

Reimagining Reality: A Comprehensive Survey of Video Inpainting Techniques

Shreyank N Gowda^a, Yash Thakre^a, Shashank Narayana Gowda^a, Xiaobo Jin^b

^a *Vidiosyncracy, India*

^b *Department of Intelligent Science, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China*

Abstract

This paper offers a comprehensive analysis of recent advancements in video inpainting techniques, a critical subset of computer vision and artificial intelligence. As a process that restores or fills in missing or corrupted portions of video sequences with plausible content, video inpainting has evolved significantly with the advent of deep learning methodologies. Despite the plethora of existing methods and their swift development, the landscape remains complex, posing challenges to both novices and established researchers. Our study deconstructs major techniques, their underpinning theories, and their effective applications. Moreover, we conduct an exhaustive comparative study, centering on two often-overlooked dimensions: visual quality and computational efficiency. We adopt a human-centric approach to assess visual quality, enlisting a panel of annotators to evaluate the output of different video inpainting techniques. This provides a nuanced qualitative understanding that complements traditional quantitative metrics. Concurrently, we delve into the computational aspects, comparing inference times and memory demands across a standardized hardware setup. This analysis underscores the balance between quality and efficiency—a critical consideration for practical applications where resources may be constrained. By integrating human validation and computational resource comparison, this survey not only clarifies the present landscape of video inpainting techniques but also charts a course for future explorations in this vibrant and evolving field.

Keywords: Video Inpainting, Deep Learning, Survey

1. Introduction

In the realm of video understanding, technologies like action recognition [1, 2, 3] and video object segmentation [4, 5, 6] have significantly advanced, enhancing our ability to analyze and interpret dynamic scenes. Building upon these developments, video inpainting has emerged as a captivating research field at the intersection of computer vision, image processing, and machine learning. The exponential growth of multimedia content, coupled with the increasing demand for seamless video editing, has underscored the need for automatic methods that can effectively fill in missing or corrupted regions in videos. Video inpainting techniques aim to address this challenge by reconstructing these regions in a visually plausible manner, enabling a wide range of applications such as video restoration [7, 8], object removal [9, 10, 11], and content creation [12, 13, 14].

The fundamental objective of video inpainting is to infer missing or damaged information based on the available surrounding content, often utilizing temporal and spatial cues within the video sequence. This process involves sophisticated algorithms that analyze the context of the video frames and synthesize new content to seamlessly blend with the existing visual information. By leveraging the inherent structure and patterns within the video frames, inpainting algorithms strive to create visually coherent and continuous video streams where the inpainted regions are virtually indistinguishable from the original content.

This survey paper aims to provide a comprehensive overview of the state-of-the-art video inpainting techniques developed to date. By exploring the advancements in this field, we seek to shed light on the diverse methodologies, innovations, and challenges encountered by researchers. We categorize video inpainting methods based on their underlying principles and highlight their key contributions, enabling readers to grasp the evolution of techniques. To the best of our knowledge, an existing survey for video inpainting does not exist and hence believe this contribution is vital to the field.

We categorize these methods into distinct groups, each characterized by its unique set of approaches and underlying principles. Firstly, we discuss Patch-Based Methods, which primarily involve substituting missing areas in a video frame with patches extracted from other parts of the same video. This category includes Exemplar-Based Inpainting [15, 16, 17, 18], leveraging best matching patches for filling holes, and Spatio-Temporal Patch Matching [19, 20, 21], which extends this concept by considering both spatial and temporal

elements in patch selection.

Next, we explore Motion-Based Methods. These techniques utilize the motion information inherent in video sequences to guide the inpainting process. Notable approaches within this category include Optical Flow Estimation [22, 23, 24] and Motion Compensation [25, 26, 10], both striving to maintain temporal coherence in dynamic scenes.

Finally, we delve into the emerging category of Diffusion-Based Methods for Video Inpainting. These methods represent a novel approach, simulating the diffusion of pixel information from adjacent areas to seamlessly restore damaged or missing regions. This section also touches upon hybrid techniques that combine diffusion-based strategies with other methods for enhanced effectiveness [27, 28].

