Paper Review

LISA: Localized Image Stylization with Audio via Implicit Neural Representation

2023. 3. 24 (Fri)

Chanyoung Kim

Overall Review

- LISA: Localized Image Stylization with Audio via Implicit Neural Representation

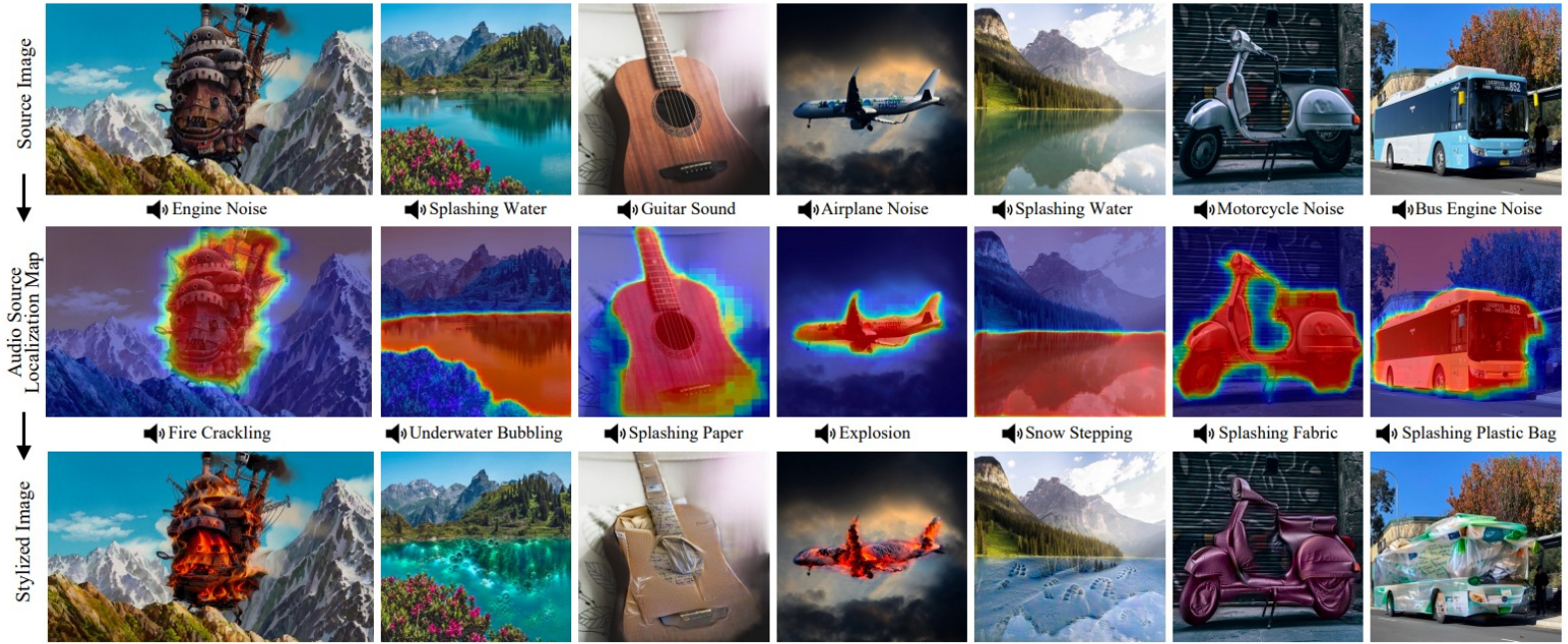


Figure 1. Examples from our localized image stylization based on audio inputs. Our model first localizes image regions corresponding to an audio input (e.g., given the sound of splashing water, our model localizes water from a source image). Conditioned on such a localization map, our model further stylizes the source image driven by another audio source (e.g., given an underwater bubbling sound, our model stylizes water as bubbling water).

Related Work (Sound-driven image editing)

- Sound-Guided Semantic Image Manipulation (CVPR 2022, Lee *et al.*)

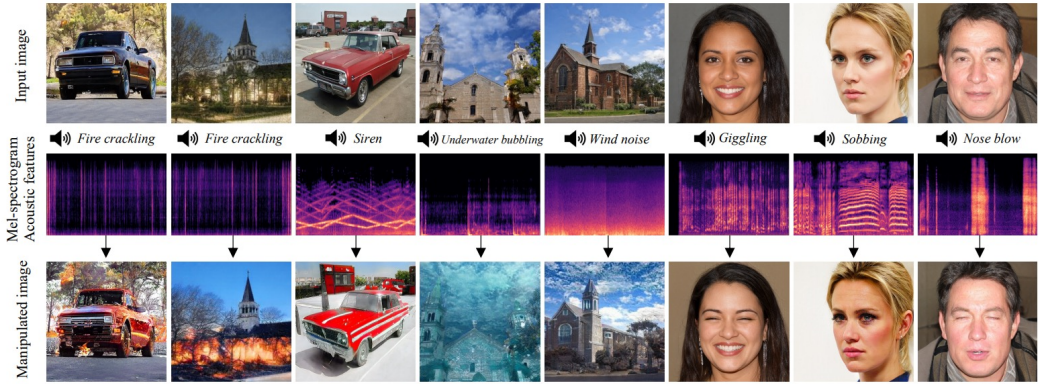
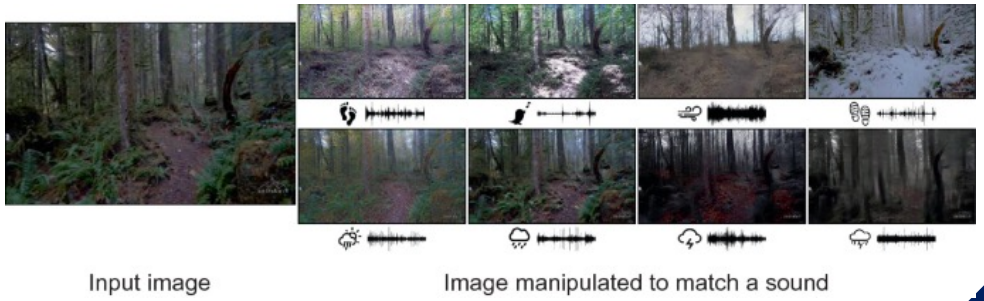


Figure 1: Modified images with sound-guided semantic image manipulation. Our method manipulates source images (top row) given user-provided sound (middle row) into semantic images (last row).

