Audio-Guided Image Manipulation for Artistic Paintings

Seung Hyun Lee[1], **Nahyuk Lee**[1], **Chanyoung Kim**[1], **Wonjeong Ryoo**[1]
Jinkyu Kim[2], **Sang Ho Yoon**[3*], **and Sangpil Kim**[1*]
[1]Department of Artificial Intelligence, Korea University
[2]Department of Computer Science and Engineering, Korea University
[3]Graduate School of Culture Technology, KAIST

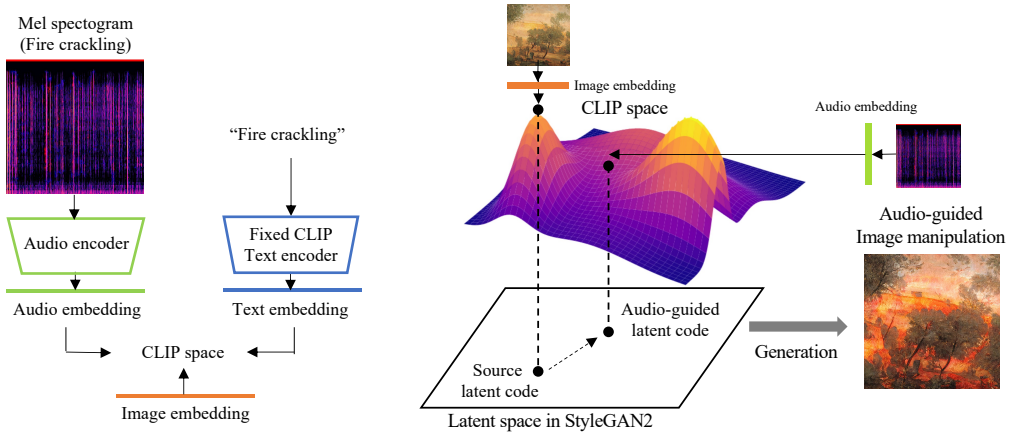

Figure 1: An overview of our proposed approach that consists of two main steps: (i) the *CLIP-based Contrastive Latent Representation Learning* step and (ii) the *Audio-Guided Image Manipulation* step. In (i), we train a set of encoders with a different modality (i.e. audio, text, and image) to produce the matched latent representations. E.g. representations for a triplet pair (the audio input "Fire crackling", the text input "Fire crackling", and the corresponding image) are pulled together in the embedding space (or CLIP space) (see left). In (ii), we use a direct code optimization so that a source latent code is modified in response to a user-provided audio input (see right).

1 Introduction

Generative Adversarial Networks (GANs) [2] can synthesize realistic high-quality images in diverse domains, and recent style-based generated models [5, 9, 10, 3] further suggest that its latent space can be utilized to manipulate synthetic and real images. A text-driven image manipulation method, called StyleCLIP [6], focused on leveraging the representational power of Contrastive Language-Image Pre-training (CLIP) models [7] to produce semantically meaningful latent manipulations given a text input (e.g. given the text prompt "surprised", a facial expression of a real celebrity portrait can be manipulated from neutral to surprised).

In this work, following this stream of work, we propose a novel audio-guided image manipulation approach for artistic paintings, which can generate semantically meaningful latent manipulations give an audio input (e.g. given an audio input "fire crackling", our model manipulates a landscape painting to a "fired" landscape painting as shown in Figure 1). To our best knowledge, our work is the first to explore generating semantically meaningful image manipulations from a variety of audio sources.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

*Corresponding authors: S. H. Yoon (sangho@kaist.ac.kr) and S. Kim (spk7@korea.ac.kr)

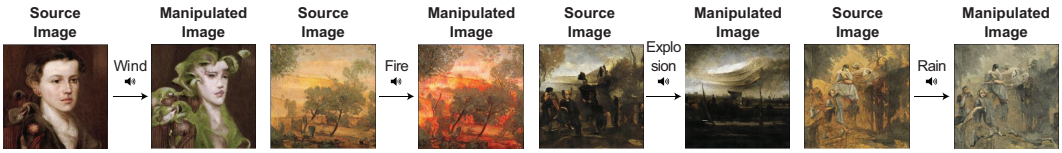

Figure 2: Examples of audio-guided image manipulations using our proposed method. Audio inputs used for conditioning each manipulation are explained above each arrow. Audible sounds and more diverse examples are available at `https://kuai-lab.github.io/aiartist`.

