Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Comparative Analysis of Video Frame Interpolation from Optical Flow to Diffusion Models

Tianlang Yin

Faculty of Science, University of Melbourne, Parkville, Australia

tianlangy1@student.unimelb.edu.au

Abstract. Video Frame Interpolation (VFI) is essential in handling video processing to fill in the gaps between the initial and final frames and increase temporal resolution. This method is critical in applications like frame rate up-sampling, slow-motion rendering, and video improvement. This work compares and evaluates the merits and limitations of several different VFI methods based on their structures and interpolation performance. This paper summarizes conventional optical flow-based methods, kernel-based models, hybrid models based on depth estimation, flow-agnostic convolutional models, Transformer models, and new generative diffusion models. In particular, this paper compares each method's structural form, movement handling ability, and efficiency. Experimental evaluation demonstrates that transformer models, as well as diffusion models, are superior in treating large and complicated motions. By comparison, models such as Flow-agnostic video representations (FLAVR) balance efficiency and accuracy, making them ideal for real-time processing. Experimental evaluations indicate that the development of VFI methods shifts toward data-driven and globally conscious structures to capture the richness of motions better. Such findings inform future research and advance the real-time handling of video applications.

Keywords: Video Frame Interpolation; Optical Flow; Transformer; Diffusion Model.

1. Introduction

Video frame interpolation (VFI) creates intermediate frames between two successive video frames. Because of its significant practical uses, VFI has been an important area of research in video processing for extended periods, such as speeding up game frame rates, making slow-motion playback smooth, and raising the limit regarding video frame rates [1,2]. By interpolating high-quality intermediate frames, VFI considerably improves visual perception—motions become smoother and natural, minimizing blur in motion and jittering. For instance, adding extra frames makes normal frame-rate content playable at higher refresh rates, especially for smooth-motion applications, such as sports replay or gaming content [1]. Critically, such enhancements may be made without costly high-speed cameras or dedicated hardware. Recent developments in Artificial intelligence-assisted VFI made high-frame-rate effects both affordable as well as remarkably accessible through software interpolation for smooth slow-motion as well as low-latency videos, as opposed to hardware capture, significantly reducing requirements in terms of hardware as well as visual enhancement in applications ranging from gaming to Virtual Reality (VR) screens [1].

Early VFI methods were based primarily on explicit motion compensation and estimation methods. Traditional methods generally estimate inter-frame movement (e.g., using optical flow or block matching), then warp and combine input frames to produce intermediate frames. While acceptable outcomes can be generated with motion-compensated interpolation, it tends to perform poorly in demanding scenarios—large motions, complicated motions, occlusions, or sudden illumination changes—causing inaccurate optical flow with noticeable artifacts such as ghosting. For these limitations, research has tried various enhancements. One method applies convolutional kernel-based interpolation, synthesizing new frames directly using learning pixel-wise convolution kernels rather than depending on optical flow maps [3]. These are slower, however, and perform poorly with large motions due to the constraints of their related kernel size. Another tendency is mixing various methodologies, for instance, linking optical flow with learned convolution kernels to harness their

respective strengths. Recently, deep learning changed VFI using end-to-end models that implicitly learn about movement and interpolation [4]. Most state-of-the-art algorithms utilize deep convolutional neural networks, frequently enhanced using advanced optical flow or transformer-formulated attention modules, with state-of-the-art accuracy [4,5]. Several use diffusion models, combining frames using bidirectional diffusion conditioned on initial and target frames [6], as another promising motion-management approach. Generally, research has moved on from hand-tailored algorithms towards data-driven neural designs. Each class of methods has weaknesses: optical flow methods retain details but are prone to errors in their estimate; convolution kernel-based methods provide accuracy at the cost of having coverage for movement; fully learned models are tough on their training data but provide high-quality outcomes. Recent research indicates that VFI's main challenge is achieving high-quality interpolation despite controlling computational costs.

This research offers an in-depth summary of video frame interpolation methods and their development over the years. This paper explains key concepts and methodological improvements in VFI in sequence, explaining its evolution in history and present times. This paper systematically compares the available VFI methods, ranging from classic algorithms to state-of-the-art deep learning algorithms, and discusses each category's technical principles. This paper compares representative methods on standard datasets, presenting their strengths and limitations regarding interpolation quality, speed, and robustness. This paper detects future trends through comparisons of methods, presenting the state-of-the-art in action. Also, this paper exhaustively explains the strengths and weaknesses of diverse methodologies, shedding light on why they succeed or fail depending on the circumstances. Ultimately, this paper identifies future research directions and open issues in VFI, proposing conceivable enhancement possibilities and innovative suggestions (like new network models or combinations with other video improvement tasks). This thorough review is an informative guide encapsulating previous and current work and pointing toward directions for future innovation in VFI.