By categorizing these methods in this manner, we not only simplify the complexity inherent in video inpainting but also shed light on the research trajectory and potential advancements within each category. This classification allows for a clearer understanding of the field's current state and the various approaches being explored to address the challenges in video inpainting.

Throughout this survey, we discuss the influence of various video characteristics, including scene complexity, motion dynamics, and temporal coherence, on the performance of video inpainting methods. Additionally, we address the challenges and open problems that need to be tackled in the field. These challenges encompass improving the handling of large and irregular occlusions, preserving fine details and textures, adapting to dynamic scenes with camera motions, and effectively handling diverse video content. By identifying and understanding these challenges, we hope to inspire researchers to develop novel solutions and push the boundaries of video inpainting. In summary, this survey aims to provide researchers and practitioners with a comprehensive understanding of video inpainting techniques, their advancements, and the challenges that lie ahead. By consolidating the knowledge accumulated in this rapidly evolving field, we hope to stimulate further research and innovation.

In addition to reviewing existing literature, we contribute to the field by conducting a practical evaluation of video inpainting methods. We reimplement five prominent papers from different categories of techniques discussed in this survey. These papers are selected based on their impact, representation of different approaches, and availability of code and models. We carefully compare these methods on a selected number of test samples, taking

into consideration both the inference speeds and the quality of inpainted outputs. To ensure a comprehensive evaluation, we employ human annotators to validate the quality of the inpainted videos. Human feedback plays a critical role in assessing the perceptual quality and determining the extent to which the inpainted regions blend seamlessly with the surrounding content. By incorporating these human-annotated scores, we aim to provide a more holistic assessment of the inpainting methods, capturing both objective and subjective aspects of the results.

2. Fundamentals of Video Inpainting

Video inpainting is a process used in the field of computer vision and graphics to modify a video by filling in missing or undesired parts of the visual data. The term "inpainting" originally comes from the art restoration field, where it refers to the process of reconstructing damaged parts of paintings or photographs. In the context of video, inpainting serves several purposes:

- Restoration: This is similar to the original meaning in art restoration. In video, inpainting can be used to restore old or damaged footage, removing artifacts such as scratches, dust, or other types of noise that degrade the quality.
- Object Removal: One of the most common uses of video inpainting is to remove unwanted objects or features from a video. This could be anything from an accidental passerby in a shot, to a piece of modern infrastructure in a historical movie scene, or even watermarks.
- Special Effects: In filmmaking and video production, inpainting can be used to achieve specific visual effects. For instance, it can be used to create the illusion of magic tricks, disappearing objects, or to manipulate the video in ways that would be difficult or impossible to achieve during filming.
- Editing: Video inpainting can also be used in post-production editing to alter the content of a scene without having to reshoot it. This can be more cost-effective and time-efficient.
- Privacy and Security: Inpainting can be used to obscure faces, license plates, or other sensitive information in videos to protect privacy or for security reasons.

The challenge in video inpainting, as opposed to still image inpainting, is maintaining consistency across frames. The inpainted regions must match the surrounding area in terms of texture, color, and lighting, and must also look natural as the scene moves and changes over time. Advances in AI and machine learning have significantly improved the capabilities and results of video inpainting in recent years.

Video inpainting involves complex techniques that combine the principles of computer vision, image processing, and machine learning. The underlying goal is to create a visually plausible and consistent output that fills in missing or unwanted parts of a video. One of the earlier techniques used for video inpainting is patch-based inpainting [29, 16], which involves copying and pasting patches (small blocks of pixels) from other parts of the video to fill in the missing or undesired areas. The algorithm searches for the most similar patches in terms of texture and color to ensure a visually coherent result, although this technique is more effective for static backgrounds and can struggle with complex, dynamic scenes.

