2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) | 979-8-3503-6897-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICITISEE63424.2024.10730211

Reconstructing Missing Frame In Video Transmission With Frame Interpolation

1st Muhamad Wisnu Pangestu
Magister of Informatics Engineering
Universitas AMIKOM Yogyakarta
Yogyakarta
wisnup2902@students.amikom.ac.id

2nd Arief Setyanto
Magister of Informatics Engineering
Universitas AMIKOM Yogyakarta
Yogyakarta
arief_s@amikom.ac.id

***Abstract*— This study addresses the issue of missing frame in video transmission, a critical factor affecting Quality of Service (QoS) in video communications. Using spatial interpolation techniques, the research generates replacement frames by analyzing pixels and patterns from surrounding frames, evaluating their accuracy in enhancing visual quality, playback smoothness, and user satisfaction. The FILM Model simulates frame loss scenarios, testing an algorithm designed to recover missing frames and assessing its effectiveness through the Structural Similarity Index (SSIM). Six different videos with four types of transitions were examined, spanning various video types. T-Test analysis across diverse video scenarios indicates that the model performs exceptionally well in Low Lighting and HDR conditions, maintaining high SSIM values and visual quality even at greater frame distances. However, challenges arise with AI-generated and camera phone footage, where SSIM values significantly decline during complex transitions like Cross Dissolve. These differences are statistically significant, highlighting the model's need for improvement in handling diverse video types. The findings underscore the importance of refining the model to better manage dynamic scenarios and complex transitions, enhancing its applicability in broader video production contexts and potentially improving real-time video communication systems.**

Keywords*— *Framedrop, Frame Interpolations, Video Restoration, Generated Frame, FILM Model , SSIM.

I. Introduction

The rapid advancement of video communication technologies has led to a surge in the demand for seamless and high-quality video transmission over networks. However, the transmission of video content over these networks is inherently susceptible to interruptions, which can result in the loss of several frames during the delivery process. This frame loss poses a significant challenge, particularly in real-time video communications such as live broadcasts and two-way video conferencing, where the Quality of Service (QoS) is critically important[1]. The loss of frames can cause the received video to appear fragmented or disjointed, disrupting the viewing experience and diminishing user satisfaction. Viewers often experience discomfort when exposed to video content that is not fluid or continuous, which negatively impacts their overall experience.

To mitigate the effects of frame loss, frame reconstruction techniques have been developed with the aim of filling in the missing frames by generating plausible replacements. Among the various approaches, spatial interpolation has emerged as a widely used method. Spatial interpolation leverages the spatial information from the frames surrounding the missing frame to predict and recreate its content. By analyzing the pixels and patterns in adjacent frames, this method can generate a new frame that fits seamlessly into the video sequence, thereby maintaining the visual continuity of the video[2]. However, while spatial interpolation is effective, it is not without its challenges. One of the primary issues in video communication is the speed at which frame reconstruction must be performed. The process needs to be rapid enough that the viewer does not perceive any interruptions in the video stream. This necessitates that the reconstruction of missing frames be accomplished with both high quality and speed, ideally within a timeframe shorter than the duration of a single frame, which is typically around 30 frames per second.

Moreover, as video resolutions and frame rates continue to increase, the demand for more sophisticated and efficient frame reconstruction techniques also grows. Traditional methods may struggle to keep pace with these advancements, leading to a degradation in performance, especially in scenarios involving high-resolution video or complex motion patterns. As such, there is an ongoing need for innovative approaches that can address these limitations and deliver superior performance in real-time video applications. In this context, the development of advanced frame interpolation methods, particularly those that can handle large motions and complex scenarios, becomes crucial. The focus of this research is on enhancing the performance of frame interpolation techniques to better manage the challenges associated with framedrop in video conferencing, with a specific emphasis on utilizing the FILM model for this purpose[3].

