Journal of Artificial Intelligence and Legal Technology
SAH PUBLISHER
e-ISSN 3123-786X
Vol. 2, No. 2, August 2026

Evaluating Film Model For Controlled Missing-Frame Reconstruction In Low-Light Smartphone Video

Muhamad Wisnu Pangestu[1]*, Fauzia Dhiyaa' Farros[2]
[1,2] Computer Science, Sugeng Hartono University, Sukoharjo, Indonesia

Abstract. This study evaluates Frame Interpolation for Large Motion (FILM) for the controlled reconstruction of intentionally withheld frames from a researcher-owned low-light video recorded with a smartphone camera. The source video was standardized to 1280 x 720 pixels and prepared in four editorial transition conditions: no transition, cross dissolve, dip to black, and dip to white. Ten target frames were selected from each condition and evaluated at frame gaps of 1, 3, and 5, producing 120 primary Structural Similarity Index Measure (SSIM) observations. For each observation, FILM synthesized an intermediate frame from two retained boundary frames; the synthesized image was compared with the withheld original frame. The aggregate mean SSIM was 0.9504 at frame gap 1, 0.9354 at frame gap 3, and 0.9152 at frame gap 5. Across the evaluated conditions, the mean SSIM decreased from 0.9504 at FG1 to 0.9152 at FG5. The 0.0352-point decrease suggests that structural preservation becomes less reliable when the retained boundary frames are farther from the withheld target frame. These descriptive findings apply only to the single researcher-owned low-light smartphone video evaluated in this study and should not be generalized to other low-light videos, devices, scenes, or transition designs. The study does not evaluate network packet loss, end-to-end video quality of service, or real-time deployment performance.

Keywords: Video Frame Interpolation, Missing-Frame Reconstruction, FILM Model, Low-Light Video, Smartphone Video, Structural Similarity Index Measure, SSIM, Optical Flow
Received Jul 2026 / **Revised** Aug 2026 / **Accepted** Aug 2026

This work is licensed under a Creative Commons Attribution 4.0 International License.

INTRODUCTION

Low-light video sequences recorded using smartphone cameras frequently exhibit a range of visual degradations that reduce the reliability of the captured information. These degradations commonly include limited illumination, unstable dark-region details, sensor noise, luminance fluctuations, reduced contrast, and motion-related artifacts. In practical recording conditions, especially when illumination is insufficient, the camera sensor may fail to preserve fine structural information consistently across consecutive frames. As a consequence, the visual continuity of the video can be disrupted when one or more frames are unavailable, corrupted, or removed from the sequence. Such discontinuities may negatively affect both human perception and subsequent computational processing tasks that rely on temporal coherence.

Video frame interpolation (VFI) is a computational technique that generates intermediate frames from available neighboring frames in order to improve motion smoothness or restore missing visual information within a video sequence[1]-[4]. The primary objective of VFI is to estimate the appearance of a frame that would plausibly exist between two known boundary frames. However, an interpolated frame should not be interpreted as an exact replacement for the original missing frame. Instead, the generated frame represents an estimation produced by the model based on the spatial and temporal information contained in the available frames. Consequently, the accuracy of the interpolation process depends on the quality, consistency, and completeness of the input information.

Several factors can make the interpolation task inherently ambiguous. Large object motion, occlusion, disocclusion, complex texture patterns, and abrupt illumination changes may reduce the model's ability to

[1]* Muhamad Wisnu Pangestu.
Email addresses: wisnup2902@gmail.com (Pangestu)

[2]*Corresponding author.
Email addresses: fauziadhiyaaf@gmail.com (Farros)

infer the correct intermediate representation[1],[2],[5]. These challenges become particularly significant in low-light video because the amount of reliable visual information available to the model is generally more limited than in well-lit or controlled illumination conditions. Noise may obscure fine details, dark regions may contain insufficient texture information, and luminance variations between frames may complicate motion estimation. As a result, low-light video provides a relevant and demanding context for evaluating the performance of VFI methods.