Maintaining temporal consistency is crucial in video inpainting [30, 31]. The inpainted areas must not only blend seamlessly with the spatial surroundings in a single frame but also across the temporal dimension. Techniques used for maintaining temporal consistency often involve tracking the movement and changes in the scene over time to ensure that the inpainted regions behave in a physically plausible manner. With advancements in AI, deep learning models, particularly Generative Adversarial Networks (GANs) [32], have been applied to video inpainting. These models are trained on large datasets and can generate new content that fills in the missing parts of a video, and are particularly effective in handling complex scenes and dynamic objects, offering more realistic and coherent results.

Optical flow [33] algorithms are used to estimate the motion between video frames. This information is crucial for ensuring that the inpainted content aligns [20, 34, 35] correctly with the movement in the video, maintaining the illusion of a continuous, unedited scene. Exemplar-based [15, 16] inpainting extends the idea of patch-based inpainting by considering the entire structure and texture of the area surrounding the hole or undesired object. The algorithm finds similar regions in the video and uses them as references to fill in the missing parts, which helps in maintaining both spatial and temporal coherence. In some cases, especially in professional video editing and restoration, inpainting involves a level of manual input or guidance. Users may define the areas to be inpainted, set motion paths, or make

adjustments to ensure the desired outcome.

Advanced inpainting systems incorporate elements of scene understanding [36, 37], where the algorithm interprets the context of the scene to make more informed decisions about how to fill in gaps. This can include predicting the trajectory of moving objects or understanding the overall geometry and depth of the scene. Each of these techniques has its strengths and is chosen based on the specific requirements of the video inpainting task, such as the complexity of the scene, the nature of the missing or unwanted content, and the desired level of automation and control. The field is continuously evolving, with ongoing research aiming to improve the realism, efficiency, and versatility of video inpainting methods.

3. Taxonomy of Video Inpainting Techniques

Video inpainting is a dynamic field that encompasses various techniques designed to reconstruct missing or corrupted parts in video frames. This section elaborates on a more comprehensive taxonomy of these techniques.

3.1. Patch-Based Methods

Patch-based methods involve replacing missing regions in a video frame with patches from other parts of the video.

3.1.1. Key Ideas

These methods rely on the assumption that similar patterns or textures exist within different parts of the video, which can be used to fill in missing regions.

3.1.2. Algorithms and Approaches

- *Exemplar-Based Inpainting:* These methods involve filling the holes by copying the best matching patches from known regions, often using a priority scheme. Shih et al. [15] improves video inpainting by extending an image inpainting algorithm with enhanced patch matching and robust tracking, thus reducing "ghost shadows" and ensuring temporal continuity in different motion segments. Koochari et al [16] use large patches to effectively fill missing parts in video frames, separating moving objects from a stationary background and creating a mosaic for better representation of occluded objects. Lee et al. [17] introduce

a deep neural network-based "Copy-and-Paste Networks" for video inpainting, using content from reference frames to effectively fill missing regions while maintaining temporal coherency.

- *Spatio-Temporal Patch Matching*: These extend exemplar-based inpainting by considering spatial and temporal dimensions for patch matching. Newson et al. [38] enhance PatchMatch [39] for spatio-temporal use, significantly speeding up high-resolution video inpainting while addressing over-smoothing issues. Kim et al. [40] use flow sub-networks and mask sub-networks to model PatchMatch. Le et al. [20] utilize motion matching ensuring that the inpainted areas align seamlessly with the surrounding video content's motion patterns. Meanwhile, Liu et al. [41] propose Temporal Adaptive Alignment Network that aligns and adapts temporal features using patch matching. More recently, Cai et al. [21] utilize deformed vision transformers to effectively reconstruct and inpaint missing or corrupted areas in video content by using Deformed Patch-based Homography (DePtH), which improves patch-level feature alignments.

3.2. Motion-Based Methods

Motion-based methods utilize the movement information in videos to achieve inpainting, particularly effective in scenes with consistent motion.

3.2.1. Key Ideas

These techniques use motion estimation to guide the inpainting process, ensuring temporal coherence and consistency in dynamic scenes.

