Generative AI-based VFX Production Pipeline

1st Jung-Jae Yu
Content Research Division
ETRI
Daejeon, Republic of Korea
jungjae@etri.re.kr

2nd Soonchul Jung
Content Research Division
ETRI
Daejeon, Republic of Korea
s.jung@etri.re.kr

3rd Dae-Young Song
Content Research Division
ETRI
Daejeon, Republic of Korea
eadyoung@etri.re.kr

***Abstract*—The biggest challenge when creating VFX using generative AI is that the background area is prone to changing in addition to the VFX elements being created. To address these issues, we propose a two-step approach consisting of 'Image-to-Image VFX' and '(Image+Video)-to-Video VFX'. And to realize this approach on the training data generated by high-speed real-time rendering, we also propose a method to automatically generate ROI areas for Image-to-Image VFX. Despite applying these methods, the generated image contained subtle flickering in the background. To address this, a final post-processing was applied: blending the background region of the original image. The resulting image was composed solely of VFX, preserving the background.We applied the proposed method to real-life footage and confirmed that it yielded good results.**

***Index Terms*—Generative AI, VFX, production, pipeline, image-to-image, image-to-video, automatic**

I. Introduction

Visual Effects (VFX), which replaces scenes that are difficult to implement in reality, such as fire, explosions, and floods, is a major issue in the film industry and has been a traditional research topic in the field of computer graphics.Conventional VFX methods that utilize particle simulation-based computer graphics technology required skilled workers to perform precise parameter setting work to achieve high-quality results. Recently, as generative AI technology has developed, try to use it to produce VFX has started. However, the biggest challenge when trying to utilize generative AI in VFX production is that the background area of the original footage also changes and flickering frequently occurs. This paper aims to solve these problems and proposes the following three main methods.

- To generate VFX from input live-action videos, we propose a two-step approach consisting of Image-to-Image and (Image+Video)-to-Video.
- To train an Image-to-Image generation model from training data built with high-speed real-time rendering, we propose a method to automatically extract ROIs for generating VFX.
- In the final generated image, we propose a method of blending the original background area excluding the VFX area to remove flickering in the background area.

II. Proposed Method

A. Two-step Approach for VFX Production

To achieve high-quality video VFX results, we first generate a VFX image from the starting frame using Image-to-Image generation method, and then apply (not exactly,) Image-to-Video generation using this as the starting frame. To be more specific, the worker first manually masks the ROI region in the first frame, and then performs Image-to-Image generation(I2I VFX) of [1] within that region. Using this VFX image as the starting frame, Image-to-Video processing is performed, and the SVD model of [2] is used to generate the video. Since the original video is input to an additional input channel during this process, strictly speaking, it can be said to be (Image+Video)-to-Video generation. ((I+V)2V VFX) Fig. 1 shows this two-step approach. The currently released SVD model can generate videos of 14 or 25 frames at a time. And let me add one more thing, if you repeatedly use the two step approach above, you can theoretically composite VFX for infinitely long footage. However, after the first clip, you can use the last frame generated from the previous clip as the starting frame to apply the (Image+Video)-to-Video VFX of the next clip without having to use image-to-image VFX using manual masking.

B. ROI Masking for Fire-flame in Learning Data

The two-step approach described above requires manual masking to apply I2I VFX. While this can be accomplished by manually drawing and inputting the masking area during the inference phase, creating such masking areas manually every time from the large amount of training data is challenging during the training phase. Of course, when creating learning data with computer graphics, you can also consider the method to generate masking information using the VFX layer. However, when using a high-speed real-time rendering service such as Ambergen [3] to reduce costs during the process of building large-scale learning data, there is a limitation that such VFX layers cannot be extracted separately. In these cases, we need a post-processing technology that robustly extracts the area where VFX is expressed by comparing the source and target image created with CG. In this paper, we focus on fire expression among various types of VFX and propose a ROI masking method to extract the flame area from VFX synthetic images.