In this study, FILM was selected as the interpolation model because it was specifically designed to address large-motion interpolation through the use of a multi-scale feature extractor and a scale-agnostic motion estimator[1]. The architecture of FILM enables the model to capture motion information at multiple spatial resolutions, which is advantageous when objects move significantly between frames. The research scope was intentionally restricted to a single low-light video asset recorded using a smartphone camera. This simplification was adopted to provide a clearer definition of the research object, the data source, and the boundaries of generalization. By focusing on one representative video sequence, the analysis can concentrate on the interpolation behavior of the model under a specific low-light condition without introducing additional variability from multiple datasets.

RELATED WORKS

Deep learning-based VFI approaches can generally be categorized into several methodological groups, including optical flow-based methods, adaptive kernel-based methods, and approaches that employ more complex feature representations. Optical flow-based methods estimate pixel-wise motion between frames and use this information to synthesize intermediate frames. DAIN, for example, incorporates depth information in order to improve interpolation performance in the presence of occlusion[2]. RIFE estimates intermediate flow in an end-to-end manner and supports arbitrary interpolation time steps, making it suitable for flexible frame generation scenarios[3]. Super SloMo utilizes bidirectional optical flow to generate multiple intermediate frames between two input frames[6].

Kernel-based methods adopt a different strategy by learning local convolutional kernels for output frame synthesis. SepConv and AdaCoF are representative examples of this approach, as they learn adaptive local kernels that are applied to neighboring pixels in order to reconstruct the interpolated frame[7],[8]. In contrast, CAIN employs channel attention mechanisms to emphasize informative feature channels, while Transformer-based VFI methods attempt to capture long-range pixel correlations and global contextual relationships across frames[4],[9]. These methodological differences indicate that interpolation performance may depend strongly on the motion representation strategy, feature extraction mechanism, and appearance modeling approach used by the model.

Previous VFI research has also demonstrated that single-frame evaluation metrics do not fully capture temporal interpolation quality. An interpolated frame may appear visually plausible when evaluated independently, yet still introduce temporal inconsistencies when viewed as part of a continuous video sequence. For this reason, perceptual metrics specifically designed for VFI have been developed to account for interpolation artifacts and spatiotemporal information[10]. These metrics aim to evaluate not only spatial similarity but also the temporal coherence of generated frames.

Nevertheless, Structural Similarity Index Measure (SSIM) remains an appropriate evaluation metric for the present study because each removed target frame is available as ground truth and can be directly compared with the corresponding output generated by FILM[11]. Under this experimental setup, SSIM provides a practical measure of structural similarity between the interpolated frame and the original frame. Therefore, SSIM is suitable for assessing how closely the generated frame preserves luminance, contrast, and structural information relative to the available ground-truth reference."

METHODS

This study adopts a controlled quantitative experimental design to evaluate the capability of the Frame Interpolation for Large Motion (FILM) model in reconstructing intentionally removed frames from a low-light video recorded with a smartphone camera. The methodology is organized as an end-to-end reconstruction pipeline consisting of source-video preparation, frame extraction, controlled target-frame removal, selection of two boundary frames, FILM-based intermediate-frame synthesis, and objective quality evaluation. The experimental design evaluates four transition conditions - no transition, cross dissolve, dip to black, and dip to white - under three temporal separation settings: frame gap 1, frame gap 3, and frame gap 5. For each test case, the original target frame is withheld from the model input and retained

as the ground-truth reference. FILM uses the two retained boundary frames to estimate the missing intermediate frame through multi-scale motion-aware interpolation[1]. The reconstructed frame is subsequently compared with the withheld original frame using the Structural Similarity Index Measure (SSIM), which assesses luminance, contrast, and structural similarity[11]. This methodological structure enables the study to examine how reconstruction quality changes across transition conditions and increasing temporal distances while maintaining a consistent source-video resolution, model configuration, and evaluation procedure.