II. Related Works

Building on previous research, the 'Time Lens' method for event-based video frame interpolation marks a significant step forward by merging the advantages of event-based and frame-based technologies, particularly in dynamic scenarios. This approach boosts PSNR by up to 5.21 dB over existing methods, showcasing its effectiveness. A newly released dataset further supports research in this area. However, a notable 'domain gap' between synthetic and real data suggests the need for better calibration to improve performance. Additionally, the method's inconsistent results when more frames are skipped indicate areas for enhancing result consistency. Validation through the Structural Similarity Index (SSI) underscores differences in models and validation techniques used.[4][5].

The 'Event-Based Frame Interpolation with Ad-hoc Deblurring' method utilizes a bidirectional recurrent network to address both sharp and blurred video deblurring. It

introduces the HighREV dataset for testing and demonstrates superior performance in frame interpolation and deblurring compared to existing methods on the GoPro and HighREV datasets. However, the REFID method struggles with generalization, relies on high-quality datasets, and is computationally intensive. In contrast, this research focuses on the FILM Model, which offers a practical solution for frame drops in video conferences. The FILM Model, featuring a unified network with a feature extractor, provides promising results but faces limitations such as dependence on limited training data, challenges with extreme scenarios, and insufficient comparative data. The study measures FILM's performance using normal videos with intentionally omitted frames, comparing it against other methods to evaluate its effectiveness[3].

This research introduces a two-stage deep learning system designed to reconstruct high-resolution slow-motion video from two input sources: one high-frame-rate but low-resolution, and another low-frame-rate but high-resolution. The system uses alignment and appearance estimation techniques to enhance video quality and accuracy, effectively combining the two inputs for frame interpolation. While the system shows impressive results with synthetic videos and practical dual-camera setups, it faces limitations such as poor performance with extremely low-resolution frames, issues with significant misalignment causing motion boundary artifacts, and challenges in estimating flow from saturated videos, which introduces minor artifacts. Additionally, the system cannot produce videos with a higher frame rate than the additional video. The research aims to improve visual quality in slow-motion video production and focuses on rapid frame recovery for real-time video conferencing scenarios[6]. This research develops a new video frame interpolation method using generalized deformable convolution mechanisms to address the limitations of traditional flow-based and kernel-based methods. The approach improves motion learning and sampling flexibility in space-time, outperforming existing techniques in handling complex motion. Despite its advantages, the method relies solely on data-based training for motion estimation, which can lead to blurred outputs when motion patterns in the evaluation dataset differ from the training data, especially with atypical movements. While this study focuses on generalized deformable convolution, it contrasts with research targeting frame interpolation for missing frames in video conferencing, using the FILM Model for more specific applications[7].

III. Methodology

This research addresses the challenges of video frame interpolation, particularly in reconstructing missing frames during transmission interruptions[7]. It evaluates the effectiveness of frame interpolation in enhancing playback smoothness, visual quality, and user satisfaction. The study is limited to using a single frame before and after the drop, with a resolution of 1280 x 720, and employs Google Colaboratory for processing resources.

A. FILM Model

This research focuses on quantitative measurement, involving the design and implementation of an interpolation algorithm using the FILM Model[3]. The effectiveness of the algorithm is assessed through the Structural Similarity Index (SSIM), which evaluates video quality based on luminance, contrast, and structure. SSIM is a newer measurement tool that is designed based on three factors i.e. luminance, contrast, and structure to better suit the workings of the human visual system[5]. Which will assess the generated images by the FILM Model against the actual frames that were intentionally removed.

The steps taken in this research involve extracting each frame from a video experiencing missing frame, then identifying which frames contain the frame transition gap. Once the pre and post frame drop pairs are obtained, the FILM model can implement the missing frames to visualize the missing frames.