- LearningVisualStylesfromAudio-VisualAssociations (ECCV 2022, *Li et al.*)



Related Work (Sound-driven image editing)

- Sound-Guided Semantic Image Manipulation (CVPR 2022, Lee *et al.*)

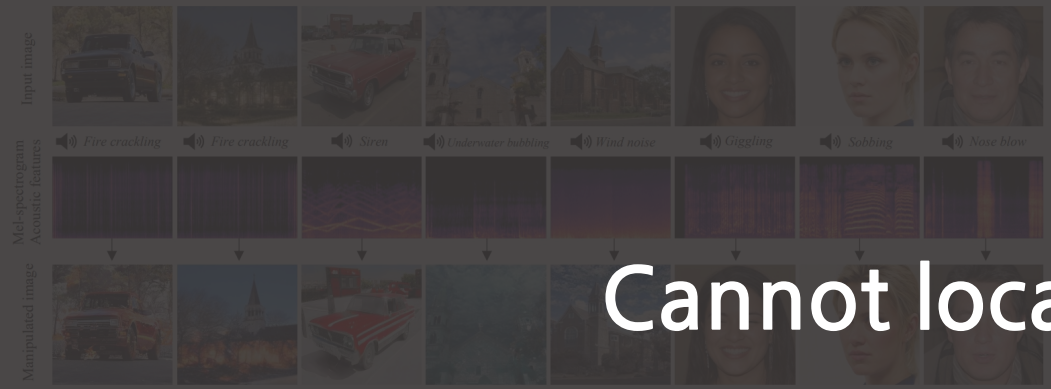
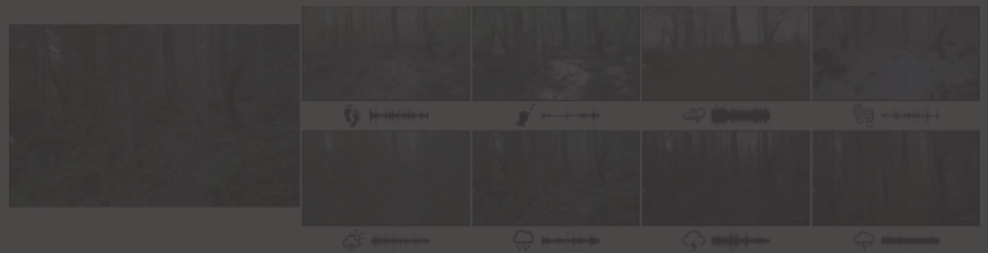


Figure 1: Modified images with sound-guided semantic image manipulation. Our method manipulates source images (top row) given user-provided sound (middle row) into semantic images (last row).

Cannot localize!

- LearningVisualStylesfromAudio-VisualAssociations (ECCV 2022, *Li et al.*)



Input image

Image manipulated to match a sound

Related Work (Localized Image Stylization)

- CBStyling: Real-time localized style transfer with semantic segmentation (ICCVW 2019, Kurzman *et al*.)

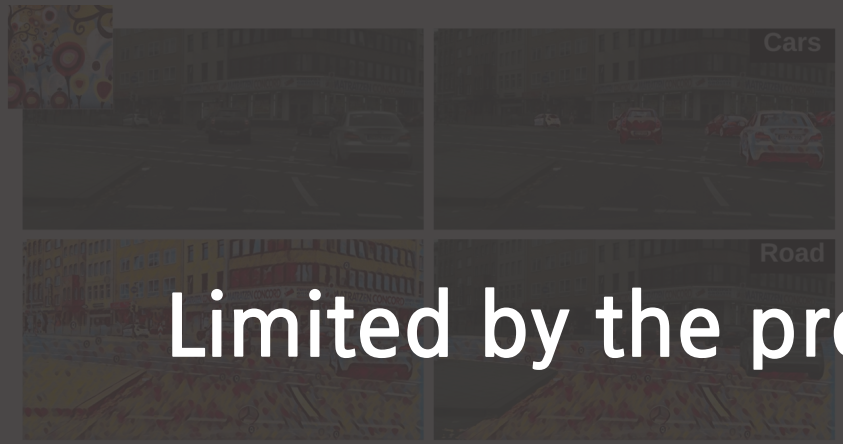
Global style transfer

Our real-time class-based styling method

- Real-time localized photorealistic video style transfer (WACV 2021)

Related Work (Localized Image Stylization)

- CBStyling: Real-time localized style transfer with semantic segmentation (ICCVW 2019, Kurzman *et al.*)

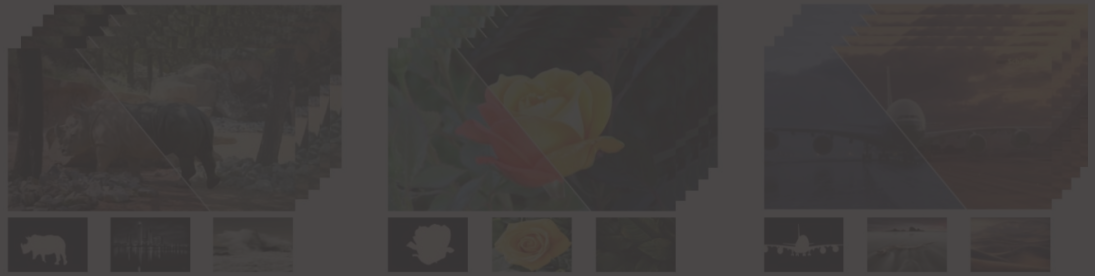


Global style transfer

Our real-time class-based styling method

Limited by the pre-defined categories!

- Real-time localized photorealistic video style transfer (WACV 2021)



Related Work (Text-driven image stylization)

- CLIPstyler: Image Style Transfer with a Single Text Condition (CVPR 2022, Kwon *et al*.)

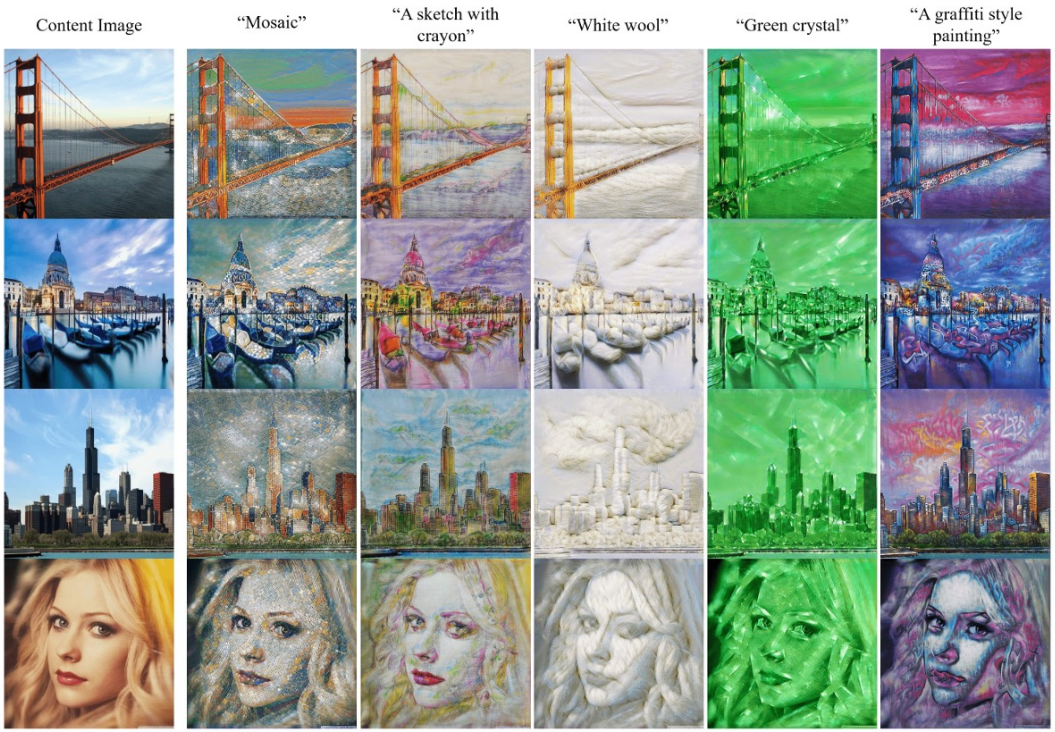


Figure 3. Style transfer results on various query text conditions. Our method can synthesize realistic textures which reflect the text conditions. Additional results are in our Supplementary Materials.