Model Architecture

The research object is the FILM-generated reconstruction of intentionally removed frames from a single researcher-owned low-light video recorded with a smartphone camera. The working resolution is fixed at 1280 x 720 pixels to keep frame extraction, model inference, and quality evaluation consistent.

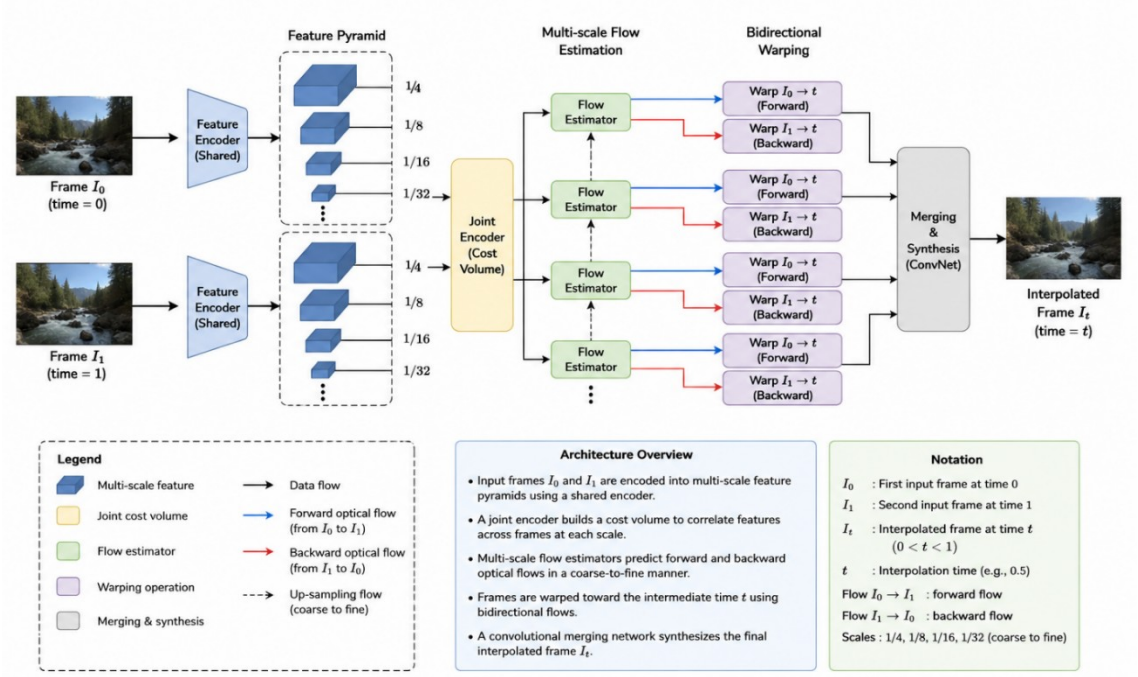


Figure 1. Film Architecture

Figure 1 presents the conceptual architecture of the FILM model used to reconstruct a missing frame from two retained boundary frames. Let denote the first input frame at time 0 and denote the second input frame at time 1. The objective of the model is to synthesize an intermediate frame at interpolation time , where . In this study, the target frame is positioned at the temporal midpoint; therefore, . The overall interpolation process can be expressed as:

(1)

Where denotes the pre-trained FILM model, and are the retained input frames, and is the reconstructed intermediate frame. For midpoint reconstruction, the formulation becomes:

(2)

The frame gap used in the experiment determines the temporal distance between the two boundary frames, whereas specifies the relative position of the target frame between those boundaries. Thus, FG1, FG3, and FG5 represent different reconstruction difficulties while retaining the same midpoint interpolation objective.

