original [34]. However, StyleGAN's latent code has different properties per each layer, so different weights should be applied to each layer if the user-provided attribute changes. We use layerwise masking to keep compact content information within style latent code. In StyleGAN2 [23], the latent code represents as $w \in \mathbb{R}^{L \times D}$, where L is the number of the network layers, and D is the latent code's dimension size. We declare a parameter vector g in L dimension. In latent optimization step, g and w are multiplied per layer. g is iteratively updated, which adaptively manipulates the latent code.
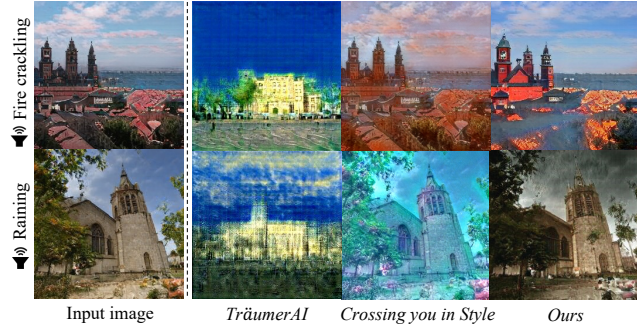

Figure 4: Comparison of sound-guided manipulation results. Given fire crackling (top) and raining (bottom) audio inputs, we manipulate the input image with TräumerAI [20], Crossing you in style [25], and our method.

Sound and Text Multi-modal Style Mixing. Multi-modal manipulation of audio and text is based on style mixing of StyleGAN. Different layers of w latent code in StyleGAN represent different properties. Because audio and text share the same new multi-modal embedding space, selecting a specific layer of each latent code guided by audio and text can manipulate the image using properties of audio and text.

4. Experiments

Implementation Details. We use the pre-trained CLIP [37] as image and text encoder. For our audio encoder, we use ResNet50 [19] as backbone, where we employ the same output dimension, 512, as the image and text encoder. First, we convert audio inputs to Mel-spectrogram acoustic features. Then, our audio encoder takes these features as an input to produce a 512-dimensional latent vector. The details about the train dataset are explained in supplemental material. For the manipulation step, we leverage Style-GAN2 [23]'s pre-trained generator. We set the size of latent code based on the resolution of the learned image. Here, we set 18×512 for images of size 1024×1024 and 14×512 for 256×256.

We train our model for 50 epochs using the Stochastic Gradient Descent (SGD) with the cosine cyclic learning rate scheduler [41]. We set the learning rate to 10^{-3} with the momentum 0.9 and weight decay 10^{-4}. The batch size is set to 384. For audio augmentation, we use SpecAugment [32] with the frequency mask ratio of 0.15 and time masking ratio of 0.3. For direct latent code optimization, λ_{sim} and λ_{ID} in in Eq. (6) are set to 0.008 and 0.004 for the FFHQ dataset; and 0.002 and 0 for the LSUN dataset.

4.1. Qualitative Analysis

Sound-guided Image Manipulation. We first compare our sound-guided image manipulation model with existing sound-based style-transfer models including TräumerAI [20] and Crossing you in Style [25]. Fig. 4 showcases image manipulation results in response to given audio inputs, including fire crackling and raining.

(a) Input image

(b) TediGAN (Text → Image)

(c) StyleCLIP (Text → Image)

(d) Ours (Sound → Image)

🔊 Baby crying   🔊 People coughing   🔊 People giggling   🔊 People screaming
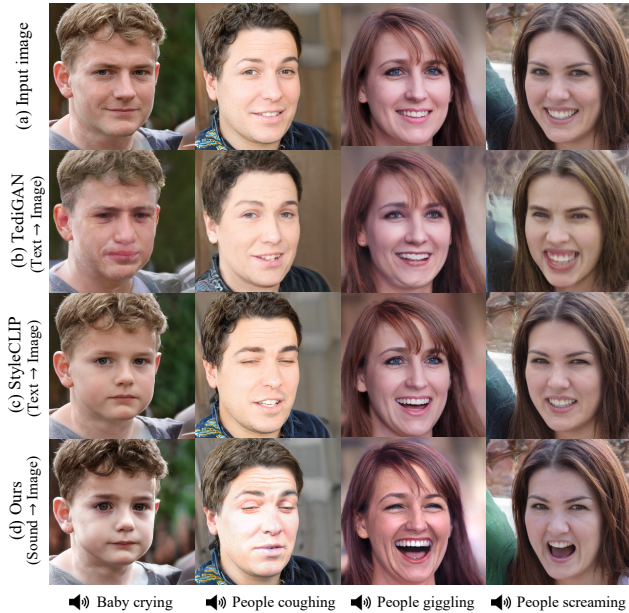
Figure 5: Given the (a) input image, we compare the image manipulation results between (b-c) text-driven image manipulation approaches (i.e. TediGAN [49] and Style-CLIP [34]) and (d) ours. Attributes for driving such manipulations include baby crying, people coughing, people giggling, and people screaming.

