Seamless Style Transfer in Videos: Enhancing Clarity and Consistency with Stable Diffusion, EBSynth, and ImageBind

1st Wei Hao Lu
Dept of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
r11944049@ntu.edu.tw

1st Po Yen Chou
Dept of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
r11922131@ntu.edu.tw

1st Jun Chen Hong
Dept of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
r11944050@ntu.edu.tw

1st Hong Yang Chang
Dept of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
r11944037@ntu.edu.tw

***Abstract*—In this work, we address the challenges of generating naturalistic translated videos and achieving accurate video style transfer. We propose a modular pipeline that combines techniques from computer vision and image manipulation to enhance the quality and texture consistency of translated videos. Leveraging a neural network trained on unpaired data, our approach maps elements of one video to another, resulting in high-quality translated videos.**

***Index Terms*—video style transfer, image similarity**

I. Introduction

With the widespread use of social media platforms and the increasing demand for visually captivating content, style transfer techniques have gained significant attention in the field of computer vision and graphics. Style transfer allows the transformation of the visual style of an image or video while preserving its content, leading to stunning and artistic visual effects.

In the realm of image style transfer, there have been notable advancements in generative approaches. These models have gained popularity for their ability to generate high-quality images. Stable diffusion has demonstrated impressive results in producing image fields. However, the generation of transfer style videos poses significant challenges. Two main obstacles arise in this context. The first challenge involves maintaining consistency across video frames, ensuring smooth and coherent transitions between each frame. The second challenge is accurately transferring the desired style to the video, ensuring that the style remains faithful throughout the entire sequence. Addressing these challenges is an active area of research in the field of video style transfer.

Motivated by EbSynth, which synthesizezing images based on reference images. We ensure the consistent of reference images by ImageBind. In this work, we present an innovative pipeline for seamless artistic style transfer in videos. Our approach combines the power of stable diffusion, EbSynth, and ImageBind to achieve optimal image clarity throughout the style transfer process.

II. Related work

A. Image style transfer

Stable Diffusion [6] is a deep learning, text-to-image model released in 2022. It is primarily used to generate detailed images conditioned on text descriptions, though it can also be applied to other tasks such as inpainting, outpainting, and generating image-to-image translations guided by a text prompt. Stable Diffusion is a latent diffusion model, a kind of deep generative neural network. We can use different models to stylize generated images. In addition to that, the utilization of extensions like ControlNet [1] and LoRA [5] allows us to enhance the quality of generated images further.

Controlnet [1] is a end-to-end neural network architecture that can enhance large image diffusion models (like Stable Diffusion) with task-specific conditions. The ControlNet clones the weights of a large diffusion model and connects them with a convolution layer called "zero convolution", where the convolution weights progressively grow from zeros to optimized parameters. Since the zero convolution does not add new noise to deep features, the training is as fast as fine tuning a diffusion model.

B. Video style transfer

Gen-1 [3] primarily focuses on transforming existing video footage, allowing users to modify object colors, image styles, and more using text commands. It employs an end-to-end approach, directly applying style transfer to the original video, transferring any given style or image style to each frame

of the video. However, it may be more challenging to have precise control over the specific style being applied during the conversion process.

EbSynth [4] is a versatile tool designed for synthesizing images based on reference image. It offers a wide range of applications, including guided texture synthesis, artistic style transfer, content-aware inpainting, and super-resolution.

The primary focus of EbSynth is to maintain the fidelity of the source material. Unlike recent approaches that heavily rely on neural networks, EBSynth employs state-of-the-art non-parametric texture synthesis algorithms. By leveraging its patch-based methodology, EbSynth achieves sharp and detailed results, faithfully preserving the fine details of the original image.

C. Similarity

ImageBind [2] is an innovative research achievement. It introduces a novel approach that enables direct calculation of similarity between different objects without the need for prior conversion into specific formats like Multimodal Learning. This means it can leverage different types of data, such as images, text, and audio, without the need to convert them into a unified representation. This approach better preserves the characteristics and information of the original data, leading to more accurate and comprehensive similarity calculations.

III. Method

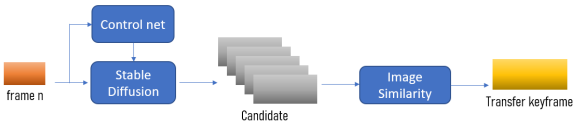

Fig. 1. Generate the style-transferred images and select the keyframe.