B. Equation

FILM Model can complete its task, requiring 2 frames (f_1, f_2). Which will be used as a reference to create a generated frame $\hat{f}_t$, with $t = 0.5$ and the symbol for the model is $\mathcal{M}$ [3]. Which is formulated as follows:

$$\hat{f}_t = \mathcal{M}(f_1, f_2) \tag{1}$$

By Equation (1), demonstrates how the FILM model can identify and create the transition frame, effectively filling in the visual gap caused by the frame drop. Based on Equation (1), the FILM model can find the transition frame with an increase in frame count exactly at the length of 0.5 frames. If this research seeks the frame drop between frames 1 and 2, the FILM model can create frame 1.5.

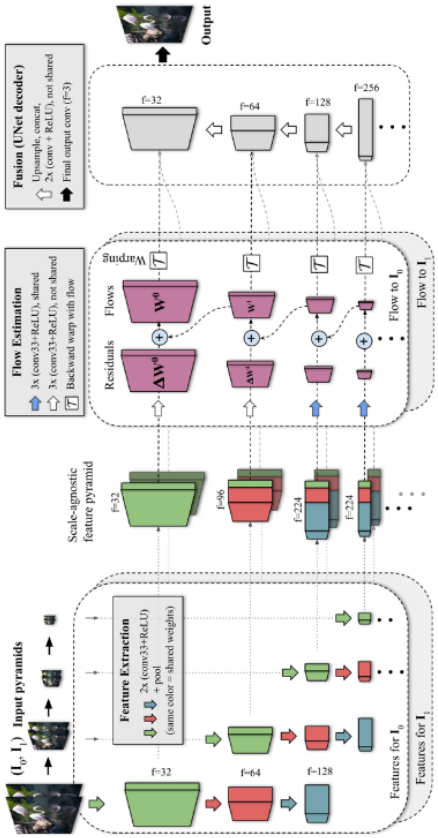

Figure 1: FILM Model Architecture

Based on Figure 1, It is explained that the architectural scheme used can generate two-way current calculations that are "scale agnostic" by utilizing a feature pyramid extracted

through joint weighting[3]. Once the generated in-between image is obtained, the visual quality of the generated result can be assessed to determine if it looks sufficiently good or not.

The SSIM method assesses image similarity by comparing structural characteristics, aligning closely with the Human Visual System's (HVS) quality perception. SSIM requires both an experimental and a reference image, with higher similarity indicating better image quality. It effectively distinguishes structural details from lighting and reflections, evaluating image quality based on structural similarity rather than brightness alone[8].

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \quad (2)$$

Where the two photos that are being compared are x and y The average of pictures x and y is represented by the parameters μ_x and μ_y respectively, while the variances of the two images are σ_x^2 and σ_y^2. σ_x^2 represents the covariance between pictures x and y. As the denomination value gets close to zero, the division is stabilized using two constants, C_1 and C_2.

A number between -1 and 1, which denotes structural equivalency between the two pictures, is determined via SSIM. A higher SSIM score denotes a better degree of structural, lighting, and contrast similarity between the two pictures. To evaluate the quality of compressed or processed images, this measure is highly helpful in the fields of image processing and compression[9].

C. Experiment Setup

This study examines various video conditions, including scene transitions, low-light[10], HDR[11], animated sequences[12], (QTE) sequences[13], and AI-video[14]. Due to upload limitations on Google Colab, videos are reduced to 720p resolution. The research uses essential Python libraries (Keras, TensorFlow[15][16], NumPy[16], Pandas, Matplotlib[17]) along with additional tools (OpenCV[18], mediapy[19], FFmpeg, imageio[20], scikit-image[21]) for processing and evaluating video frames[17]. The process involves parsing videos into frames, identifying missing frames, and using the FILM Model to generate interpolated frames Figure 2 outlines the procedure for this experiment.

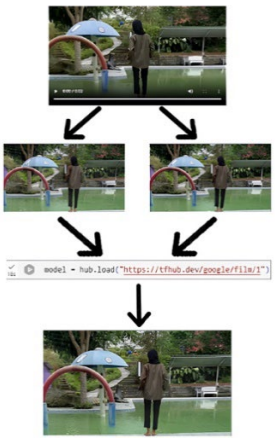

Figure 2: Outline Process

Based on Figure 2, the test involves parsing a video into individual frames. The missing frame is identified when the visual difference between consecutive frames, such as from frame a to frame a+1, is significant. Once identified, frames a and a+1 are used as reference points in the FILM Mode [3]. These reference frames are loaded into the model, which then generates the missing target frame as the output.

