Sound-Guided Semantic Image Manipulation

Seung Hyun Lee[1], **Sang Ho Yoon**[2], **Jinkyu Kim**[1*], **and Sangpil Kim**[1*]
[1]Korea University and [2]KAIST

Abstract

Semantically meaningful image manipulation often involves laborious manual human examination for each desired manipulation. Recent success suggests that leveraging the representation power of existing Contrastive Language-Image Pre-training (CLIP) models with the generative power of StyleGAN can successfully manipulate a given image driven by textual semantics. Following this, we explore adding a new modality, *Sound*, which can convey a different view of dynamic semantic information and thus can reinforce control strength over the semantic image manipulation. 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 same CLIP embedding space. Given such aligned embeddings, we use a direct latent optimization method so that an input image is modified in response to a user-provided sound input. We quantitatively and qualitatively demonstrate the effectiveness of our approach, and we observe our sound-guided image manipulation approach can produce semantically meaningful images.

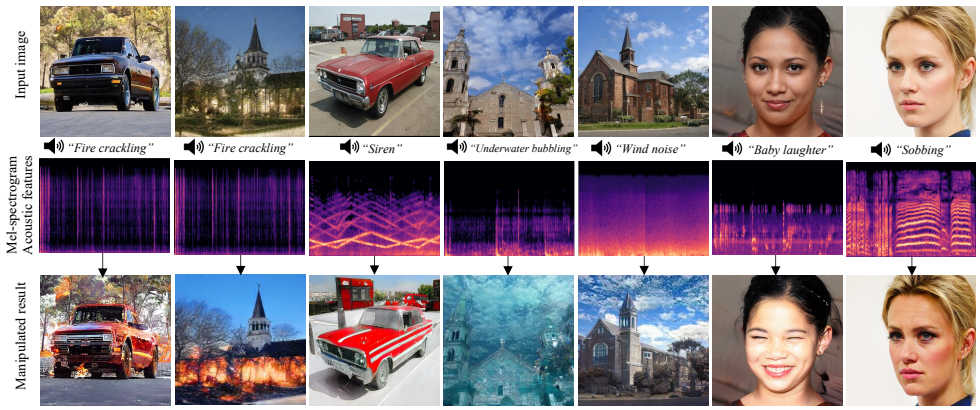

Figure 1: Our sound-guided image manipulation examples. Our model manipulates input images (top row) based on the user-provided sound inputs (e.g. fire crackling, siren, underwater bubbling, or wind noise), which produces sound-driven manipulated results (bottom row).

1 Introduction

Semantic image manipulation requires modifying the image appearance in response to a user-controlled way, while preserving contextual information as well as the realism of the result. Producing semantically meaningful images are further challenging as it involves laborious manual human examination for each desired manipulation. Recently, several approaches have been proposed for manipulating images through guidance of diverse styles [3, 11, 10, 12, 17]. Especially, a text-driven image manipulation method called StyleCLIP [12] considered leveraging the representational power of Contrastive Language-Image Pre-training (CLIP) [14] models to produce semantically meaningful manipulations given text input.

*Corresponding authors: J. Kim (jinkyukim@korea.ac.kr) and S. Kim (spk7@korea.ac.kr)

1st CtrlGen: Controllable Generative Modeling in Language and Vision Workshop at NeurIPS 2021.

In this work, we introduce a novel image manipulation approach that is driven by audio semantics, i.e. as shown in Figure 1, an image of an old car is manipulated into an old car with fire truck-like exterior appearance. Our model consists of two main stages: (i) the CLIP-based Contrastive Latent Representation Learning where an audio encoder is trained to produce a latent representation that is 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 input.

Our main contributions are listed as follows: (i) Expanding the modality space of CLIP to make shareable embedding space among audio, image, and text. (ii) Enabling semantic-level image manipulation solely based on the given audio information. (iii) Providing dynamic image manipulation using style-mixing latent codes guided by multi-modal data (audio and text).