2 Method

We closely follow the existing text-guided image manipulation model, called StyleCLIP [6], which first trains the image and audio encoders to produce similar latent representations. After this 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 explore another source modality, audio. As shown in Figure 1, our audio encoder is first trained (using a contrastive learning loss) to produce a latent representation that is aligned with pre-trained StyleCLIP's text and image encoders. Such aligned representations can be used for image manipulation with the target audio input in a similar way to StyleCLIP.

CLIP-based Contrastive Latent Representation Learning. Our audio encoder takes mel-spectrogram acoustic features as an input and produces a d-dimensional latent representation. We use ResNet-50 [4] architecture as a backbone, and we use the contrastive learning approach: i.e. 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. Formally, we use the following contrastive loss function $\mathcal{L}_{con}$ to train our audio encoder:

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

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 set $y = 1$ for the positive pairs, otherwise $y = 0$. N represents the batch size and $\cdot$ is the Euclidean dot product. We use VGGSound video dataset [1] to train our audio encoder.

Audio-guided Image Manipulation. We use direct latent code optimization for audio-guided image manipulation by solving the following optimization problem:

$$\underset{w_a \in W+}{\operatorname{argmin}} \left(1 - \frac{I(G(w_a)) \cdot \mathbf{a}}{||I(G(w_a))||_2 \cdot ||\mathbf{a}||_2} \right) + \lambda_{\text{sim}} ||w_a - w_s||_2 \tag{2}$$

where a given source latent code w_s, we use a StyleGAN-based generator G and a CLIP-based image encoder I. With such an optimization scheme, we minimize the cosine distance between the embeddings of the manipulated image $G(w_a)$ and the audio input $\mathbf{a}$ so that an input latent code w_a is modified in response to a user-provided audio input. The $\mathcal{L}_2$ distance in latent space is used to regulate the similarity to the input image with a hyperparameter λ_{sim}. Note that our generator G is pre-trained with WikiArt [8] fine-art paintings dataset.

3 Result and Conclusion

As shown in Figure 2, our model can produce a variety of manipulations for art paintings conditioned on driving audio inputs, such as wind, fire, explosion, thunderstorm, rain, folk music, and Latin music. We observe in our experiment that an audio input can successfully provide a semantic cue to manipulate images accordingly. For example, given a wind sound, a portrait is manipulated with wind brown hair; a landscape painting is manipulated with fire and/or rain. We provide more diverse examples produced by ours at `https://kuai-lab.github.io/aiartist`.

Ethical Implications

The proposed model utilized given sound as an input source and the generated results could be biased from original paintings. Although this could bring satisfactory output for users, the generated results might imply intention contrary to original painters. Thus, users should be aware of it and respect the painter's intentions.

Acknowledgments and Disclosure of funding

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 Institute of Information & Communications Technology Planning & Evaluation (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. 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.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

Supplementary Document

Please visit our project page[2] for more information containing manipulated paintings with audio. We present experimental results of generated paintings using our proposed method. We selected seven audio clips: Fire, Rain, Wind, Explosion, Folk Music, Latin Music, and Thunderstorm for the experimental results. We use ten seconds of audio per clip for the artwork generation.