D. Validation Procedures

The CLi will provide output in the form of visual differences and SSIM values. [5]. Using structural data from an image with a parameter range of 0 to 1, with a middle value of 0.5 to 0.8, SSIM may be used to analyze picture differences and identify abnormalities when comparing two photos.[8], [22].

The validation process using SSIM can be started by inputting CLI Figure 3.

```
from skimage.metrics import structural_similarity as ssim

# Menghitung SSIM antara dua gambar
ssim_value, ssim_image = ssim(image1, image2, full=True, multichannel=True)

# Menampilkan hasil SSIM
print(f'SSIM antara Frame Aktual dan Frame Generated: {ssim_value}')

# Menampilkan gambar SSIM
plt.imshow(ssim_image, cmap='gray')
plt.title(f'SSIM Image (SSIM: {ssim_value:.4f})')
plt.axis('off')
plt.show()
```

Figure 3: Image Vadilation using SSIM

The CLi will provide output on Figure 4 in the form of visual differences and SSIM values.

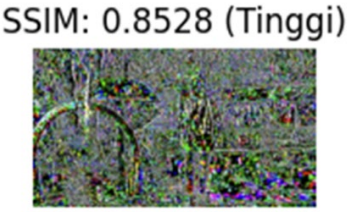

Figure 4: SSIM Value & Visual Comparison with Actual Frame

The result produced indicates that the SSIM value is 0.8528. It indicates that the final picture produced by the Film Model bears a strong resemblance to the Actual Frame, which was purposefully deleted at the outset of the study.

With additional frames that have been deleted from the video used for this study, the interpolation and validation processes must be redone in a similar manner.

IV. Result & Analysis

The study findings are reported in this chapter. First, it discusses the accuracy of the FILM Model output. Second, provides the similarity accuracy in a table that will be closed in Third, addressing the presentation of the study findings, where the accuracy results obtained using SSIM seem high.

A. Model Accuracy

The accuracy of the FILM Model in generating frames that closely resemble the original frames was assessed using the SSIM. SSIM values, ranging from 0 to 1, indicate the degree of similarity between two images, with values closer to 1 representing higher similarity. Our evaluation conducted with various scenes including Camera Phone, AI Generated, Quick Time Event, Animation, Low Light Video & HDR Video with three variations of frame gap and transition variations such as Non-Transition, Cross Dissolve, Dip to Black & Dip to White. To validate and minimize the factor (coincidence) each test was repeated ten times in different scenes. The total testing conducted was 720 times for each Scene, Transition, and Frame Gap.

Table 1. Avg. SSIM Value

Scene & Transition	Frame Gap 1	Frame Gap 3	Frame Gap 5
CPNT	0.8476	0.7645	0.6777
CPCD	0.7670	0.5871	0.5295
CPDTB	0.8297	0.6851	0.6439
CPDTW	0.8267	0.7030	0.6246
AINT	0.7930	0.7253	0.6877
AICD	0.7930	0.7253	0.6877
AIDTB	0.8363	0.7932	0.7306
AIDTW	0.8728	0.8309	0.7706
QTENT	0.6646	0.6181	0.5916
QTECD	0.7772	0.7407	0.7293
QTEDTB	0.8519	0.7884	0.7556
QTEDTW	0.8583	0.8193	0.7961
ANT	0.9180	0.8712	0.8575
ACD	0.8684	0.8114	0.7545
ADTB	0.8472	0.7862	0.7223
ADTW	0.8417	0.7943	0.7404
LLNT	0.9428	0.9410	0.9159
LLCD	0.9591	0.9506	0.9435
LLDTB	0.9396	0.9231	0.9118
LLDTW	0.9600	0.9270	0.8896
HDRNT	0.9102	0.9047	0.8555
HDRCD	0.9227	0.9028	0.8837
HDRDTB	0.9204	0.8615	0.8392
HDRDTW	0.9395	0.9029	0.8844