2. Methodology

2.1. Dataset Description

Video frame interpolation research extensively uses benchmark datasets like Vimeo90K, University of Central Florida (UCF) 101, and Adobe240fps. Vimeo90K, designed by the Massachusetts Institute of Technology (MIT) and other research institutions, consists of about 64,612 training videos and 7,824 testing videos, each having a sequence of 7 frames at a resolution of 448×256 [7]. It is generated based on online video samples and is used for multiple video processing tasks, such as super-resolution, denoising, deblocking, and especially frame interpolation [7]. UCF101, initially made publicly available at UCF University, includes 13,320 real-world YouTube video samples belonging to 101 activity categories [8]. With a constant frame rate of 25 Frames Per Second (FPS) at a resolution of 320×240, UCF101 features rich scenarios and complicated motions [8]. Adobe240fps, made publicly available at Adobe Research, is composed of high-frame-rate recordings at the rate of 240 FPS, having 1,132 (about 376,000) samples for training and testing purposes [9]. Its high frame rate and high resolution prove helpful in measuring interpolation quality, especially in dealing with fine details and target blurring [9].

2.2. Overview

Video frame interpolation methods have evolved from conventional block-matching methods to hybrid, end-to-end deep learning and the most recent generative models. The progression in video frame interpolation methods would be as follows (shown in Fig. 1): conventional optical flow/block matching methods, hybrid models (with explicit structural information incorporated), end-to-end convolutional neural network models, modern models like Transformers, and diffusion models. Earlier methods mainly utilized motion estimation methods, where optical flow or block-matching methods were used for interpolation [10, 11]. Methods later utilized deep neural networks but

maintained explicit structural information (e.g., optical flow and depth in Depth-Aware Video Frame Interpolation (DAIN) [12]). Most recently, end-to-end deep learning models were based on convolutional networks and attention-based models like Transformers [6]. Recent developments bring about diffusion models based on stochastic generation and the prediction of frames to enhance interpolation quality [13]. This section conducts an in-depth review of different representative methods on this progressive path.

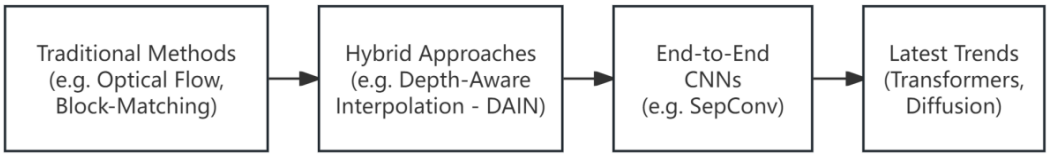

Fig. 1 The Pipeline of the Study (Picture credit: Original).

2.2.1. Traditional Methods

Conventional optical flow-based interpolation is based on estimating the motion fields between two neighboring frames. Warping and interpolating between them to produce an intermediate frame, as illustrated in Fig. 2. Traditional methods like TV-L1 optical flow and Edge-preserving interpolation of correspondences for optical flow (EpicFlow) try to estimate the motion vectors between the two frames and then transform them via warping operations to produce the intermediate frame [10,11]. These methods are computation-friendly owing to their explicit physical modeling, making them efficient for smooth and straightforward motions. However, their performance drops in complicated scenes with substantial motions, occlusions, or high-speed objects, where optical flow estimation errors are cumulative and result in apparent artifacts like blurring or ghosting.

For example, while EpicFlow does a great job of modeling motion boundaries correctly, localized blur and distortion can still be caused by even minor optical flow estimation errors. This illustrates the limitations of conventional optical flow algorithms in handling complicated motion and severe occlusion. Another group of traditional methods is based on kernel-based interpolation. Instead of directly estimating optical flow, these methods use deep neural networks that predict a spatially adaptive convolution kernel for each pixel, which is then used for sampling pixels in the input frames to produce the middle frame. Adaptive separable convolution (SepConv) is an example of predicting two separable 1D convolution kernels per pixel to convolve over the input frames [3]. This reduces the explicit need for motion estimation so the network can better generalize across different scenarios. Kernel methods are limited; enormous convolution windows are necessary for high motion, increasing computation costs and memory usage. Also, hybrid approaches melding classical methods and deep models have emerged. An excellent case is DAIN, where optical flow and depth estimation are combined in a deep learning framework [12]. PWC-Net is utilized in DAIN for extracting optical flow, and it uses an auxiliary depth estimation sub-network for deriving scene depth maps. Optical flow and depth are combined in DAIN with the help of a depth-aware projection layer to create a compensated flow field for rendering intermediate frames. The method performs better occlusion handling as it explicitly models depth-driven occlusion relations. Unfortunately, the higher complexity of DAIN, involving several modules (optical flow, depth estimation, kernel prediction), raises the computational cost and optical flow estimation sensitivity.