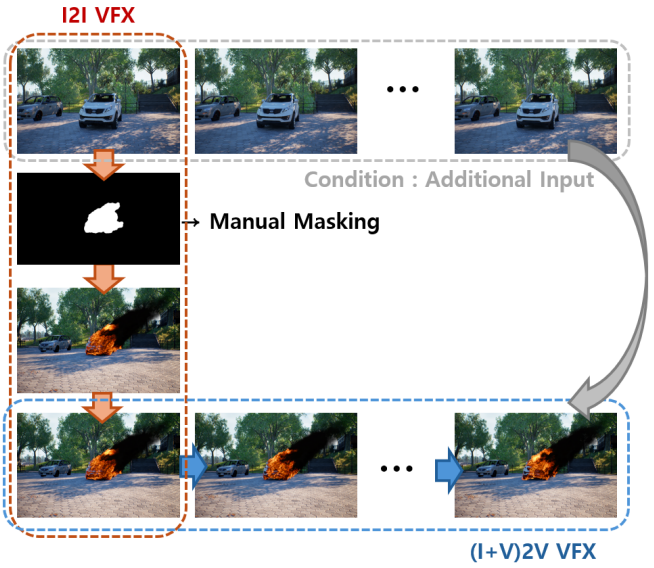

Fig. 1. Two-step VFX production pipeline composed of I2I VFX(Image-to-Image VFX) and (I+V)2V VFX((Image+Video)-to-Video VFX)

For example, the goal is to extract only the flame area by comparing the source and target image of the learning data represented in Fig. 2 (a) and (b), and create a masking area like (d) in a shape as if drawn by the user in a scribble manner. Algorithm 1 shows how to get this ROI mask. All parameters are explained based on images with a resolution of 1024p. Fn(.) is a function that assigns 1 to pixels of the flame color and 0 to pixels of other colors. The modeling of flame color was experimentally estimated and used with fixed parameters. As described in Algorithm 1, it was experimentally confirmed that the method is robust to some extent to errors in flame color estimation because it involves a process of removing noise and refining the masking area. 'Masking expansion' is the process of creating extra area so that the masking area sufficiently covers the flame area. The purpose of the opening and closing morphology operations is to create a masking area that resembles a human-drawn scribble mask.

C. VFX Blending

In the (I+V)2V VFX stage above, although the original input video was provided as a condition, unnecessary flickering still occurs in the background area as well as the VFX area when looking at the generated video. If VFX-generated footage is to be used in commercial content, these background artifacts must be removed. We note that since VFX such as fire and smoke are expressed sporadically, there is no need to precisely extract and composite only those areas. Extracting the area to be composited wider than the actual VFX area and applying blurring to the alpha map so that the border between the composited areas is not visible is a useful method that can easily solve the current task. Algorithm 2 describes this VFX Blending strategy. First, a difference image map between the generated image and the original image is obtained,

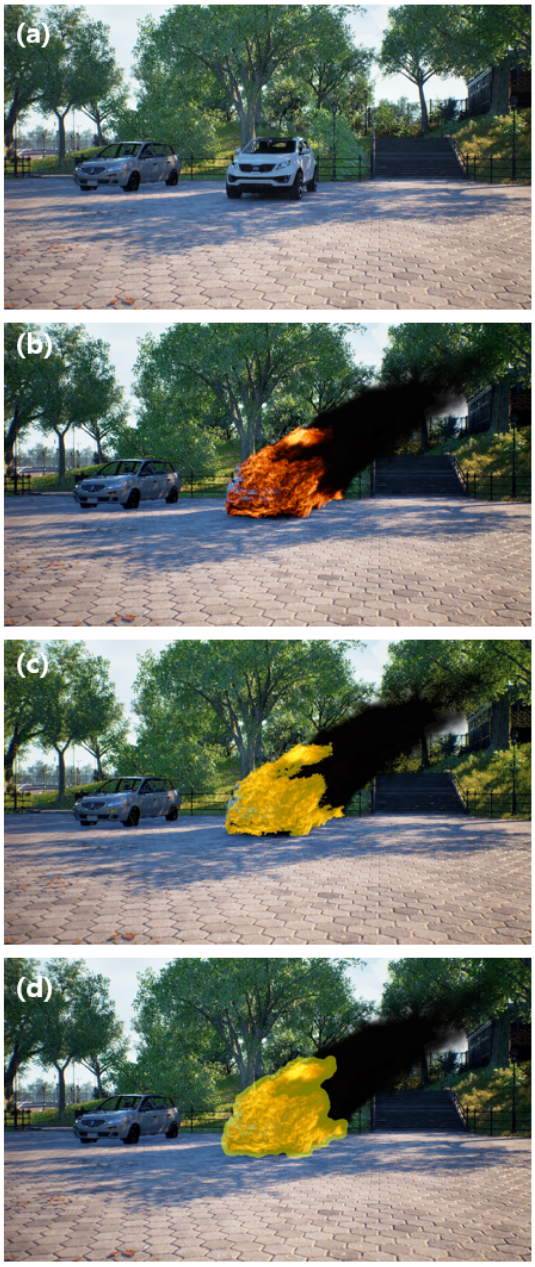

Fig. 2. Example of automatic extraction of ROI masking in the flame area: (a) A source image (CG generated), (b) A target image (CG generated), (c) Results of extracting pixels of flame color, (d) The final extraction result of ROI masking in a shape that looks like a person drawing.