2 Related Work

Interpreting Latent Space in StyleGAN. The intermediate latent space in pre-trained StyleGAN [8] solves the disentanglement issue and allows the generated images to be manipulated meaningfully according to changes in the latent space. GANSpace [6] and StyleSpace [16] allow image manipulation with interpretable controls from a pre-trained GAN generator. However, these works have not invested on the audio sequences. Our approach controls StyleGAN2's generator given an input audio to perform an sound-guided image manipulation.

Text-guided Image Manipulation. Text-guided image manipulation is the most widely studied among guidance based tasks. SISGAN [3] employed the GAN-based encoder-decoder structure to preserve the features of the image while presenting image manipulations corresponding to the text description. Several local text-adaptive discriminators in TAGAN [11] classify attributes independently and provide feedback to the generator. ManiGAN [10] proposes a multi-stage network for cross-modality representations of text and image through a text-image affine combination module. Unlike above works, StyleCLIP [12] and TediGAN [17] utilize the latent space of the pre-trained StyleGAN and the prior knowledge from CLIP. StyleCLIP performed image manipulation using three techniques including latent optimization, latent mapper, and global direction. TediGAN enabled image generation and manipulation using GAN inversion technique using multi-modal mapping.

Sound-guided Image Manipulation. Few approaches have been introduced for the sound-guided image manipulation task. Prior work mainly focuses on music, (instead of using sound semantics), including music-to-visual style transfer with cross-modal learning strategy [9], a neural music visualizer by mapping music embeddings to visual embeddings from StyleGAN [7], and audio-reactive latent vector interpolation [1]. Instead of using music, we focus on sound, which can convey more rich semantic information. Moreover, we leverage the representation power of existing CLIP models to train our audio encoder, which gives benefit of producing semantically meaningful results.

3 Method

We follow the existing text-guided image manipulation model, called StyleCLIP [12], which first trains the image and audio encoders to produce similar latent representations. After the pre-training step, encoders are frozen and used to manipulate images according to a target text input (e.g. images with different facial expressions can be manipulated with different text inputs). In this work, we add audio as extra modality source. Our audio encoder is trained (using a similar contrastive learning loss) to produce a latent representation that is aligned with pre-trained StyleCLIP's text and image encoders. As shown in Figure 2, such aligned representations can be used for image manipulation with the given audio input like StyleCLIP. Our model consists of two main steps: (i) the CLIP-based Contrastive Latent Representation Learning and (ii) the Sound-guided Image Manipulation. Details of each step are explained as follows.

CLIP-based Contrastive Latent Representation Learning Our audio encoder takes melspectrogram acoustic features as an input and produces a d-dimensional latent representation. We use ResNet-50 architecture as a backbone and adopt the contrastive learning approach. Here, the latent representations of a positive multi-modal pair (e.g. sound of fire and a text of "fire crackling") from audio and text encoders are pulled together while latent representations of a negative pair are pushed apart from each other. We employ following contrastive loss function $\mathcal{L}_{\text{con}}$ to train our audio encoder:

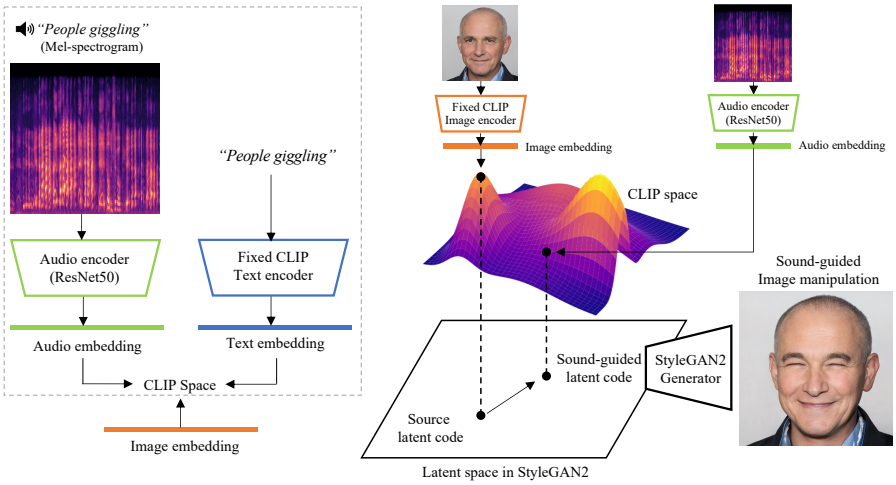

Figure 2: An overview of our proposed approach, which consists of two main steps: (i) the CLIP-based Contrastive Latent Representation Learning step and (ii) the Sound-Guided Image Manipulation step. In (i), we train a set of encoders with different modalities (i.e. audio, text and image) to produce the matched latent representations. Exemplary representations for a triplet pair (audio input: "Giggling", text input: "people giggling", and corresponding image) are pulled together in the embedding (CLIP) space (Left). In (ii), we use a direct code optimization approach. Here, a source latent code is modified in response to a user-provided audio input (Right).

$$\mathcal{L}_{\text{con}} = \sum_{i=1}^{N} \sum_{j=1}^{N} y \cdot \log(\mathbf{a}_i \cdot \mathbf{t}_j) + (1-y) \cdot \log(1 - \mathbf{a}_i \cdot \mathbf{t}_j) \tag{1}$$

where we optimize cosine similarity between the audio embedding $\mathbf{a} \in \mathbb{R}^d$ and the text embedding $\mathbf{t} \in \mathbb{R}^d$. We use an indicator variable y where we set $y = 1$ for the positive pairs, otherwise y = 0. N represents the batch size and $\cdot$ is the Euclidean dot product. We use over 200k audio sources from VGG Sound dataset [2] and over 2.1M from Audioset [5] for training our audio encoder. Note that the corresponding text prompts are extracted from the ground truth labels. The text prompt describes the occurrence of dynamic actions or events, such as "playing violin" or "baby laughing". Furthermore, we augmented text data by (i) replacing words with synonyms, (ii) applying a random permutation of words, and (iii) inserting random words. We find synonyms of given word from Wordnet [4] and insert the synonym anywhere in the sentence. For example, we augmente original texts ("rowboat, canoe, kayak rowing") to produce new texts ("row canoe, kayak quarrel rowboat").

Sound-guided Image Manipulation We employ the direct latent code optimization for sound-guided image manipulation by solving the following optimization problem:

$$\mathcal{L}_{man} = \underset{w_a \in \mathcal{W}}{\text{argmin}}\, d_{\text{cosine}}(G(w_a), a) + \lambda_{L_2} ||w_a - w_s||_2 + \lambda_{\text{ID}} L_{ID} \tag{2}$$

where a given source latent code $w_s \in \mathcal{W}$, input latent code $w_a \in \mathcal{W}$, and StyleGAN-based generator G. With such an optimization scheme, we minimize the cosine distance between the embedding vectors of the manipulated image $G(w_a)$ and the audio input. Therefore, the input image is modified in response to a user-provided audio input. The L_2 distance in latent space is used to regulate the similarity to the input image with a hyperparameter λ_{L_2}. Furthermore, the similarity to the input image is controlled by the identity loss $\mathcal{L}_{\text{ID}}$, which only changes a person's visual characteristics (e.g. giggling or crying) without shifting identity. Also, we disable the identity loss for other objects by setting $\lambda_{\text{ID}} = 0$.

4 Results

Qualitative Analysis. In Figure 3 (a), we provide examples of our audio-driven manipulation results (middle row). We adopt six different audio clips including "baby crying", "people giggling", "noise blowing", "explosion", "fire crackling", and "thunderstorm". Moreover, we compare our result with