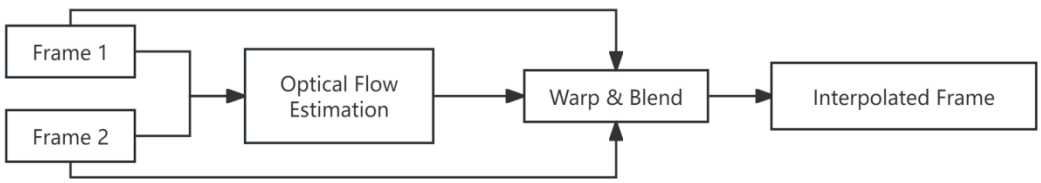

Fig. 2 Flowchart Illustrating the Traditional Frame Interpolation Method (Picture credit: Original).

2.2.2. Modern Methods

With advances in deep learning, end-to-end interpolation models eliminating explicit motion estimation have been the focus of recent studies (shown in Fig. 3). For instance, flow-agnostic video Representations (FLAVR) introduce a flow-agnostic, one-pass inference paradigm using a 3D spatiotemporal convolutional network [14]. In contrast with conventional methods, FLAVR learns motion representations directly from data without estimating optical flow explicitly. It thus enjoys faster inference rates at comparable efficacy compared with previous methods. However, it tends to fall behind, especially in the case of scenes with complex motions or interactions, where convolutional networks are prone to not capturing far-reaching dependencies.

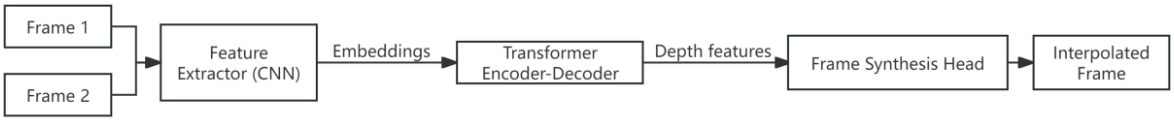

Fig. 3 Flowchart Illustrating the Modern Frame Interpolation Method (Picture credit: Original).

Researchers have developed Transformer-based frameworks for video frame interpolation in response to these limitations. Techniques like Video frame interpolation transformer (VFIformer) leverage the transformer's ability to capture long-term relationships by implementing self-attention mechanisms on spatiotemporal embeddings [6]. VFIformer effectively captures informative correlations between frames, making it possible to improve motion estimation and interpolation. VFIformer incorporates multi-scale feature fusion, increasing computational effectiveness and interpolation quality. Interestingly, compact models like Video frame interpolation transformer (VFIT-S) outperform large convolutional models like FLAVR but with significantly lower parameters [6].

One of the newer trends is using Diffusion Models for video frame interpolation. In contrast to deterministic methods, diffusion models use a stochastic generation pipeline in that intermediate frames are continually improved with iterative denoising and optimization. An example is Motion-aware latent diffusion models for video frame interpolation (MADiff), utilizing a motion-conditioned latent diffusion model for iterative refinement of frames based on input conditions [13]. This method ensures robustness even for complicated motion patterns and occlusions, resulting in realistic interpolations even in harsh environments. Though diffusion models offer an exciting new paradigm, their high computational demand still limits actual implementation.