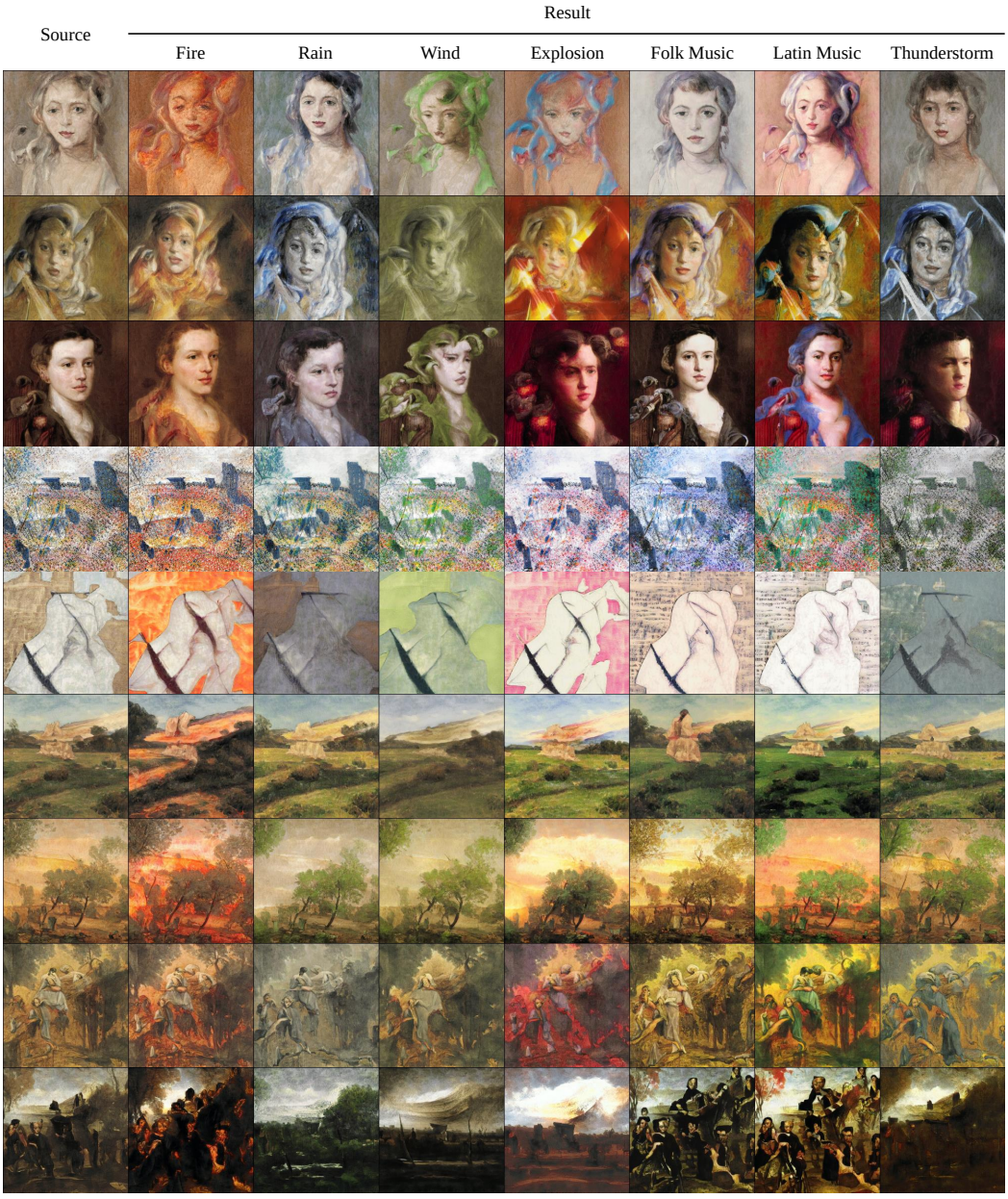

Figure 3: Source Column: Source paintings from WikiArt, Result Columns: Generated paintings with seven different kinds of audio clip

[2] https://kuai-lab.github.io/aiartist