We observe that our model produces a better quality of manipulated images where existing models often fail to capture semantic information of the given audio input (See 2nd and 3rd columns).

Comparison of Text-guided Image Manipulation. We use the latest text-guided image manipulation models as a baseline, including TediGAN and the latent optimization technology of StyleCLIP. As shown in Fig. 5, the proposed sound-guided image manipulation (proposed method) shows more radical results than text-guided manipulation (TediGAN [49] and StyleCLIP [34]). Unlike text-guided methods, the audio-guided approach achieves natural image style transfer while capable of reflecting multiple labels. For example, TediGAN emphasizes crying, whereas StyleCLIP focuses on the baby when "baby crying" context is given. On the contrary, our proposed method is capable of handling "baby" and "crying" simultaneously.

We demonstrate that each audio sample has its own context, which makes the guidance richer than text (Fig. 6). If the magnitude of *Thunder* is altered or a specific attribute like *Rain* is added to the audio, the manipulation context becomes more diverse than text-guided image manipulation.

We visualize the direction vector with t-SNE [46] in a supplemental document. By subtracting the vectors of the latent code guided by each modality and the source latent code, we show the distribution of manipulating direction. We select the attributes in VGG-Sound [6] and randomly

Source Image   Raining w/o thunder   Raining w/ weak thunder   Raining w/ strong thunder
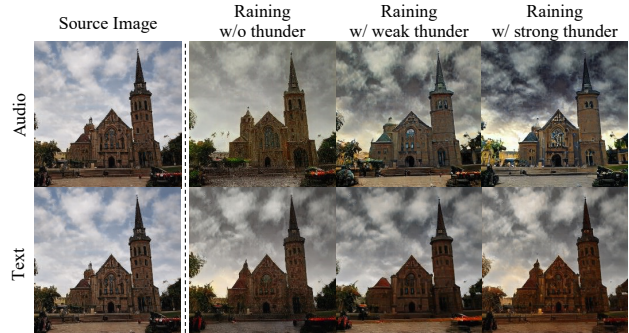
Audio

Text

Figure 6: Comparison of manipulation results between ours (top) and the existing text-driven manipulation approach, StyleCLIP [34] (bottom). Unlike the text-driven approach, ours can produce more diverse manipulation results in response to different intensities of raining, i.e. raining, raining with weak thunder, and raining with strong thunder.
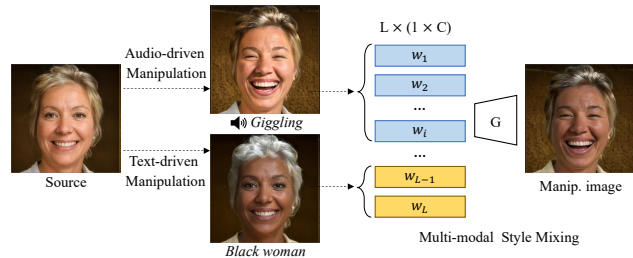

Figure 7: An example of image style mixing jointly with the audio (*people giggling*) and text input (*black woman*).

manipulate the audio and text prompts. Although we randomly sample the audio and text in the same labels, the sound-guided latent code shows a more significant transition than the text-guided latent code. We use various text synonyms for a fair comparison, but text-guided latent code seems less effective with changes.

Multi-modal Image Manipulation. Our method ensures that audio, text, and image share the same embedding space. To demonstrate that multi-modal embedding lies in a same latent space, we interpolated text and sound-guided latent code (see supplementary document). Constructing multi-modal shareable latent space enables joint modification of the target image with user-provided text and audio inputs from the same embedding space. We further perform multi-modal style mixing experiments by selecting a specific layer of latent code and mixing style with audio and text. We find that the sound source can effectively manipulate facial emotion aspects such as "giggling" on the face and text information controls the background color of the target image (Fig. 7). For the style-mixing details, we follow TediGAN's StyleGAN layerwise analysis [49]. In the 18×512 latent code, the style-mixing technique selects the 1st to 9th layers of the sound-guided latent code and the 10th to 18th layers of the text-guided latent code to mix the dynamic characteristics of sound and human properties of text.