Related Work (Text-driven image stylization)

- CLIPstyler: Image Style Transfer with a Single Text Condition (CVPR 2022, Kwon *et al*.)

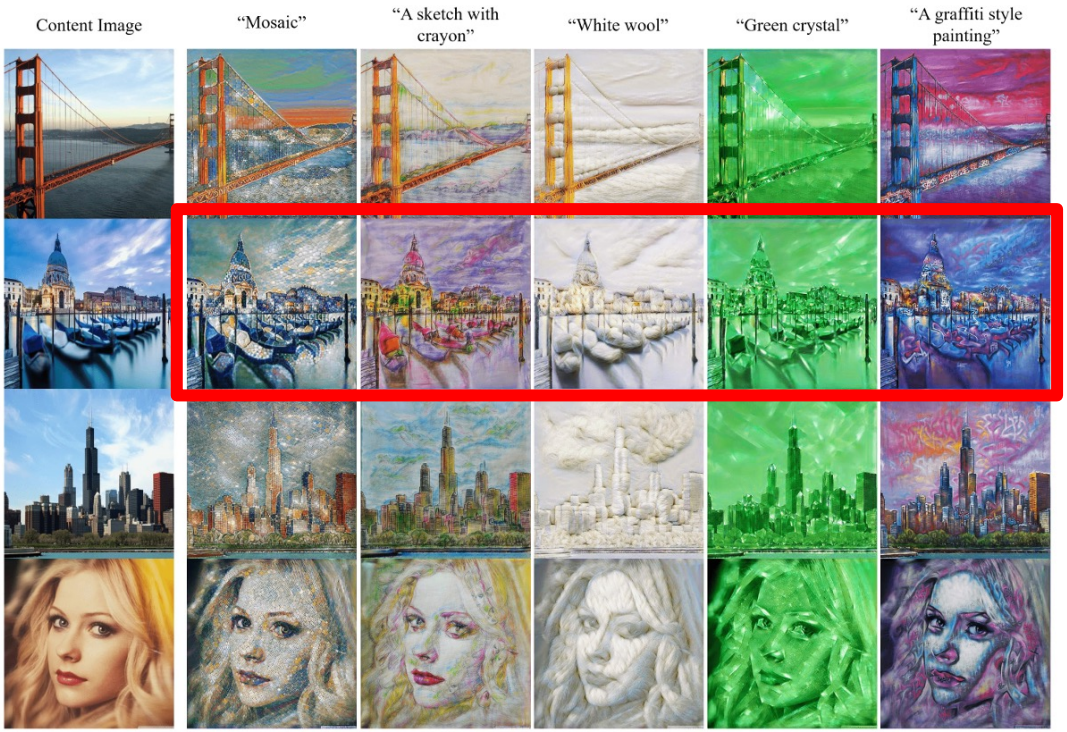
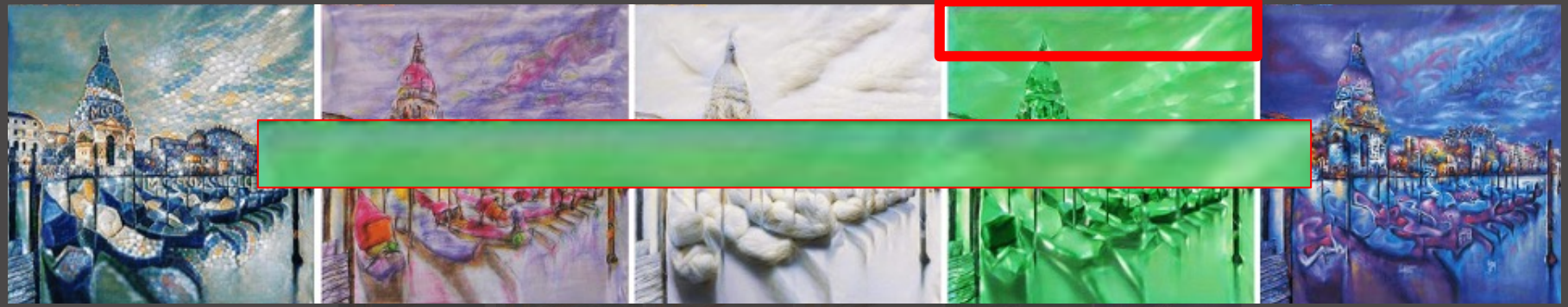


Figure 3. Style transfer results on various query text conditions. Our method can synthesize realistic textures which reflect the text conditions. Additional results are in our Supplementary Materials.



Border Effect!

Figure 3. Style transfer results on various query text conditions. Our method can synthesize realistic textures which reflect the text conditions. Additional results are in our Supplementary Materials.

Main Contributions

- We propose a novel **audio-guided image localized stylization**. In particular, our method is able to perform partial stylization **without manually producing a segmentation mask** by using sound source localization.

- We carefully design **implicit neural representation for style transfer to achieve naturalness**. We demonstrate that this structure is **effective in removing boundary artifacts**. With INR, our method can stylize the source image at **any arbitrary resolution**.

- We achieve **state-of-the-art performance** compared with the existing sound source visual localization methods. Using a pseudo-ground-truth segmentation mask from images and text joint representation, we **improve the quality of the audio-visual source localization** that generates concise object or scene boundaries.

Method

- Overview

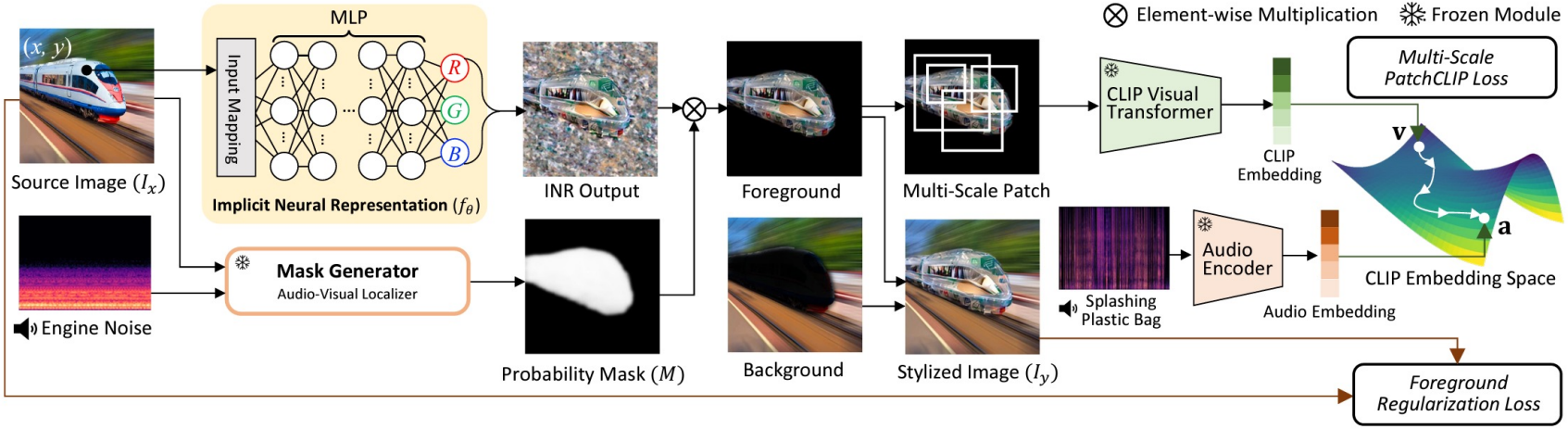
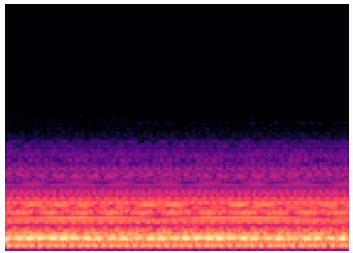