and small noise elements are removed through morphology operations. After sufficiently expanding this difference image map, Gaussian blurring is applied to it to obtain a blending map, and finally the generated image and the original image are synthesized. Fig. 3 shows this process.

III. Experimental Results

As explained above, in implementing VFX of the two-step approach, we implemented I2I VFX by applying the method of [1] in the manual masking region, and implemented (I+V)2V

Algorithm 1: ROI Masking for Fire-flame

Input: Rs, Gs, Bs, Rt, Gt, Bt
Output: $Mask$

1 **for** ***each i*** **do**
2 $Mt(i) \leftarrow Mean(Rt(i), Gt(i), Bt(i))$;
3 $Ms(i) \leftarrow Mean(Rs(i), Gs(i), Bs(i))$;
/* Extract pixels of flame color */
4 $Mask(i) \leftarrow$
$Fn(Rs(i), Rt(i), Gs(i), Gt(i), Ms(i), Mt(i))$;
/* Noise deletion */
5 $Mask \leftarrow Morphology(Mask, erosion, 5)$;
/* Masking expansion */
6 $Mask \leftarrow Morphology(Mask, expansion, 21)$;
/* Opening operation */
7 $Mask \leftarrow Morphology(Mask, erosion, 21)$;
8 $Mask \leftarrow Morphology(Mask, expansion, 21)$;
/* Closing operation */
9 $Mask \leftarrow Morphology(Mask, expansion, 21)$;
10 $Mask \leftarrow Morphology(Mask, erosion, 21)$;
11 **return** $Mask$;

Algorithm 2: VFX Blending

Input: $Result, Source$
Output: $Composite$

1 $Diff \leftarrow Abs(Result - Source)$;
2 $DiffMap \leftarrow Diff > th$;
/* Operation for noise deletion */
3 $DiffMap \leftarrow Morphology(DiffMap, erosion, 5)$;
4 $DiffMap \leftarrow Morphology(DiffMap, dilation, 5)$;
/* Expansion of Diff Map */
5 $DiffMap \leftarrow Morphology(DiffMap, dilation, 21)$;
6 **for** $k = 0$ **to** 10 **do**
/* Blurring to remove border effect */
7 $Blending \leftarrow$
$GaussianSmoothing(DiffMap, 5)$;
8 $Composite \leftarrow$
$Blending \times Result + (1 - Blending) \times Source$;
9 **return** $Composite$;