FILM was used as a pre-trained video frame interpolation model to estimate a withheld target frame from two retained boundary frames. The model uses multi-scale feature processing and bidirectional motion-aware synthesis; its complete architecture and training procedure are described by Reda[1]. In the present

study, FILM was treated as a fixed evaluation instrument rather than retrained or modified. The experimental focus was therefore the effect of transition condition and temporal separation on reconstruction quality. This stage integrates complementary information from both directions and reconstructs regions that may be unreliable after warping, such as areas affected by occlusion, motion ambiguity, or incomplete visual correspondence. The final output is subsequently compared with the withheld ground-truth frame using SSIM[11].
The same source video was edited into four transition conditions. The no-transition condition preserves continuity within one shot. Cross dissolve gradually mixes two visual states. Dip to black and dip to white introduce transitions dominated by dark and bright values, respectively. These conditions were selected because the visual evidence surrounding a target frame may differ across transitions, affecting how well an intermediate image can be inferred from two boundary frames.

Experimental Design
Ten target frames were evaluated in each transition condition. Every target frame was retained as a ground-truth reference but excluded from model input. Three temporal separations were evaluated: frame gap 1 (FG1), frame gap 3 (FG3), and frame gap 5 (FG5). The primary dataset therefore comprises 4 transition conditions x 10 target frames x 3 frame gaps = 120 SSIM observations. The independent variables are transition condition and frame gap; the dependent variable is the SSIM score between the FILM output and the original target frame.
The experiment simulates visual frame removal under controlled offline conditions. It does not emulate packet transmission, bitrate adaptation, codec behavior, jitter, consecutive packet loss, or end-to-end latency. Accordingly, the findings should be interpreted as evidence about reconstruction quality for the defined video asset, not as a direct measurement of network quality of service.

Reconstruction Workflow
For each target frame f1, two retained boundary frames, f_1 and f_2, were selected from the preceding and succeeding sides of the target frame. For a target frame at time t, FGg used retained boundary frames at $t - g$ and $t + g$, where $g \in \{1, 3, 5\}$. Thus, the target frame remained at the temporal midpoint, while increasing g increased the temporal separation and reconstruction difficulty. FILM then estimates the intermediate image as:

(3)

Where denotes the pre-trained FILM interpolation model[1]. The original target frame was withheld from the model input and stored separately as the ground-truth reference. This separation prevents information leakage and ensures that the reconstructed output is evaluated only against information that was not used during inference.

The source video was recorded using a Low Light video at 30 fps, with an original duration of 7 minutes and 34 second and an original resolution is 1280 × 720 pixels before frame extraction. Target frames were selected based on frame between three type of frame gap, and the same selection procedure was applied across all transition conditions.



Figure 2. Outline Process

Figure 2 summarizes the operational sequence used to reconstruct a missing frame. First, the source low-light smartphone video is decoded into an ordered frame sequence. Next, one frame is selected as the target frame and deliberately removed from the model input. The two remaining boundary frames are supplied to FILM, which produces an estimated intermediate frame at the desired temporal position. The generated frame is subsequently compared with the withheld original frame. This controlled procedure enables direct assessment because the target frame is known and can serve as an objective ground-truth reference.

The workflow is repeated for every transition condition, target-frame sample, and frame-gap configuration. Consequently, the procedure produces a consistent set of reconstruction results that can be compared across no-transition, cross-dissolve, dip-to-black, and dip-to-white conditions. Although the workflow represents controlled frame removal, it does not simulate actual packet-level network loss, transmission delay, jitter, or adaptive video encoding. The experiment therefore evaluates offline visual reconstruction quality rather than network Quality of Service.

The workflow can be implemented in Google Colab[12]. OpenCV and FFmpeg support video decoding, frame extraction, and image-sequence operations[13],[14]. TensorFlow and TensorFlow Hub are used to load and execute the pre-trained interpolation model[15],[16], while NumPy supports array operations[17]. Matplotlib, MediaPy, ImageIO, and scikit-image support visualization, media handling, image input/output, and SSIM computation[18]-[21]. The FILM implementation follows the official Google Research repository[22]. For reproducibility, each run should record the source asset identifier, transition condition, target-frame index, boundary-frame indices, frame gap, resolution, model version, and inference parameters.