Figure 2. An overview of our proposed method called Localized Image Stylization with Audio (LISA). Our model consists of two main parts: (i) Audio-Visual Localizer, which outputs a pixel-level localization mask conditioned on an audio input (e.g., given a sound of engine noise input, our model localizes a train from the source image, producing a probability mask) and (ii) Audio-Guided INR Stylizer, which outputs stylized images by taking pixel locations as input and producing RGB pixel values as output. Conditioned on a new user-provided sound input (e.g., splashing plastic bag), our model is optimized with multi-scale PatchCLIP loss to generate an audio-guided "locally" stylized image. We also use Foreground Regularization Loss to make the stylized image and a source image perceptually look similar.

Method

- Mask Generation Process

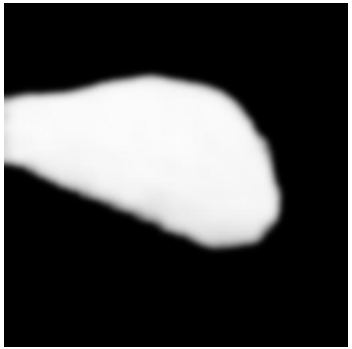
Source Image (I_x)



Engine Noise

Mask Generator

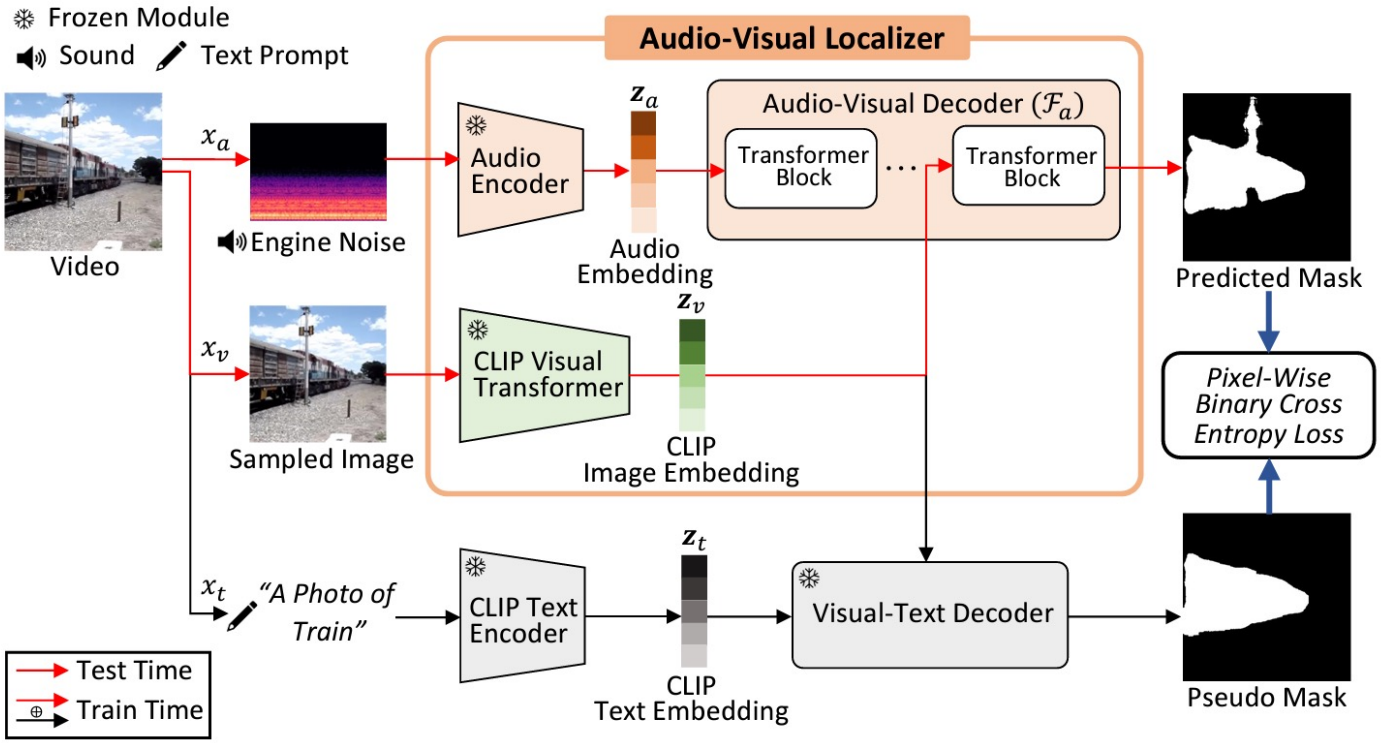
Audio-Visual Localizer



Probability Mask (M)