Overall, the development of VFI methods reveals an unmistakable tendency towards incorporating learning-based structures to enhance interpolation performance. While earlier methods involve explicit modeling of structure and motion, deep learning models offer better handling of varying scenes and sophisticated motions. Further, the introduction of Transformers, as well as diffusion models, reflects an emphasis on structures that can efficiently capture long-range dependencies, as well as probabilistic generation processes, for enhanced interpolation quality and robustness. Video frame interpolation methods have evolved from classical optical flow estimating methods towards hybrid models, end-to-end deep learning structures, and the most recent generative paradigms. Early methods were based on optical flow estimates, utilizing methods such as TV-L1 and EpicFlow [10, 11]. This entails estimating the motion vectors between two adjacent frames with a warping operation to generate the in-between frames. Although effective for smooth motions, such methods were insensitive towards occluding objects and objects with significant motions. Convolutional kernel-based approaches came as an alternative solution. SepConv utilized adaptive separable convolutional kernels for interpolating frames without explicit optical flow calculation [3]. Hybrid approaches such as DAIN later utilized depth information to enhance occlusion-free and high-quality interpolation in occluded or complex cases [12]. Contemporary methods such as FLAVR, designed purely based on convolutional networks, and VFIformer, based on transformer structures, have extended the limit of interpolation performance [6,14]. Generative methods such as MADiff, with diffusion models, are an advanced line of research promising progress in handling complicated and ambiguous motions [13].

3. Results and Discussion

3.1. Results

There are substantial variations in performance among different video frame interpolation algorithms on scene categories and motion types. Conventional optical flow algorithms (e.g., TV-L1 and EpicFlow) are stable in handling smooth and straightforward motions. Still, they are susceptible to ghosting or blurring artifacts in the case of strong motions or the presence of occlusions due to the compounding of optical flow approximations' errors. For example, EpicFlow is superior in preserving edges but fails to appropriately handle occlusions and nonlinear motions [11]. Kernel-based convolutional methods avoid direct optical flow estimation through learning adaptive convolutional kernels, thus better handling moderate amplitude motion fields. SepConv uses separable one-dimensional convolutions, performing better on most datasets than light-flow algorithms with more accurate edges and fewer artifacts during motion [3]. The hybrid model DAIN combines optical flow and depth information, outperforming in handling occlusions. By leveraging depth perception, it selects a more reasonable pixel mapping path during intermediate frame generation, effectively mitigating ghosting and blurring caused by occlusions [12]. Its interpolation error on Vimeo90K and UCF101 datasets is lower than comparable methods, particularly at moving boundaries, where results appear more natural and clear [12]. The pure convolutional model FLAVR adopts a 3D convolutional structure, dispensing with optical flow estimation and directly learning spatiotemporal features from video data. This single-inference framework accelerates interpolation while maintaining accuracy, making it especially suitable for real-time applications [14]. FLAVR surpasses DAIN on several benchmarks and demonstrates computational efficiency advantages [14]. Transformer-based methods such as VFIformer enhance large-scale motion modeling capabilities through global self-attention mechanisms. In complex motion scenarios, VFIformer preserves structural integrity and texture clarity, achieving superior structural similarity index (SSIM) and Peak signal-to-noise ratio (PSNR) metrics compared to traditional convolutional neural network (CNN) methods like FLAVR [6].

Additionally, lightweight models such as VFIT-S achieve similar or even better performance than FLAVR while maintaining a low parameter count [6], enabling high-quality interpolation under mobile devices or resource-constrained environments. The diffusion model MADiff represents an emerging generative interpolation approach that progressively optimizes intermediate frames via a hidden variable diffusion process informed by motion perception. On the UCF101 dataset, this method produces highly realistic results, particularly in video clips featuring complex textures and fast motion, significantly reducing blur and ghosting artifacts [13]. Despite its slower production speed, MADiff stands out in visual evaluations for its detailed presentation and natural transitions [13].

3.2. Discussion

Contemporary video frame interpolation methods exhibit diversity and technological evolution in structural design and performance outcomes. Traditional optical flow methods such as TV-L1 and EpicFlow rely on explicit motion estimation, characterized by clear structures and ease of implementation, but lacking robustness, especially in occlusion and large displacement scenarios where artifacts and ambiguities are likely to occur [10, 11]. Although optical flow estimation accuracy has improved recently, its adaptability to complex dynamic content remains limited. Convolution kernel-based methods (e.g., SepConv) effectively address the dependency of traditional methods on accurate optical flow estimation. Through pixel-level convolution kernel prediction, these methods excel in detail processing and edge preservation and are well-suited for medium-motion scenes. However, their local nature within the convolution window restricts their ability to capture global changes in large-motion scenarios [3, 14]. Hybrid models such as DAIN introduce depth estimation modules to assist optical flow inference, performing exceptionally well in occlusion areas. This architecture's strength is combining scene geometric information to enhance pixel mapping accuracy.