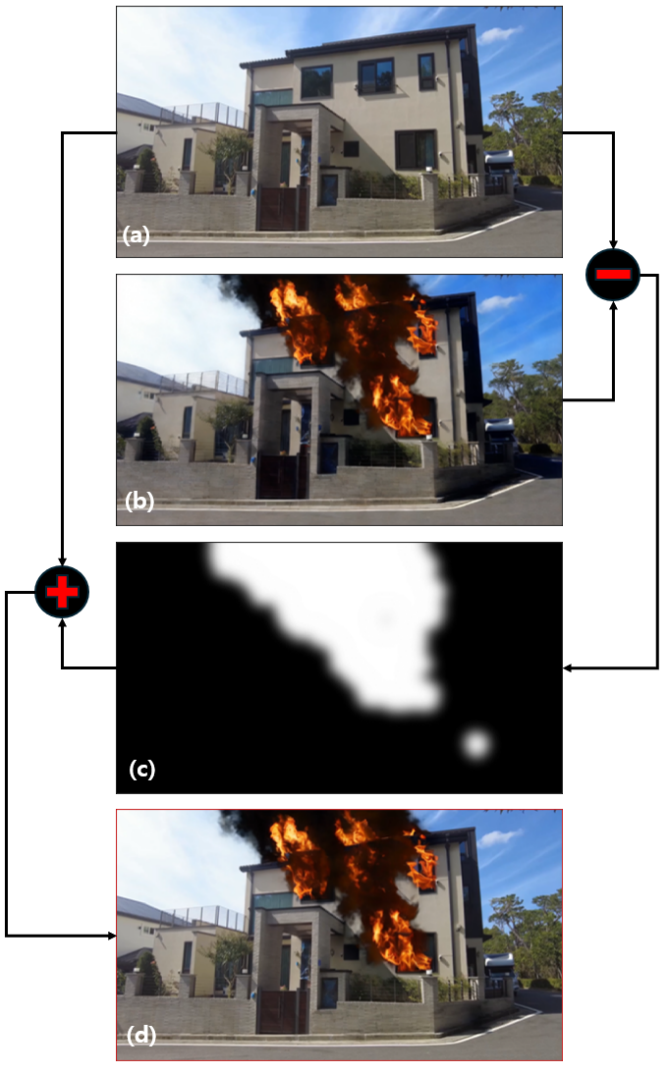

Fig. 3. Example of VFX Blending: (a) a source image (a real image), (b) a (Image+Video)-to-Video VFX Result, (c) Blending Map (line 7 in Algorithm 2), (d) final blending result (line 8 in Algorithm 2.)

VFX by using the video generation model of [2]. Among various VFX, we only implemented the expression of flames, and built learning data that could generate fire in three types of background spaces: cars, buildings and forests. We constructed these three background spaces with CG models and rendered the fire with Ambergen [3], and generated 600 video clips for each kind of background. Each clip is composed of 30fps and 20sec, and a total of 1800 clip videos and 1,080,000 frames of target and source images were made. The training data was constructed at 2K resolution, but was downscaled to 1024p resolution for actual training and inference. The purpose of this paper is to produce high-quality VFX synthetic images that minimize artifacts in the background area, so computation time is not considered.

Fig. 4 shows the result of (I+V)2V VFX. The left column is the input video and the right column is the VFX-generated video. The car and building videos are real-life test videos, and the forest video is a CG-created test video. Since it is a picture in a thesis, it may be difficult to distinguish flickering, but you can see that there is a slight difference in color in the background area. The color difference in the background area that occurs during the VFX creation process is also resolved in the final blending process described above. Fig. 5 compares the images before and after blending by overlaying them on the original input image. As you can see by comparing the area within the red dotted box, after the blending process, the colors that were distorted in the background area are now consistent with the original image.

Fig. 4. Example of (Image+Video)-to-Video VFX Result: Results of creating fire VFX in car, building, and forest background footage (car and building are real footage, forest is CG footage)

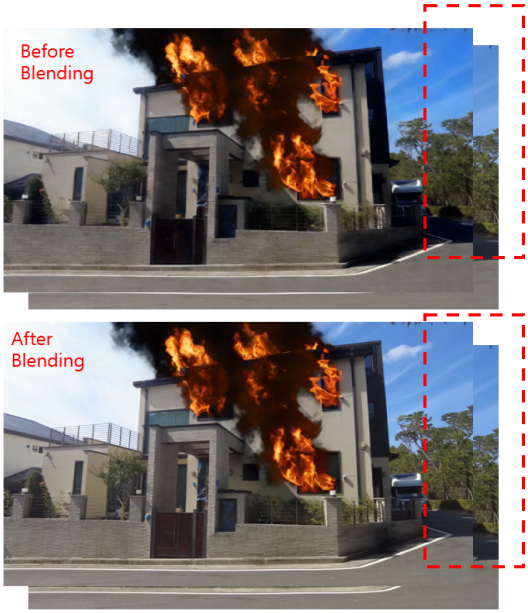

Fig. 5. Comparison of background color before and after final VFX blending: With the input original image laid on the floor, compare the background color before (top) and after (bottom) blending.

IV. Conclusion

In this paper, we addressed a two-step VFX production pipeline for creating VFX based on generative AI, and proposed two image processing technologies essential for its practical implementation. Although the development technology is not yet fully developed, it is believed to have made a significant contribution to the problem of preserving background areas and maintaining consistency when utilizing generative AI technology in VFX work. We expect that its usability will increase further if we address the issue of computational speed, which has not yet been taken into account, and expand the applicable image range.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00395401, Development of VFX creation and combination using generative AI)

References

[1] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

[2] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[3] JangaFX, "Embergen." https://jangafx.com/software/embergen/download, 2025. accessed: 2025-08-18.