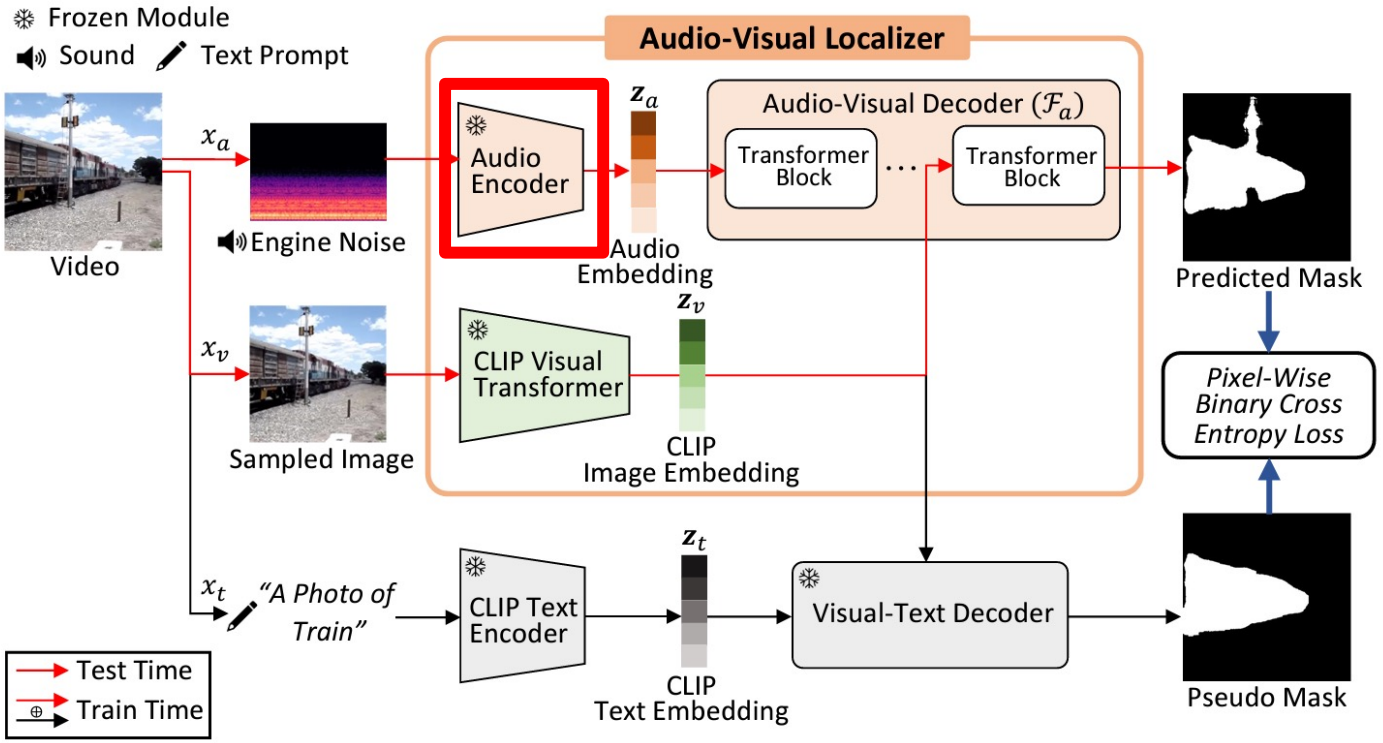
Method

- Pre-Training Audio-Visual Localizer by Weakly Supervised Learning



Method

- Pre-Training Audio-Visual Localizer by Weakly Supervised Learning



Method

- Pre-Trained Audio Encoder (from Sound-Guided Semantic Image Manipulation, CVPR 2022 *Lee et al.)*

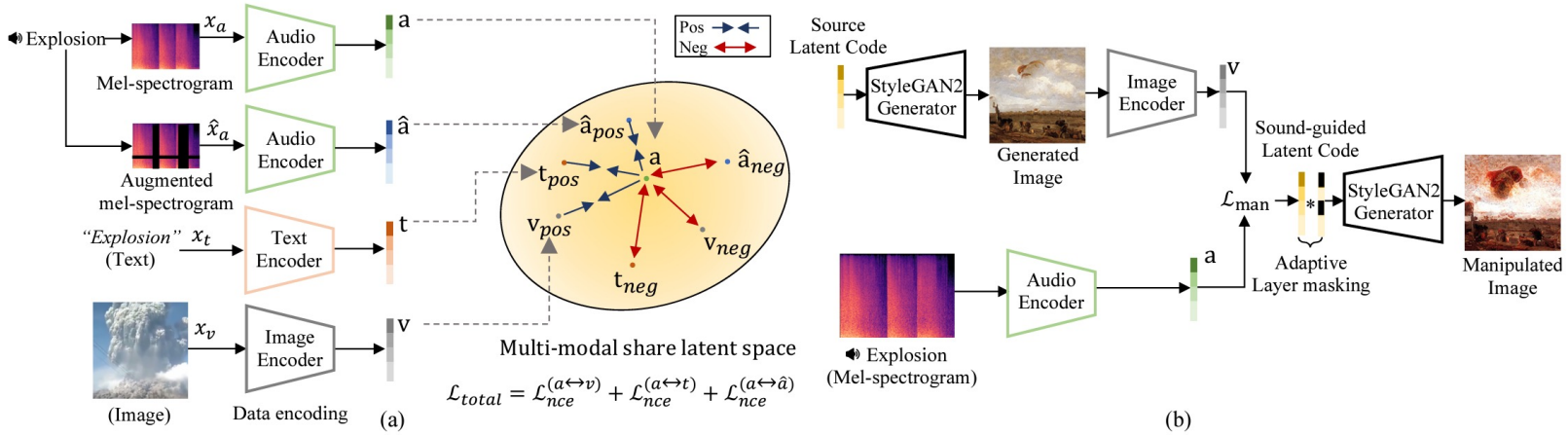
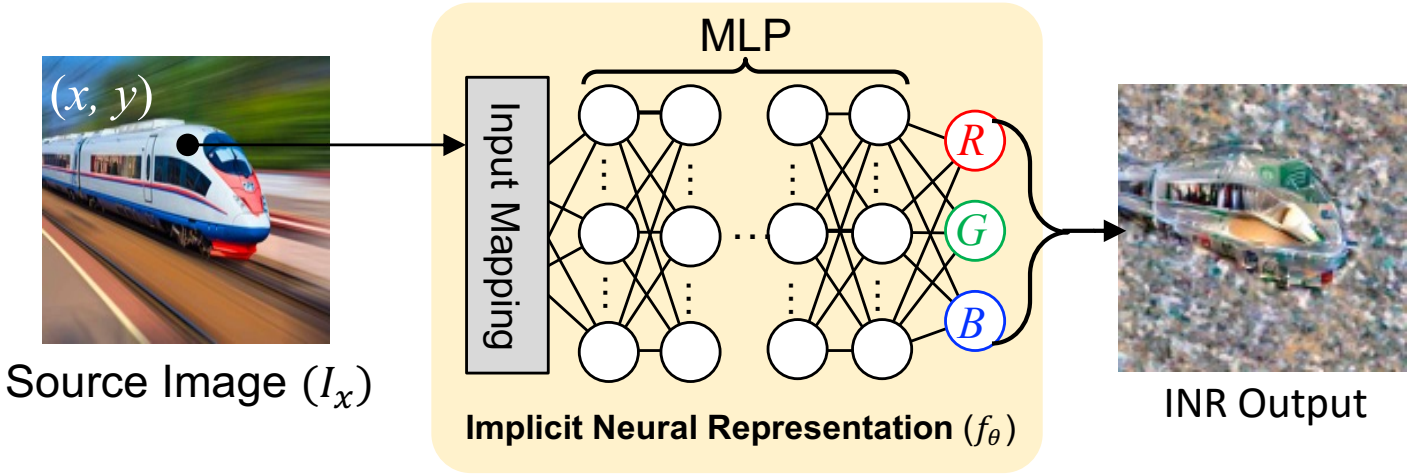


Figure 2: Our model consists of two main steps: (a) the *CLIP-based Contrastive Latent Representation Learning* step and (b) the *Sound-Guided Image Manipulation* step. In (a), we train a set of encoders with three different modalities (audio, text, and image) to produce the matched latent representations. 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).