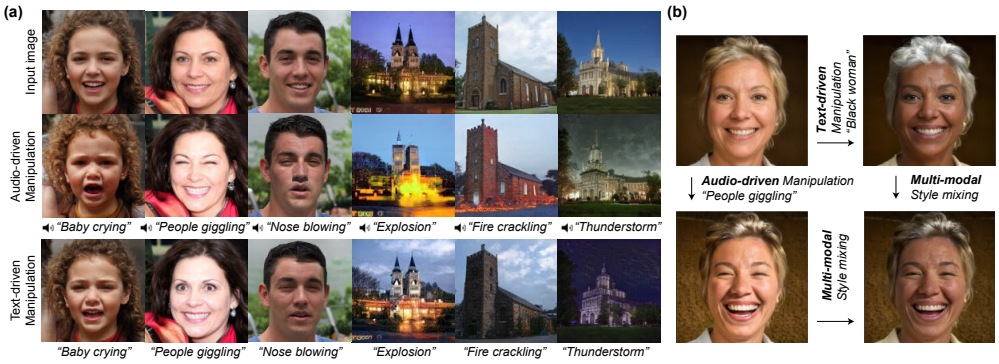

Figure 3: (a) Examples of our generated audio-driven manipulation (2nd row) from an input image (1st row). For comparison, we also provide results of text-driven manipulation (bottom row). The audio and text prompts are explained at the bottom of each image. (b) An example of image manipulation jointly with the audio ("people giggling") and text inputs ("black woman").

text-driven approach (bottom row). We observe that the proposed method successfully carry out image manipulation with the audio input while preserving identity and key visual attributes. Figure 3(b) demonstrates that proposed audio-driven approach could be used with text-driven manipulation as well. For instance, a text prompt "Black woman" and an audio input "people giggling" could manipulate a white woman portrait into a giggling black woman portrait. We showcase more diverse examples in the supplemental material (see Supplemental Figure 1 and Figure 2).

Quantitative Evaluation. We further analyze the effectiveness of our proposed audio-driven image manipulation approach quantitatively. First, we measure performance on the downstream 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, wind, fire crackling, underwater bubbling, explosion, and thunderstorm. As shown in Supplemental Table 1, we generally outperform existing text-driven manipulation approach with better semantically-rich latent representation. Then, we compare the cosine similarity between text-guided and sound-guided latent representations. As demonstrated in Supplemental Table 2, the latent representations generally exhibit a high-level characteristic of the content.

We also evaluate the distinguishability of the feature vector from the proposed audio encoder by comparing the downstream zero-shot classification task. As a baseline model, we use a ResNet50-based classifier, which is trained end-to-end from scratch (i.e. random initialization). We use two datasets: (i) Environment Sound Classification dataset (ESC-50) [13], which comprises of 2000 clips of 5 s length from 50 classes. Note that each of these clips was sampled at 44.1 kHz, with a length of 5 s. (ii) The UrbanSound8k dataset [15] contains 8732 clips from 10 classes. Each audio is less than 4 s long and is sampled at frequencies of 16 to 48 kHz. As shown in Supplemental Table 3, our audio encoder shows better classification performance than the baseline.

5 Conclusion

We propose a method for guiding the image manipulation direction with the intrinsic meaning of given audio input. We take the user-provided audio input into the latent space of StyleGAN2 and the CLIP embedding space where the latent code is aligned with the audio, enabling meaningful image manipulation reflecting the content of the audio. Our model produces a variety of manipulations based on various audio inputs relating to wind, fire, explosion, thunderstorm, rain, giggling, and nose blowing. We observe that an audio input can successfully provide a semantic cue to manipulate images accordingly.

Acknowledgement. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2017-0-00417, openholo library technology development for digital holographic contents and simulation) and IITP of the Korean government under grants No. 2019-0-00079 (Artificial Intelligence Graduate School Program of Korea University). J. Kim is partially supported by the National Research Foundation of Korea grant (NRF-2021R1C1C1009608), Basic Science Research Program (NRF-2021R1A6A1A13044830), and ICT Creative Consilience program (IITP-2021-2020-0-01819).

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020.

[7] D. Jeong, S. Doh, and T. Kwon. Tr\" aumerai: Dreaming music with stylegan. *arXiv preprint arXiv:2102.04680*, 2021.

[8] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

[9] 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.

[10] 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.

[11] S. Nam, Y. Kim, and S. J. Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. *arXiv preprint arXiv:1810.11919*, 2018.

[12] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021.

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

[14] 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. *arXiv preprint arXiv:2103.00020*, 2021.

[15] 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.

[16] Z. Wu, D. Lischinski, and E. Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021.

[17] 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.