Evaluation and Statistical Interpretation

SSIM compares two images through luminance, contrast, and structural components. It can be written as:

(4)

Where denotes image mean, denotes variance or covariance, and C1 and C2 are stabilizing constants[11]. Scores closer to 1 indicate greater structural similarity between the reconstruction and the ground truth. Means and standard deviations were calculated for every transition-gap combination. The supplied dataset supports descriptive comparison. Inferential claims would require a fully specified analysis, including the pairing rule, test statistic, degrees of freedom, p-value, confidence interval, and effect size. Because these outputs are not recorded in the supplied workbook, this paper does not claim unverified statistical significance.

RESULT AND DISCUSSION

This section presents and interprets the reconstruction results obtained from the low-light smartphone video experiment. The analysis focuses on the structural similarity between FILM-generated frames and their corresponding ground-truth frames under four transition conditions and three frame-gap settings. The results are first described using mean SSIM and standard deviation values, followed by an interpretation of how transition type and temporal separation influence reconstruction quality.

Descriptive SSIM Results

Table 1 reports the mean SSIM and standard deviation for the four low-light transition conditions. Across all 24 condition-by-gap means, the aggregate SSIM is 0.9504 for FG1, 0.9354 for FG3, and 0.9152 for FG5. The pattern is consistent with the general VFI observation that larger temporal separation makes motion correspondence and synthesis more uncertain[5],[23].

Tabel 1. Mean SSIM and standard deviation for the low-light smartphone video

Condition	FG1		FG3		FG5	
	Avg.	Std.Dev	Avg.	Std.Dev	Avg.	Std.Dev
LLNT	0.9428	0.0614	0.9410	0.0605	0.9159	0.0819
LLCD	0.9591	0.0271	0.9506	0.0288	0.9435	0.0287
LLDTB	0.9396	0.0355	0.9231	0.0557	0.9118	0.0607
LLDTW	0.9600	0.0517	0.9270	0.0989	0.8896	0.1301

The aggregate mean decreased by 0.0150 SSIM points from FG1 to FG3 and by 0.0352 points from FG1 to FG5. In practical terms, the model preserved the main structure of the withheld frames more consistently when the retained boundary frames were temporally closer to the target. The effect of increased frame gap was not uniform across conditions. LLCD showed the smallest decline from FG1 to FG5 (0.9591 to 0.9435; a decrease of 0.0156), whereas LLDTW showed the largest decline (0.9600 to 0.8896; a decrease of 0.0704). Within this video asset, this pattern indicates that the selected cross-dissolve targets were comparatively stable, while the selected dip-to-white targets were more sensitive to wider temporal separation.

Observation-Level SSIM Results

Table 2 presents the complete observation-level SSIM results used to calculate the descriptive statistics in Table 1. Each row represents one withheld target frame from a specific low-light transition condition, while the FG1, FG3, and FG5 columns report the SSIM obtained at the corresponding frame-gap setting. The complete table contains 40 target-frame records across four transition conditions. Since each target frame is evaluated at three frame gaps, the table represents all 120 SSIM observations included in this study.

Tabel 2. Complete observation-level SSIM results for the low-light smartphone video