3.2.2. Algorithms and Approaches

- *Optical Flow Estimation*: Zhang et al. [35] propose repairing or reconstructing missing or damaged areas in video sequences using internal learning mechanisms, leveraging the video's own data and optical flow for inpainting. Work by Ding et al. [22] incorporates ConvLSTM and optical flow for efficient, real-time processing of large and arbitrary length videos. Wu et al. [23] develop an approach for predicting future video frames based on a sequence of past continuous video frames using optical flow. More recently, Li et al. [42] use a comprehensive framework for video inpainting, utilizing three key modules: flow completion, feature propagation, and content hallucination, to facilitate

an efficient and effective inpainting process. Kang et al. [43] propose a framework that enhances flow-based video inpainting methods by introducing a newly designed flow completion module and an error compensation network that leverages an error guidance map, aiming to offset the weaknesses of traditional flow-based methods. Zhang et al. [44, 45] enhance video inpainting using optical flow guidance in a transformer-based framework. Gao et al. [24] propose video completion that combines optical flow and edge information.

- *Motion Compensation:* These methods involve aligning frames based on estimated motion to maintain consistency. Kim et al. [46] propose an approach extending deep learning-based image inpainting methods to the video domain which involves an additional time dimension. Ebdelli et al. [25] present an algorithm for video inpainting that estimates unknown pixels as a linear combination of the closest patches using motion-compensated neighbor embedding. March et al. [26] delve into the mathematical underpinnings of motion compensated inpainting in video processing, focusing on the application of Euler equations and analysis of minimizers in this context. Zhang et al. [47] introduce PFTA-Net, a novel network that incorporates a progressive feature alignment module with sub-alignments and a progressive refinement scheme for more accurate motion compensation in video inpainting. Liu et al. propose Fuseformer [10], a transformer-based architecture for video inpainting that enhances the quality of reconstructed video by integrating fine-grained details across frames.

3.3. Diffusion-Based Methods for Video Inpainting

Diffusion-based methods represent a novel approach in the field of video inpainting, utilizing advanced techniques to enhance the restoration of missing or corrupted parts in video sequences.

3.3.1. Key Ideas

Diffusion-based video inpainting leverages the concept of iteratively refining the inpainted area by simulating the diffusion of pixel information from the surrounding areas. This process gradually introduces information into the damaged or missing regions, ensuring a natural and seamless restoration.

3.3.2. Algorithms and Approaches

- *Extending Image-Based Diffusion to Video:* Recent advancements involve deep learning models that simulate diffusion processes. These models are trained to predict the natural flow of pixel information, making them highly effective for complex video sequences with varying textures and motion. The "Pix2video" work completed by Ceylan et al. [48] introduces a training-free, text-guided video editing method using image diffusion models that edit an anchor frame and then propagate changes to subsequent frames. Hoppe et al. [49] introduce the Random-Mask Video Diffusion (RaMViD) method, which extends image diffusion models to video using 3D convolutions and a new conditioning technique, enabling video prediction, infilling, and upsampling. Cherel et al. [50] present a technique using internal diffusion processes for video inpainting, focusing on improving the quality and consistency of inpainted video content. Voleti et al [9] introduce a versatile video diffusion model that excels in video prediction, generation, and interpolation by leveraging a masked conditional approach.
- *Hybrid Approaches:* Combining diffusion-based techniques with other methods, such as optical flow or patch-based inpainting, allows for more robust solutions that can handle a variety of inpainting challenges in videos. Gu et al. [27] introduce the Flow-Guided Diffusion model for Video Inpainting (FGDVI), which enhances temporal consistency and inpainting quality in videos by reusing an existing image generation diffusion model. Wang et al. [28] present a method that combines local and nonlocal flow guidance to improve the quality and consistency of video inpainting.

Whilst a lot of methods can be classified broadly within these groups, there are approaches which would fit loosely and hence we believe such a style of classification of methods would be easier to summarize and understand. This also gives us an idea on the direction of research among these methods and the potential to improve them.

4. Challenges in Video Inpainting

We identify and discuss key factors such as scene complexity, motion dynamics, and temporal coherence, which significantly impact the performance and applicability of inpainting techniques.