Table 1 summarizes SSIM values from 720 trials, SSIM values decrease as the frame gap widens, indicating a loss of structural similarity. However, transitions like LLCD and LLDTW maintain high similarity even with larger gaps, reflecting consistent visual quality. In contrast, transitions such as CPCD experience a more significant decline in similarity with larger frame gaps.

B. SSIM Comparison Values for All Validated Frames

This section compares SSIM values across validated frames, illustrating how visual similarity decreases with increasing frame distance. SSIM values decline as the distance between frames grows, indicating reduced structural similarity. However, transitions like LLNT and LLDTW maintain high SSIM values even with larger frame distances, suggesting consistent visual quality.

Conversely, transitions such as CPCD and AICD exhibit a more pronounced drop in similarity with greater frame distances, indicating higher sensitivity to visual changes and potential impacts on perceived video quality. Overall, LL and HDR scenarios show higher SSIM values compared to AI or CP videos, suggesting better preservation of visual details under these conditions. The bar charts in Figures 5, 6, and 7 illustrate the overall comparison of SSIM values as detailed in Table 1.

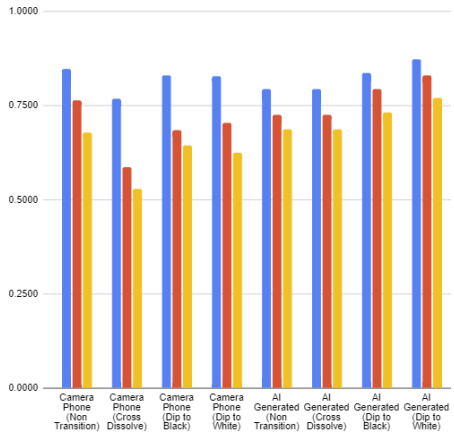

Figure 5: Average SSIM Datatest Result CP to AI

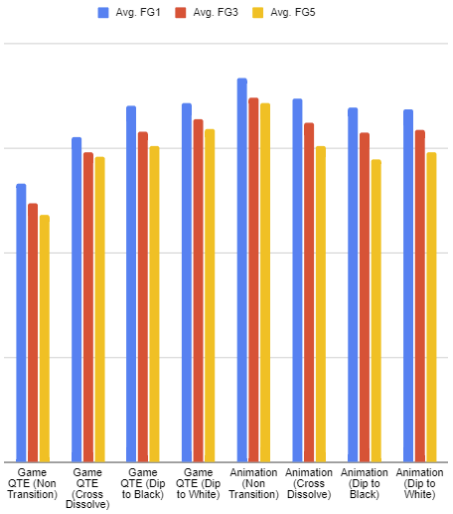

Figure 6: Average SSIM Datatest Result Game QTE & Animation

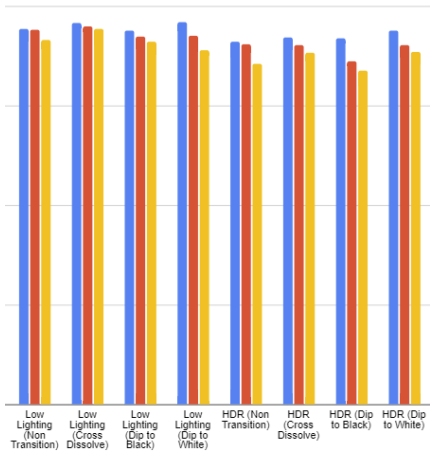

Figure 7: Average SSIM Datatest Result LL & HDR

To strengthen the SSIM analysis, a T-Test was performed to assess the significance of the differences in SSIM values between frames generated at different frame distances. The T-Test is a statistical test used to compare the means of two data groups and determine whether the observed differences between them are statistically significant.