Input image / **w/o layer masking** / **w/ layer masking**

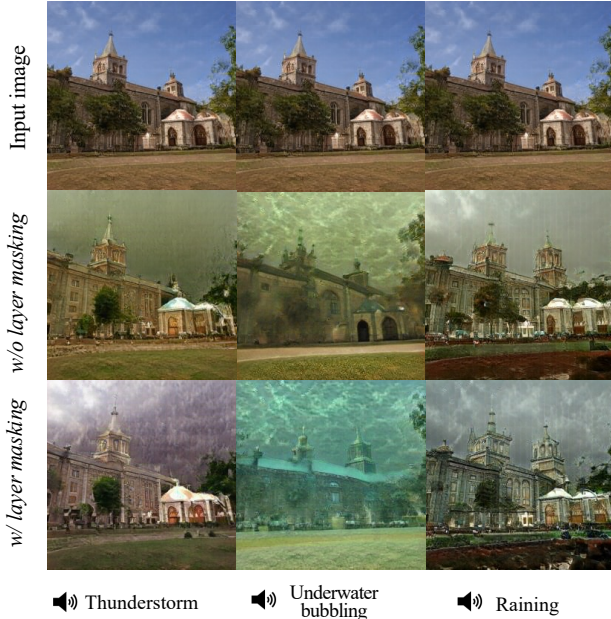🔊 Thunderstorm   🔊 Underwater bubbling   🔊 Raining

Figure 8: Ablation study of adaptive layer masking. The first row is the input image, the second row is the manipulation results without adaptive layer masking, and the third row is the sound-guided image manipulation results after applying the adaptive layer masking.

Effect of Adaptive Layer Masking. In StyleGAN [20], it is necessary to adaptively regularize style layer since each layer of latent code has different style attributes. For each layer of latent code, it multiplies the trainable parameter that controls the diversity during regularization. The ablation study shows a qualitative comparison of the mechanism for applying adaptive layer masking to the style layer, as illustrated in Fig. 8. The adaptive masking rectifies the direction by changing the latent code based on the semantic cue. When applying the gate function, sound-guided image manipulation is semantically reasonable. For example, a thunderstorm is a blend of thunder and rain sound. Although thunder and lightning are not seen in the second row, lightning and rain appear in the last row. Manipulation results according to λ_{sim} and λ_{ID} hyperparameters are added to the supplemental material.

4.2. Quantitative Evaluation

Zero-shot Transfer. We compare our model to the supervised method and the existing zero-shot audio classification method (see Table 1). First, we compare audio embeddings trained by supervised methods such as logistic regression, ResNet50 [19] supervised by random initialization of weights as a baseline model. Our audio encoder shows better classification performance than the baseline. Secondly, we compare the zero-shot audio classification accuracy with Wav2clip [48] and AudioCLIP [16]. Our proposed loss function learns three modalities in the CLIP embedding space and obtains a more rich audio representa-

Table 1: Comparison of the quality of audio representations between ours and alternatives. We report classification accuracy (top-1 in %) of a linear classifier on the ESC-50 [35] and the Urban sound 8k [39] datasets as well as their zero-shot inference results. *Abbr. S*: supervised setting.

Model	S	Zero-shot	Dataset	
			ESC-50	Urban sound 8k
ResNet50 [19]	✓	-	66.8 %	**71.3 %**
Ours w/o $\mathcal{L}_{nce}^{(a \leftrightarrow \hat{a})}$ (LR)	-	-	58.7 %	63.3 %
Ours (LR)	-	-	**72.2 %**	66.8 %
Wav2clip [48]	-	✓	41.4 %	40.4 %
AudioCLIP [16]	-	✓	**69.4 %**	**68.8 %**
Ours w/o $\mathcal{L}_{nce}^{(a \leftrightarrow \hat{a})}$	-	✓	49.4 %	45.6 %
Ours	-	✓	57.8 %	45.7%

tion through the contrastive loss whereas Wav2clip only learns the relationship between audio and visual context. Although the performance of the AudioCLIP in zero-shot task is better than ours, comparison between ours and AudioCLIP is not an apple-to-apple comparison. This is because, AudioCLIP uses pretrained ESResNeXt [17] as an audio encoder, which is different from ours. The AudioCLIP's audio encoder was pre-trained on a large-scale datsaet of audio, namely AudioSet [15], which consists of over 2M sound clips. However, ours was trained in end-to-end manner without benefits from such a large dataset. Moreover, our learned embedding space is more suitable for the image manipulation task (see the supplemental material).