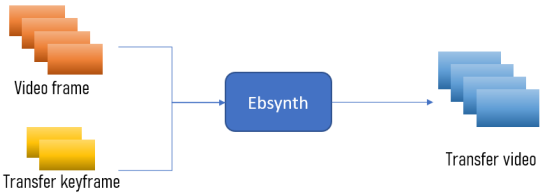

Fig. 2. Stylize entire video frames based on multiple keyframes.

In our method, we begin by generating the style-transferred images(refer to Figure 1), followed by selecting the most similar image as the keyframe (refer to Figure 3). Subsequently, we utilize Ebsynth to create the style-transferred video (refer to Figure 2).

A. Style Image Generation

Stable Diffusion is a powerful image generative model,that can generate detailed images conditioned on text descriptions or generating image-to-image translations guided by a text prompt. Compared to other generative models, advantages of Stable Diffusion is that it is open source, allowing users to run it for free on their own computers or servers and has a strong community support.

We use Stable Diffusion image-to-image model to generate images for the keyframe in the video. In order to ensure the quality of generated images,we use ControlNet to enhance the similarity between the source keyframe and generated image.

B. ImageBind on Keyframe Selection

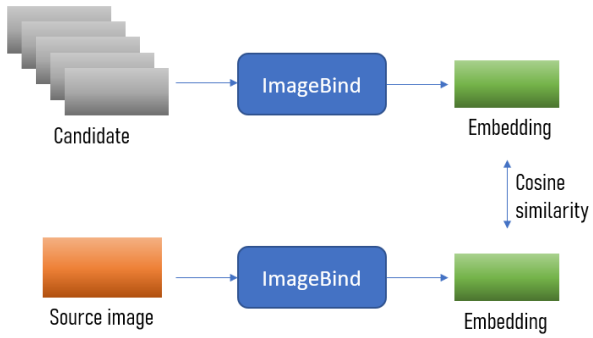

Fig. 3. Select the most similar image as the keyframe by ImageBind.

We focus on imagebind, which enables direct extraction of embeddings from different images. This approach helps us avoid information loss and saves computational time by eliminating the need for conversion processes. We utilize these embedded vectors to calculate the similarity between various candidate images generated through stable diffusion. We pick the most similar images are selected as keyframes to synthesize a style transfer video.

C. Ebsynth for Video Style Transfer

EbSynth is a powerful algorithmic video stylizer known for its high-quality results and versatility. It can stylize entire videos based on a single keyframe. However, it struggles with visual shifts in videos, often resulting in inaccurate stylization and unnatural appearances. To overcome this limitation, new stylized keyframes need to be introduced for each visual shift. Despite this drawback, EbSynth exhibits state-of-the-art video synthesis capabilities when used with appropriate keyframes, making it an impressive tool for video-to-video translation with adaptability to various content.

IV. Experiments

A. Experiments detail

Our experiments are conducted on stable-diffusion-webui[1] and Ebsynth[2]. We use stablediffusion+VAE+controlnet[3] to generate our result.We use orangemixs[4] model for stable

[1]https://github.com/AUTOMATIC1111/stable-diffusion-webui
[2]https://github.com/jamriska/ebsynth
[3]https://github.com/Mikubill/sd-webui-controlnet
[4]https://huggingface.co/WarriorMama777/OrangeMixs

diffusion and VAE,control sd15 canny[5] model for ControlNet.

```
img2img_payload = {
    "init_images": [img_base64],
    "prompt": prompt,
    "negative_prompt": negative_prompt,
    "denoising_strength": 0.7,
    "width": 512,
    "height": 768,
    "cfg_scale": 7,
    "sampler_name": "DPM++ 2M Karras",
    "restore_faces": False,
    "steps": 29,
    "batch_size":5,
    "n_iter":2,
    "alwayson_scripts": {
        "controlnet": {
            "args": [

                {
                    "input_image": img_base64,
                    "module":"canny",
                    "model": "control_sd15_canny",
                    "resize_mode": "Crop and Resize",
                    "pixel_perfect":True,
                    "weight":1.0,
                    "control_mode": "My prompt is more important",
                }

            ]
        }
    }
}
```

Fig. 4. config for generation

In Figure 4, the provided configuration for the generation process is displayed. .We use 29 steps to generate image and generate 5 images per iteration. For each keyframe, 10 candidate images are generated, and the most similar candidate is selected as the keyframe for imagebind. Furthermore, Ebsynth is utilized to generate every frame from multiple keyframes.