- **Null Hypothesis (H_0)**: There is no significant difference between SSIM values at different frame distances.

- **Alternative Hypothesis (H_1)**: There is a significant difference between SSIM values at different frame distances.

The T-Test results reveal significant differences in SSIM values across varying frame distances in many scenarios, particularly with complex transitions. However, in LL and HDR, these differences are often non-significant, indicating greater resilience to frame distance variations in terms of visual similarity. Overall, increased frame distances generally lead to reduced visual similarity, with certain transitions showing more pronounced declines. These findings offer valuable insights for selecting effective video transition techniques to preserve visual quality across different scenarios.

C. Discussion

Evaluating SSIM values shows the FILM Model excels in generating high-fidelity frames, especially in Low Light and HDR videos. While it performs well, challenges with AI and camera phone footage suggest areas for improvement, such as refining datasets and model adjustments. The FILM Model's strengths and opportunities for enhancement highlight its potential in media production and video editing. Future research should focus on improving performance in dynamic conditions and diverse video genres to maximize the model's effectiveness across various applications.

D. Conclusion

The FILM Model excels in generating frames with high structural similarity to the originals, as shown by the SSIM values across different video conditions. The analysis indicates that visual similarity between frames diminishes with increased distance, though some transitions demonstrate greater stability. LLNT and LLDTW maintain high SSIM scores even at larger distances, reflecting effective visual quality preservation. In contrast, CPCD and AICD show significant SSIM declines at greater distances. HDR and LL generally achieve higher SSIM values compared to AI or CP videos, suggesting better detail preservation. T-Test results validate these differences as statistically significant, underscoring the influence of frame distance on visual similarity and offering insights for enhancing video transition techniques. Overall, while the FILM Model shows strong performance in frame generation, the findings highlight areas for further refinement, offering a solid basis for future improvements and broader application in video content.

E. Limitation

The primary limitation of this research is that the FILM model is still in its pre-trained phase, and the study's limited timeframe prevented further training. According to Reda's test results [3], the model generated 65 frames between two tested frames. However, due to time constraints, additional training to improve performance was not possible. Future research should focus on extended development and training to fully harness the model's potential.

ACKNOWLEDGMENT

The authors extend their gratitude to Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, and Caroline Pantofaru from Google Research, as well as Brian Curless from Google Research and the University of Washington, for their invaluable contributions to the FILM Model's development. Their expertise and collaborative efforts were crucial to this research.

REFERENCES

[1] G. Konuko, S. Lathuilière, and G. Valenzise, "A Hybrid Deep Animation Codec for Low-bitrate Video Conferencing," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.13530

[2] I. A. Laksmi Dewi, N. Indra, I. M. O. Widyantara, I. A. Laksmi, and N. Indra Er, "FRAME RATE MINIMUM VIDEO DENGAN METODE NORMALIZED FRAME DIFFERENCE SEBAGAI PENDESKRIPSI INTENSITAS GERAK," 2015.

[3] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru, and B. Curless, "FILM: Frame Interpolation for Large Motion," 2022. [Online]. Available: https://film-net.github.io.

[4] S. Tulyakov et al., "Time Lens: Event-based Video Frame Interpolation," 2021. [Online]. Available: http://rpg.ifi.

[5] D. R. I. M. Setiadi, "PSNR vs SSIM: imperceptibility quality assessment for image steganography," Multimed Tools Appl, vol. 80, no. 6, pp. 8423–8444, Mar. 2021, doi: 10.1007/s11042-020-10035-z.

[6] A. Paliwal and N. K. Kalantari, "Deep Slow Motion Video Reconstruction with Hybrid Imaging System," IEEE Trans Pattern Anal Mach Intell, vol. 42, no. 7, pp. 1557–1569, Jul. 2020, doi: 10.1109/TPAMI.2020.2987316.