Semantic Accuracy of Manipulation. We quantitatively analyze the effectiveness of our proposed audio-driven image manipulation approach. First, we measure performance on the semantic-level classification task. Given the audio embeddings from our pre-trained audio encoder, we train a linear classifier to recognize eight semantic labels including giggling, sobbing, nose-blowing, fire crackling, wind noise, underwater bubbling, explosion, and thunderstorm. We use StyleGAN2 [23] weights pre-trained from the FFHQ [22] dataset when guiding with giggling, sobbing, and nose blowing attributes to compare the semantic-level classification accuracy between text and audio. Also, when guiding with fire crackling, wind noise, underwater bubbling, explosion, and thunderstorm attributes, the weights of StyleGAN2 pre-trained with the LSUN (church) [50] dataset are used. As shown in Fig. 9 (a), we generally outperform existing text-driven manipulation approach with better semantically-rich latent representation.

Distribution of Manipulation Direction. We can see how much the latent code has changed by the cosine similarity between the source latent code and the manipulated latent code. We compare the cosine similarity between text-guided and sound-guided latent representations. We evaluate the mean and variance of the cosine similarity between w_s, a source latent code, w_a, an audio-driven latent code, and w_t,
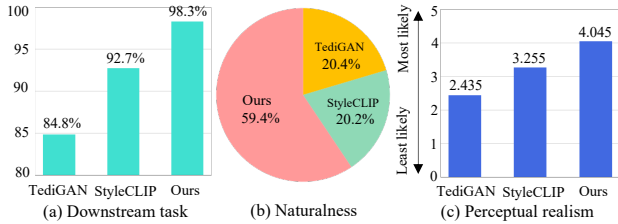
Figure 9: Quantitative evaluation and user study results. (a) Downstream task evaluation to compare the quality of representations between ours and text-driven manipulation approaches on the FFHQ [22] dataset. A linear classifier is trained to predict 8 semantic labels, e.g. giggling. Participants answered a questionnaire including (b) Naturalness ("Which image manipulation result better expresses the target attribute?") and (c) Perceptual realism ("On the scale from 1 to 5, how realistic are the images below?").

Figure 10: Failure cases of manipulation with our method.

a text-driven latent code. The latent representations generally exhibit a high-level characteristic of the content (see sup.). In the latent space of StyleGAN2, the sound-guided latent code moves more from the source latent code than the text-guided latent code, and the image generated from the sound-guided latent code is more diverse and dramatic than the text-guided method.

4.3. User Study

We recruit 100 participants from Amazon Mechanical Turk (AMT) for evaluating our proposed method. We show participants three manipulated images that are generated by TediGAN [49], StyleCLIP [34], and our model. Participants answer the following questionnaire: (i) Naturalness- *Which image manipulation result better expresses the target attribute?* and (ii) Perceptual Realism- *How realistic are the images below?* For perceptual realism, we employ Likert scale ranging from 1 (low realistic) to 5 (high realistic). Fig. 9 (b) and Fig. 9 (c) show that our method significantly outperforms other state-of-the-art approaches (TediGAN and StyleCLIP) in terms of *Naturalness* and *Perceptual Realism*. The large portion of participants (59.4%) chose generated image by our model as the best. Moreover, the result also shows that our method generated more natural images than other text-driven manipulation approaches.

5. Applications

Sound-Guided Artistic Paintings Manipulation. We propose a novel sound-guided image manipulation approach

Figure 11: Examples of sound-guided artistic paintings manipulation and music style transfer using our method.

for artistic paintings. We employ StyleGAN2 [23] generator which is pre-trained with the fine-art paintings dataset called WikiArt [40]. As shown in Fig. 11, our model could produce various manipulations for art paintings guided by given audio inputs. We observe that an audio input can successfully provide a semantic cue to manipulate artistic paintings. Given a fire crackling sound, a painting is manipulated with fire crackling. We also measured the manipulation quality for artistic painting using the Wikiart dataset using AMT. The responses showed that audio (73.3%) is better than text (26.7%) in terms of manipulation.

Music Style Transfer. Our method has a potential to reflect the mood of the music into the image style. Fig. 11 illustrates the results of image style transfer with various music genres. The source latent code is close to the keywords of each music, so the mood of the music appears in the image. For instance, *Funny* music manipulates the image with a fairy-tale style whereas *Latin* music manipulates the image with red-color theme which reflects *passion* characteristic.