Condition	Test	FG1	FG3	FG5
Low Lighting (Non-)	1	0.9645	0.9618	0.9612
	2	0.9663	0.9638	0.9628
	3	0.9665	0.9592	0.9592
	4	0.9604	0.9605	0.9570
	5	0.9620	0.9615	0.9568
	6	0.9622	0.9622	0.9531
	7	0.9629	0.9599	0.9514
	8	0.9612	0.9605	0.9527
	9	0.7588	0.7595	0.7535
	10	0.9627	0.9607	0.7510
Low Lighting (Cross Dissolve)	1	0.9729	0.9682	0.9601
	2	0.9627	0.9561	0.9457
	3	0.9635	0.9580	0.9454
	4	0.9643	0.9404	0.9374
	5	0.8786	0.8678	0.8612
	6	0.9718	0.9693	0.9664
	7	0.9692	0.9646	0.9600
	8	0.9655	0.9536	0.9545
	9	0.9701	0.9613	0.9455
	10	0.9725	0.9669	0.9584
Low Lighting (Dip to Black)	1	0.9642	0.9596	0.9555
	2	0.9366	0.9182	0.9040
	3	0.8909	0.8119	0.7990
	4	0.9653	0.9551	0.9369
	5	0.9653	0.9653	0.9369
	6	0.9472	0.9582	0.9538
	7	0.9595	0.9540	0.9432
	8	0.9375	0.9232	0.9164
	9	0.8571	0.8202	0.7920
	10	0.9690	0.9643	0.9590
Low Lighting (Dip to White)	1	0.9808	0.9631	0.9311
	2	0.9823	0.9669	0.9262
	3	0.8091	0.6339	0.5037
	4	0.9809	0.9671	0.9447
	5	0.9797	0.9556	0.9248
	6	0.9926	0.9692	0.9613
	7	0.9800	0.9568	0.9281
	8	0.9445	0.9173	0.8869
	9	0.9809	0.9756	0.9287
	10	0.9691	0.9648	0.9600

The observation-level values show that condition-level means do not fully represent the behavior of every target frame. The highest observed score was 0.9926 for LLDTW Test 6 at FG1, indicating very close structural agreement for that specific target. In contrast, the lowest observed score was 0.5037 for LLDTW Test 3 at FG5. This target also declined from 0.8091 at FG1 to 0.6339 at FG3 and 0.5037 at FG5, illustrating that wider temporal separation can substantially reduce reconstruction reliability for particular frames. LLCD remained the most stable condition overall, although LLCD Test 5 produced lower but relatively stable values from 0.8786 at FG1 to 0.8612 at FG5. Therefore, mean performance should be interpreted together with observation-level variation.

Interpretation of the Observed Pattern

LLCD produced the highest and most stable results across the three gaps: 0.9591 at FG1, 0.9506 at FG3, and 0.9435 at FG5. It also has the smallest standard deviations. Within this source asset, the selected cross-dissolve targets therefore retained visual patterns that FILM could reconstruct consistently. This finding should not be generalized to all cross dissolves, because performance also depends on the specific motion, target-frame location, and content of the underlying shots.

LLDTW produced the highest FG1 mean (0.9600) but declined to 0.8896 at FG5 and had the largest FG5 standard deviation (0.1301). This indicates less stable reconstruction when the boundary frames were farther apart. A transition toward a bright-dominant frame can reduce visible structural cues in particular parts of the sequence, but this is an interpretation rather than a causal conclusion; direct inspection of each generated frame would be required to identify specific blur, ghosting, or exposure artifacts.

LLNT and LLDTB also maintained high similarity while decreasing with larger gaps. The overall pattern is compatible with prior VFI research identifying motion magnitude, occlusion, and trajectory ambiguity as major sources of difficulty[1]-[5], [23]-[25]. The present result is intentionally narrow: it describes the behavior of FILM on the tested low-light smartphone asset, rather than demonstrating that all low-light scenes are intrinsically easier or harder to interpolate.

Finally, controlled frame removal should not be considered equivalent to real transmission loss. In an actual networked video system, missing frames may be influenced by packet loss patterns, codec settings, compression artifacts, buffering, and transmission latency. A network-oriented extension should therefore model these factors while preserving the original frames as ground truth for evaluation. Separating the visual reconstruction problem from the transmission-loss problem would allow future studies to evaluate not only whether FILM can reconstruct missing frames, but also how effectively such reconstruction performs under realistic communication conditions.

Overall, the results provide descriptive evidence that FILM can achieve high structural similarity when reconstructing selected missing frames from the evaluated low-light smartphone video, particularly when the temporal separation between available boundary frames is small. However, reconstruction quality becomes more variable as the frame gap increases, with the magnitude of this effect depending on the transition condition and individual target frame. These findings should therefore be interpreted as an evaluation of FILM under the specific experimental conditions rather than as evidence of universal performance across low-light video or real-world frame-loss scenarios.