[7] Z. Shi, X. Liu, K. Shi, L. Dai, and J. Chen, "Video Frame Interpolation via Generalized Deformable Convolution," IEEE Trans Multimedia, vol. 24, pp. 426–439, 2022, doi: 10.1109/TMM.2021.3052419.

[8] H. B. Sumarna, E. Utami, and A. D. Hartanto, "Tinjauan Literatur Sistematik tentang Structural Similarity Index Measure untuk Deteksi Anomali Gambar Systematic Literature Review of Structural Similarity Index Measure for Image Anomaly Detection," Citec Journal, vol. 7, no. 2, 2020.

[9] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in Proceedings - International Conference on Pattern Recognition, 2010, pp. 2366–2369. doi: 10.1109/ICPR.2010.579.

[10] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding, and K.-K. Ma, "Learning a Simple Low-Light Image Enhancer from Paired Low-Light Instances," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2023, pp. 22252–22261. doi: 10.1109/CVPR52729.2023.02131.

[11] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," in ACM Transactions on Graphics, Association for Computing Machinery, Nov. 2017. doi: 10.1145/3130800.3130816.

[12] R. Fernanda Lengkong and A. Budiman, "TEKNIK DIGITAL COMPOSITING DALAM ANIMASI 2D 'MENJAGA RINJANI' DIGITAL COMPOSITING

TECHNIQUE IN 2D ANIMATION 'MENJAGA RINJANI.'"

[13] "Quick time event," https://en.wikipedia.org/wiki/Quick_time_event. Accessed: Jul. 10, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Quick_time_event

[14] X. Liu et al., "NTIRE 2024 Quality Assessment of AI-Generated Content Challenge," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.16687

[15] N. Narayanan, Z. Chen, B. Fang, G. Li, K. Pattabiraman, and N. DeBardeleben, "Fault Injection for TensorFlow Applications," IEEE Trans Dependable Secure Comput, vol. 20, no. 4, pp. 2677–2695, Jul. 2023, doi: 10.1109/TDSC.2022.3175930.

[16] B. Antunes and D. R. C. Hill, "REPRODUCIBILITY, ENERGY EFFICIENCY AND PERFORMANCE OF PSEUDORANDOM NUMBER GENERATORS IN MACHINE LEARNING: A COMPARATIVE STUDY OF PYTHON, NUMPY, TENSORFLOW, AND PYTORCH IMPLEMENTATIONS."

[17] -- Anhui, Q. Meng, and R. Sun, "Frontiers in Computing and Intelligent Systems Visual Analysis of Campus Card Consumption Data based on Matplotlib", Accessed: Jul. 03, 2024. [Online]. Available: https://doi.org/10.54097/fcis.v4i2.10367

[18] T. Cut Al-Saidina Zulkhaidi, E. Maria, P. Studi Teknologi Rekayasa Perangkat Lunak, and P. Pertanian Negeri Samarinda, "Pengenalan Pola Bentuk Wajah dengan OpenCV," JURTI, vol. 3, no. 2, 2019.

[19] The mediapy Authors, "https://google.github.io/mediapy/mediapy.html."

[20] D. R. Stirling, A. E. Carpenter, and B. A. Cimini, "CellProfiler Analyst 3.0: Accessible data exploration and machine learning for image analysis." [Online]. Available: https://cellprofileranalyst.org/examples

[21] K. Navya Sri, K. Neha, K. P. P. Sudheshna, M. R. S. L. K. Gayathri, and B. M, "Robust Parking Space Allocation System Using Open CV and Scikit-learn," Journal of Image Processing and Intelligent Remote Sensing, no. 43, pp. 37–46, Apr. 2024, doi: 10.55529/jipirs.43.37.46.

[22] H. Zhao, B. Liu, and L. Wang, "Blur kernel estimation and non-blind super-resolution for power equipment infrared images by compressed sensing and adaptive regularization," Sensors, vol. 21, no. 14, Jul. 2021, doi: 10.3390/s21144820.