6. Limitation

One major limitation of the proposed method is color and contrast shift changes when performing image manipulation with a sound which is not included in the domain of the pre-trained StyleGAN (see Fig. 10). Additionally, we need more sound types to conclude that the representation of audio embeddings are better than text.

7. Conclusion

We propose a method to manipulate images based on the semantic-level understanding from the given audio input. We take the user-provided audio input into the latent space of StyleGAN2 [23] and the CLIP [37] embedding space. Then, the latent code is aligned with the audio to enable meaningful image manipulation while reflecting the context from the audio. Our model produces responsive manipulations based on various audio inputs. We observe that an audio input can successfully provide a semantic cue to manipulate images accordingly. Our method of traversing multi-modal embedding space can be used in many applications with multi-modal contexts.

References

[1] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3

[2] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020. 3, 4

[3] Y. Aytar, C. Vondrick, and A. Torralba. See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932, 2017. 3

[4] H. Brouwer. Audio-reactive latent interpolations with stylegan. In *NeurIPS 2020 Workshop on Machine Learning for Creativity and Design*, 2020. 3

[5] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[6] H. Chen, W. Xie, A. Vedaldi, and A. 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. 6

[7] L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357, 2017. 2

[8] Y. Chen, Y. Xian, A. Koepke, Y. Shan, and Z. Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7016–7025, 2021. 3

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5

[10] H. Dong, S. Yu, C. Wu, and Y. Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017. 2

[11] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, and G. W. Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10304–10312, 2019. 2

[12] C. Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010. 4

[13] L. Gatys, A. Ecker, and M. Bethge. A neural algorithm of artistic style. *Journal of Vision*, 16(12):326–326, 2016. 1, 2

[14] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1

[15] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. 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. 7

[16] A. Guzhov, F. Raue, J. Hees, and A. Dengel. Audioclip: Extending clip to image, text and audio, 2021. 7

[17] A. Guzhov, F. Raue, J. Hees, and A. Dengel. Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. *arXiv preprint arXiv:2104.11587*, 2021. 7

[18] W. Hao, Z. Zhang, and H. Guan. Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[19] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. 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, 7

[20] D. Jeong, S. Doh, and T. Kwon. Träumerai: Dreaming music with stylegan. *arXiv preprint arXiv:2102.04680*, 2021. 2, 5, 7

[21] W. Jiang, N. Xu, J. Wang, C. Gao, J. Shi, Z. Lin, and S. Liu. Language-guided global image editing via cross-modal cyclic mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2115–2124, 2021. 2

[22] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 3, 7, 8

[23] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 3, 5, 7, 8

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[25] C.-C. Lee, W.-Y. Lin, Y.-T. Shih, P.-Y. Kuo, and L. Su. Crossing you in style: Cross-modal style transfer from music to visual arts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3219–3227, 2020. 2, 3, 5

[26] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 2

[27] P. Mazumder, P. Singh, K. K. Parida, and V. P. Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3090–3099, 2021. 3

[28] A. Nagrani, S. Albanie, and A. Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings*

of the European Conference on Computer Vision (ECCV), pages 71–88, 2018. 3

[29] S. Nam, Y. Kim, and S. J. Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[30] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7539–7548, 2019. 2

[31] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, page 2613–2617, 2019. 4, 5

[33] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[34] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 1, 2, 3, 4, 5, 6, 8

[35] K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 7

[36] Y. Qiu and H. Kataoka. Image generation associated with music data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2510–2513, 2018. 2

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2. 2, 3, 4, 5, 8

[38] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 5

[39] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014. 7

[40] B. Saleh and A. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, (2), 2016. 8

[41] L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017. 5

[42] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[43] Z. Sun, P. Sarma, W. Sethares, and Y. Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999, 2020. 3

[44] D. Surís, A. Duarte, A. Salvador, J. Torres, and X. Giró-i Nieto. Cross-modal embeddings for video and audio retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[45] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Trans. Graph.*, 40(4), jul 2021. 5

[46] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-629 sne. *Journal of Machine Learning Research*, 9(2579-2605):630, 2008. 6

[47] C.-H. Wan, S.-P. Chuang, and H.-Y. Lee. Towards audio to scene image synthesis using generative adversarial network. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500. IEEE, 2019. 2

[48] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello. Wav2clip: Learning robust audio representations from clip, 2021. 7

[49] W. Xia, Y. Yang, J.-H. Xue, and B. Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021. 1, 2, 6, 8

[50] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. 2015. 2, 7

[51] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 3

[52] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, pages 1–26, 2021. 2