B. Result

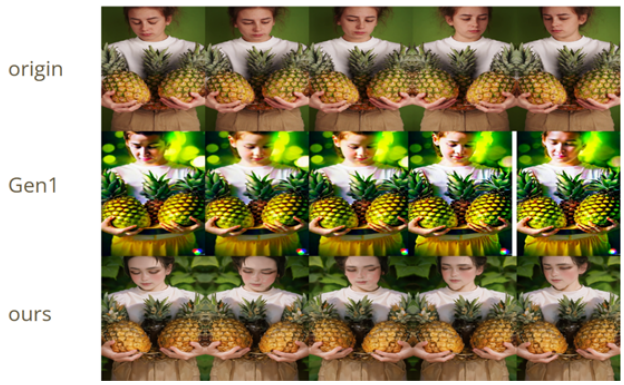

Fig. 5. continuous 5 timesteps image of different generation strategy's result

We want to transfer the origin video to animate style. In Figure 5, the image depicts the disparity between our output using EbSynth and Gen1. The result from Gen1 appears to be brighter in comparison to the original frame, and there is a noticeable alteration in the woman's appearance. Specifically, in Gen1's result, the woman exhibits features that resemble an Asian ethnicity, whereas the original frame portrays her as European. Additionally, there is a significant distinction in the clothing. In Gen1's result, the woman's pants have transformed into yellow cotton pants, whereas in the original frame, they resemble brown denim jeans.

Based on these two observations, it becomes evident that the image generated by EbSynth exhibits greater precision compared to the image generated by Gen1. Moreover, subjectively speaking, the result from EbSynth appears more natural when visually examining the picture. These findings indicate that our video generation method has made significant advancements in terms of precision and overall performance.

V. CONCLUSION

This work has explored an innovative approach to the challenges inherent in video style transfer, a burgeoning field within computer vision and graphics. We have proposed a new pipeline that amalgamates the capabilities of Stable Diffusion, EbSynth, and ImageBind. Our approach addresses the issues of frame consistency and accurate style preservation, two significant hurdles in this area. Our method begins by generating style-transferred images using Stable Diffusion, an open-source image generative model with powerful capabilities. The ControlNet ensures the quality and stylistic consistency of these generated images. The technique of ImageBind is then employed to extract direct embeddings from these images, selecting the most similar images as keyframes based on calculated similarity. These keyframes are finally synthesized into a style-transferred video with the aid of EbSynth.

Our proposed solution not only mitigates the primary challenges of video style transfer but also provides an optimal image clarity and smooth transitions between frames. The implications of this research extend beyond academic interest, with potential widespread application in the growing demand for stylistically appealing video content, particularly on social media platforms. Future work will aim to enhance this pipeline's efficiency further and explore potential applications in other areas of computer vision and graphics.

VI. CONTRIBUTION

The stable diffusion[6], controlnet[7], imagebind[8] and Ebsynth[9] part was obtained from a git clone.How to connect these function together and the transition and processing between images and videos were done by us.

REFERENCES

[1] Lvmin Zhang, Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models, arXiv.org.arXiv preprint arXiv:2302.05543, 2023.

[2] Girdhar, Rohit and El-Nouby, Alaaeldin and Liu, Zhuang and Singh, Mannat and Alwala, Kalyan Vasudev and Joulin, Armand and Misra, Ishan. Imagebind: One embedding space to bind them all. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2023

[3] Esser, P. et al. Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, Anastasis Germanidis, Structure and Content-Guided Video Synthesis with Diffusion Models.arXiv preprint arXiv:2302.03011, 2023.

[5]https://huggingface.co/lllyasviel/ControlNet/tree/main/models

[6]https://github.com/AUTOMATIC1111/stable-diffusion-webui

[7]https://github.com/Mikubill/sd-webui-controlnet

[8]https://github.com/facebookresearch/ImageBind

[9]https://github.com/jamriska/ebsynth

[4] O. Jamriska. Ebsynth: Fast example-based image synthesis and style transfer. https://github.com/jamriska/ebsynth, 2018.
[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2021.
[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models.arXiv preprint arXiv:2112.10752, 202.