Nevertheless, multi-module collaborative processing increases training complexity and reduces inference speed [12]. Pure CNN methods like FLAVR adhere to an end-to-end design philosophy, utilizing three-dimensional convolutions to capture spatiotemporal characteristics with a compact overall structure and high operational efficiency. This makes them ideal for deploying highly timed systems [14]. However, the model's generalization capability heavily depends on the quality of training data, potentially leading to performance degradation in unseen scenarios. With the advent of Transformer architectures, video interpolation tasks benefit from enhanced global modeling capabilities. VFIformer improves the model's adaptability to fast or large-scale motion by addressing long-term dependencies through the self-attention mechanism [6]. Lightweight Transformer models such as VFIT-S maintain or even exceed the interpolation quality of traditional CNN methods while reducing computational overhead [6], demonstrating a favorable balance between performance and efficiency. The emergence of diffusion models like MADiff introduces a novel generative perspective for frame interpolation. The iterative optimization mechanism enriches the details of interpolated frames, particularly suited for high-complexity scenes. These methods excel in visual quality, especially in subjective evaluations, though challenges remain regarding inference speed and deployment costs [13].

In summary, future research could focus on integrating the strengths of multiple model classes. For example, it could explore combining the Transformer's global perception with CNN's efficient structure or use FLAVR's rapid output as the initial value for diffusion models to enhance inference efficiency. Cross-domain adaptive training and time-series consistency modeling should be strengthened to improve model generalization in real-world applications. Furthermore, developing evaluation metrics more aligned with human visual perception will help better reflect the actual performance of interpolation methods in viewing experiences.

4. Conclusion

This study systematically analyzes the technical characteristics and performance of various video frame interpolation methods, aiming to elucidate the development trends in the field of video frame interpolation (VFI) and clarify the strengths and weaknesses of each approach. Specifically, the structural features and interpolation effects of traditional optical flow, convolution kernel, hybrid models, pure convolution, Transformer, and diffusion models are discussed. Experimental analysis reveals that Transformer models (e.g., VFIformer) and diffusion models (e.g., MADiff) demonstrate exceptional performance in interpolation quality, particularly in handling complex motion and dynamic texture scenes. Pure convolution models (e.g., FLAVR) balance performance and efficiency better, making them more suitable for real-time application scenarios. Future research will optimize advanced models' computational efficiency and generalization capabilities, focusing on efficiency improvements, model compression, and generalization performance optimization for Transformer and diffusion models, thereby further promoting the widespread application of video frame interpolation technology in practical contexts. Additionally, exploring richer application scenarios and developing more comprehensive evaluation indicators will become critical for subsequent research.

References

[1] MENGISTU Biruk. Deep-Learning Realtime Upsampling Techniques in Video Games. Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal. 2023, 10(2): 4.

[2] ZHU Tian Yi, et al. Generative Inbetweening through Frame-wise Conditions-Driven Video Generation. 2024.

[3] WIJMA Ruth, YOU Shao Di, and LI Yu. Multi-level adaptive separable convolution for large-motion video frame interpolation. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 1127-1135.

[4] HUANG Zhe Wei, et al. Real-time intermediate flow estimation for video frame interpolation. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 624-642.

[5] ZHAO Bin, and LI Xue Long. Edge-aware network for flow-based video frame interpolation. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(1): 1401-1408.

[6] SHI Zhi Hao, et al. Video frame interpolation transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 17482-17491.

[7] CHEN Mu, et al. General and Task-Oriented Video Segmentation. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 72-92.

[8] VRSKOVA Roberta, et al. Human activity classification using the 3DCNN architecture. Applied Sciences, 2022, 12(2): 931.

[9] ZHU Qi, et al. Exploring temporal frequency spectrum in deep video deblurring. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 12428-12437.

[10] ROMERA Thomas, et al. Optical flow algorithms optimized for speed, energy and accuracy on embedded GPUs. Journal of Real-Time Image Processing, 2023, 20(2): 32.

[11] REVAUD Jerome, et al. Epicflow: Edge-preserving interpolation of correspondences for optical flow. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1164-1172.

[12] BAO Wen Bo, et al. Depth-aware video frame interpolation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3703-3712.

[13] HUANG Zhi Lin, et al. Motion-aware latent diffusion models for video frame interpolation. Proceedings of the ACM International Conference on Multimedia. 2024, 1043-1052.

[14] KALLURI Tarun, et al. Flavr: Flow-agnostic video representations for fast frame interpolation. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023, 2071-2082.