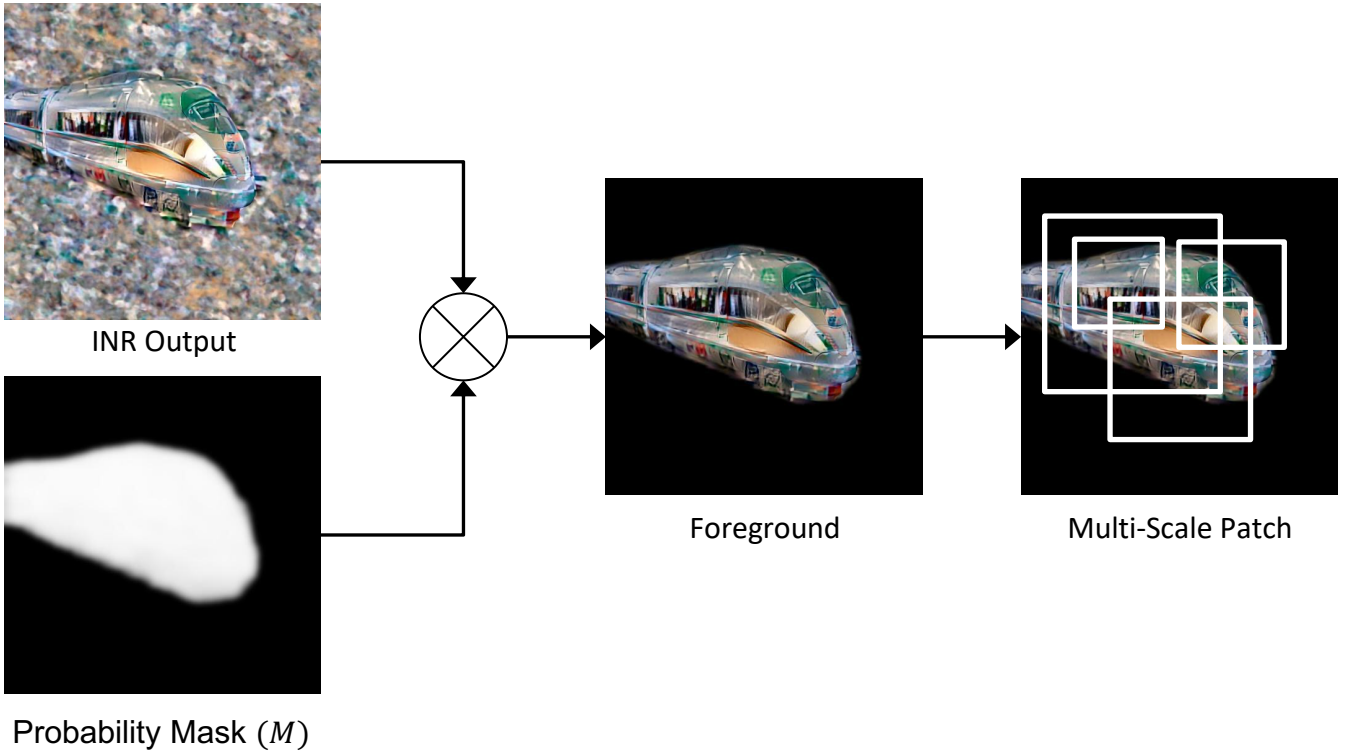
Method

- Implicit Neural Representation – Coordinate to RGB value



Method

- Multi-Scale Patch Sampling

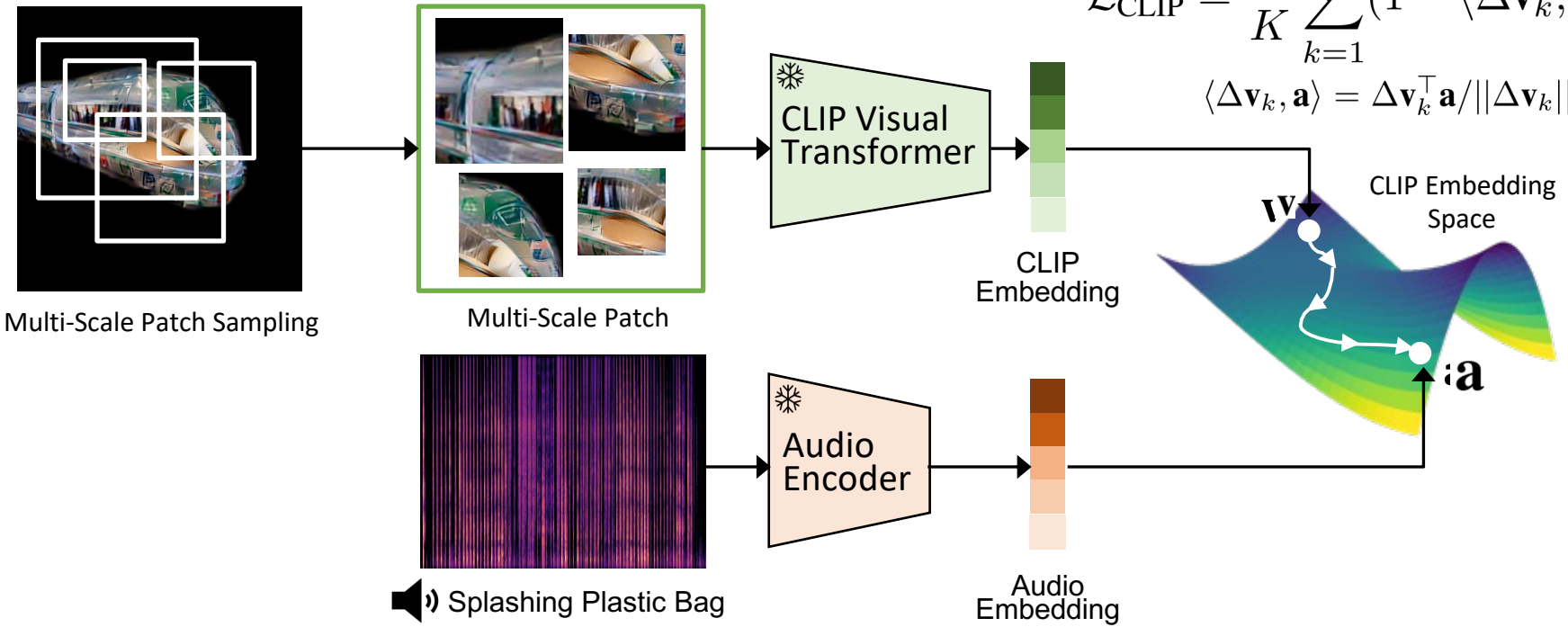


Method

- Multi-Scale PatchCLIP Loss

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{K}\sum_{k=1}^{K}(1 - \langle \Delta \mathbf{v}_k, \mathbf{a} \rangle)$$

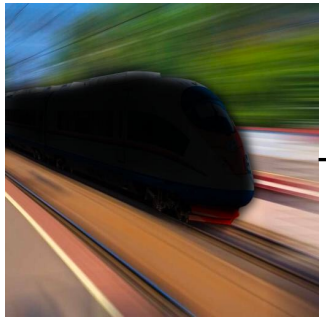
$$\langle \Delta \mathbf{v}_k, \mathbf{a} \rangle = \Delta \mathbf{v}_k^{\top} \mathbf{a} / ||\Delta \mathbf{v}_k|| ||\mathbf{a}||$$