CONCLUSION

This study provides a focused evaluation of FILM for controlled missing-frame reconstruction in one researcher-owned low-light smartphone video. Four transition conditions, ten target frames per condition, and three temporal gaps yielded 120 SSIM observations. Within the evaluated video asset, LLCD was the most stable condition, whereas LLDTW showed the largest decline and variability at FG5. These findings provide a traceable baseline for this specific researcher-owned low-light smartphone video; they should not be interpreted as general performance estimates for low-light video interpolation. Future work should evaluate multiple videos, devices, motion conditions, and interpolation models using both spatial and temporal quality measures.

REFERENCES

[1] F. Reda et al., “FILM: Frame Interpolation for Large Motion,” ECCV, 2022. doi: 10.1007/978-3-031-20047-2_3.

[2] W. Bao et al., “Depth-Aware Video Frame Interpolation,” CVPR, 2019, pp. 3703-3712. doi: 10.1109/CVPR.2019.00383.

[3] Z. Huang et al., “Real-Time Intermediate Flow Estimation for Video Frame Interpolation,” ECCV, 2022, pp. 624-642. doi: 10.1007/978-3-031-19772-7_36.

[4] L. Lu et al., “Video Frame Interpolation with Transformer,” CVPR, 2022, pp. 3532-3542. doi: 10.1109/CVPR52688.2022.00352.

[5] D. M. Argaw and I. S. Kweon, “Long-Term Video Frame Interpolation via Feature Propagation,” CVPR, 2022, pp. 3543-3552.

[6] H. Jiang et al., “Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation,” CVPR, 2018, pp. 9000-9008.

[7] S. Niklaus, L. Mai, and F. Liu, “Video Frame Interpolation via Adaptive Separable Convolution,” ICCV, 2017, pp. 261-270.

[8] H. Lee et al., “AdaCoF: Adaptive Collaboration of Flows for Video Frame Interpolation,” CVPR, 2020, pp. 5316-5325.

[9] M. Choi et al., “Channel Attention Is All You Need for Video Frame Interpolation,” AAAI, 2020, pp. 10663-10671.

[10] Q. Hou, A. Ghildyal, and F. Liu, “A Perceptual Quality Metric for Video Frame Interpolation,” ECCV, 2022.

[11] Z. Wang et al., “Image Quality Assessment: From Error Visibility to Structural Similarity,” IEEE Trans. Image Process., 13(4), pp. 600-612, 2004. doi: 10.1109/TIP.2003.819861.

[12] Google Research, “Google Colab.” https://colab.research.google.com/.

[13] OpenCV, “OpenCV Documentation.” https://docs.opencv.org/.

[14] FFmpeg, “FFmpeg Documentation.” https://ffmpeg.org/documentation.html.

[15] TensorFlow, “TensorFlow Documentation.” https://www.tensorflow.org/api_docs.

[16] TensorFlow, “TensorFlow Hub.” https://www.tensorflow.org/hub.

[17] NumPy, “NumPy Documentation.” https://numpy.org/doc/.

[18] Matplotlib, “Matplotlib Documentation.” https://matplotlib.org/stable/.

[19] Google Research, “MediaPy Documentation.” https://google.github.io/mediapy/.

[20] ImageIO, “ImageIO Documentation.” https://imageio.readthedocs.io/.

[21] scikit-image, “skimage.metrics Documentation.” https://scikit-image.org/docs/stable/api/skimage.metrics.html

[22] Google Research, “frame-interpolation: FILM implementation.” https://github.com/google-research/frame-interpolation

[23] Z. Zhong et al., “Clearer Frames, Anytime: Resolving Velocity Ambiguity in Video Frame Interpolation,” ECCV, 2024.

[24] Y. Zhang et al., “LADDER: An Efficient Framework for Video Frame Interpolation,” arXiv:2404.11108, 2024.

[25] G. Wu et al., “Perception-Oriented Video Frame Interpolation via Asymmetric Blending,” CVPR, 2024, pp. 2753-2762.