Method

- Foreground Regularization Loss

Foreground



Background

Stylized Image (I_y)

$$\mathcal{L}_{\text{reg}} = \sum_{k=0}^{K} |G_k^{\phi}(\mathcal{I}_y) - G_k^{\phi}(\mathcal{I}_x)|$$

Source Image (I_x)

Result

- Effect of INR

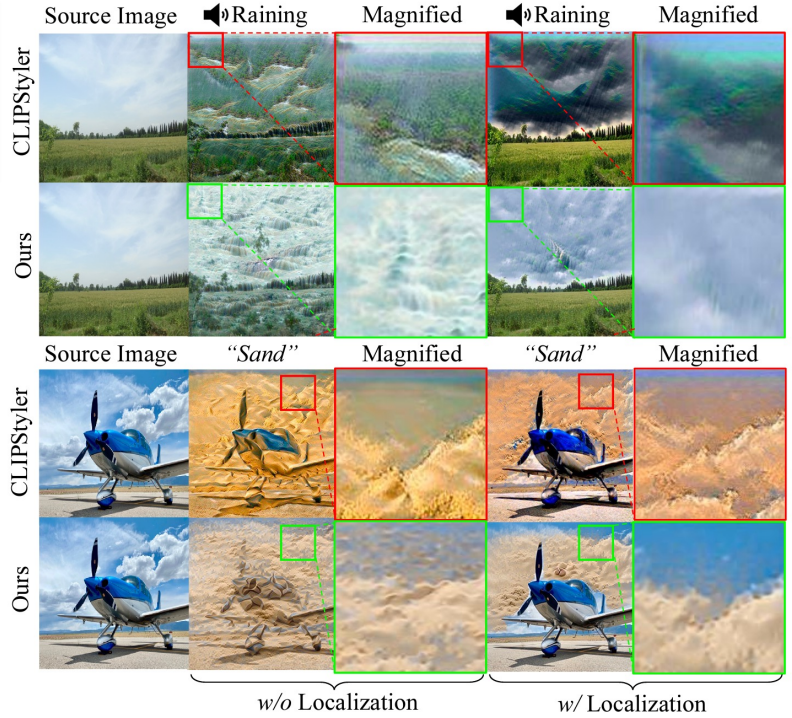


Figure 8. Stylization quality comparison between conventional U-Net-based [20] vs. INR-based approach (ours). Note that the former suffers from artifacts near borders, while the INR-based approach generates clearer images.

Result

- Localized Image Stylization

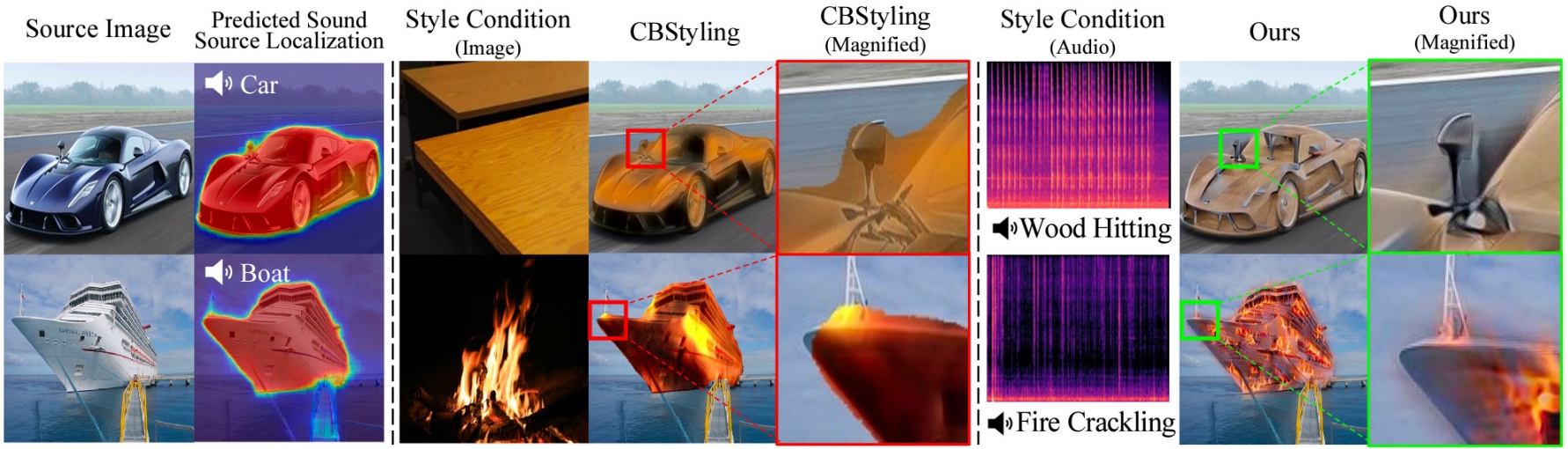
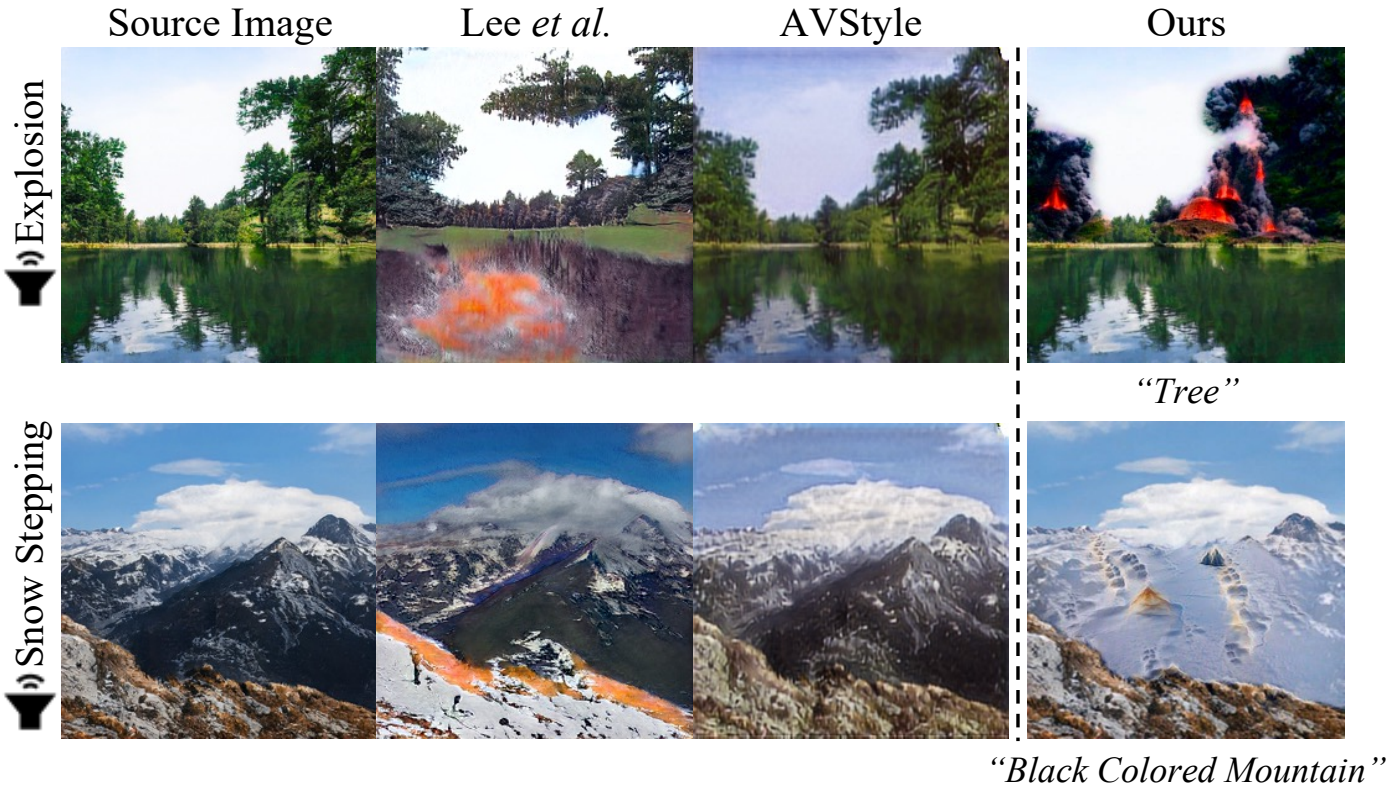


Figure 4. Comparison of our method to other existing localized neural style transfer, CBStyling [19] which both utilize localization mask. For a fair comparison, we apply the same sound source localization map created with our method to CBStyling and ours. Our method produces a more plausible effect on the edge of the predicted mask than the baseline.

Result

- Audio-Guided Image Stylization



Result

- Audio-Visual Localization Quality

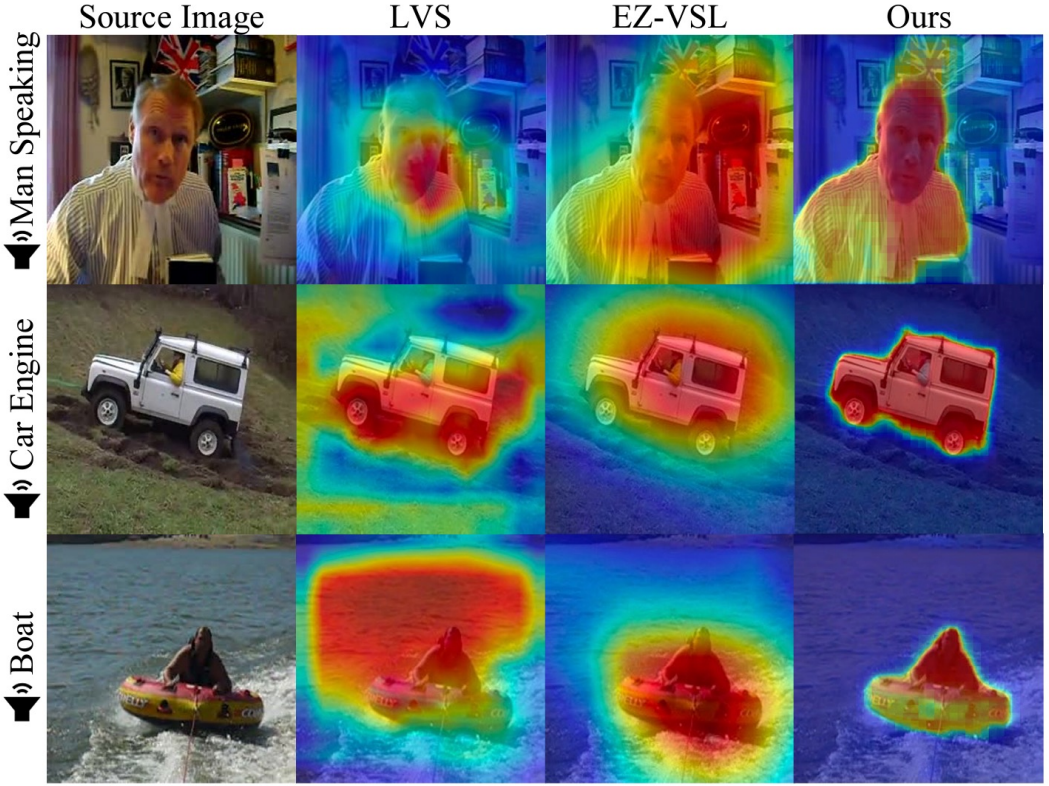


Table 1. Quantitative comparison between the existing audio-visual sound source localization methods and ours. We report CIoU and AUC for sound source localization (SSL).

Model	SSL Metric	
	CIoU (↑)	AUC (↑)
LVS [5]	73.59 %	59.00 %
EZ-VSL [27]	83.94 %	63.60 %
Ours	**85.94** %	**66.70** %

Result

- Text-Driven Localized Image Stylization

  * Our model can perform image stylization under various audio and text conditions because our audio latent representation shares with CLIP embedding space.

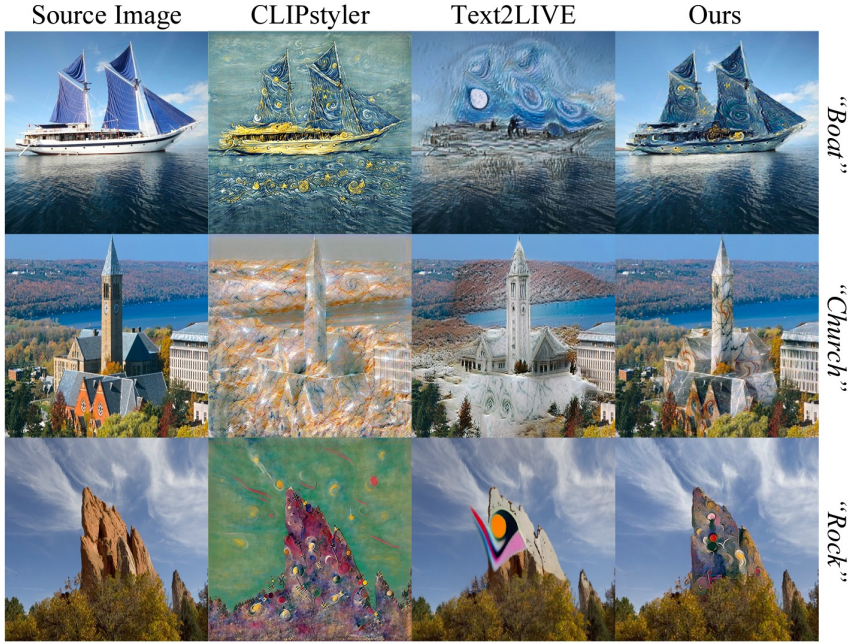


Figure 6. Stylization output comparison with existing text-driven stylization approaches (CLIPstyler [20] and Text2LIVE [3]). We set *The Starry Night by Vincent Van Gogh*, *Marble*, and *Composition VII by Wassily Kandinsky* as style conditions (respectively, from top to bottom).

Showcase


Source Image
Fire Crackling
Waterfall
Snow
"Cyberpunk"
"Bamboo"
"Stainless Steel"
Boat Sound
Train Sound
"Rock"
"Church"

Thank you

Any question?