4.1. Scene Complexity

Scene complexity in videos, characterized by varying textures, colors, and patterns, poses a substantial challenge for inpainting algorithms. The ability to accurately reconstruct complex scenes while maintaining visual realism is critical. As outlined in multiple previous works [36, 38, 19], the complexity of a scene directly influences the choice and performance of inpainting methods, necessitating more advanced techniques for detailed and intricate scenes.

4.2. Motion Dynamics and Temporal Coherence

The dynamics of motion within a video significantly affect the inpainting process. Effective inpainting must account for and replicate the natural motion patterns to ensure seamless integration of the inpainted region into the video. This aspect is particularly challenging in videos with rapid or unpredictable motion. We already speak about methods with a specific focus on motion compensation and optical flow prediction. This emphasizes the importance of incorporating motion understanding in video inpainting, especially in high-motion scenes. Maintaining temporal coherence, the consistency of visual elements over time, is essential for the believability and quality of inpainted videos. Temporal incoherencies can lead to noticeable artifacts in video inpainting, thus undermining the overall quality. Advanced methods strive to ensure that inpainted regions remain consistent and coherent throughout the video duration.

4.3. Challenges and Open Problems

We also address several pressing challenges and open problems in the field of video inpainting:

- **Handling Large and Irregular Occlusions:** Dealing with large and irregularly shaped occlusions [51, 30, 52] remains a significant challenge, requiring innovative approaches to ensure effective and realistic inpainting.
- **Preserving Fine Details and Textures:** The preservation of fine details and textures, crucial for realism, is a complex task, especially in highly textured areas. Methods for enhancing texture fidelity in video inpainting has been a strong research area [53, 35, 54].

- **Adapting to Dynamic Scenes with Camera Motions:** Dynamic scenes, especially those with camera motions, pose unique challenges. As explored previously [31, 55, 20], adapting inpainting methods to accommodate camera movements is essential for realistic reconstructions.
- **Effectively Handling Diverse Video Content:** The diversity of video content, ranging from simple to complex scenes, requires versatile and robust inpainting solutions.

By consolidating the knowledge accumulated in this rapidly evolving field, we hope to stimulate further research and innovation. Understanding these challenges and the influence of various video characteristics is crucial for the development of more effective and versatile video inpainting methods.

5. Evaluation Metrics

In the realm of video inpainting, a key challenge is to evaluate the effectiveness of various algorithms in reconstructing video sequences. The quality of video inpainting is not just about replacing missing or damaged areas; it's about doing so in a way that the viewer perceives the video as uninterrupted and authentic. To quantify this, several evaluation metrics are employed, each focusing on different aspects of video quality and perceptual integrity.

Peak Signal-to-Noise Ratio (PSNR): PSNR is a commonly used metric for measuring the quality of reconstruction in video and image processing. It calculates the ratio between the maximum possible power of a signal and the power of distorting noise that affects the fidelity of its representation. In video inpainting, PSNR is used to compare the inpainted video against a ground truth or original video to assess the quality of reconstruction.

The formula for PSNR is given by:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right)$$

where MAX is the maximum possible pixel value, and MSE is the Mean Squared Error between the original and inpainted videos.

Structural Similarity Index (SSIM): SSIM is a perception-based metric that measures the similarity between two images or videos. It considers changes in texture, luminance, and contrast, providing a more comprehensive assessment than PSNR. In video inpainting, SSIM is utilized to evaluate

how well the inpainted areas blend with the surrounding context in terms of perceived changes in structural information.

The formula for SSIM is a combination of luminance, contrast, and structure terms:

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

where μ_x , μ_y are the local means, σ_x , σ_y are the local standard deviations, σ_{xy} is the local cross-covariance, and C_1 and C_2 are constants to stabilize the division.

Frechet Inception Distance (FID): FID is a metric that compares the distribution of generated images to real images in an embedding space. It evaluates both the visual quality of individual frames (e.g., clarity, color accuracy) and the temporal consistency across frames in the video. This is crucial in video inpainting, where maintaining a seamless and consistent flow from frame to frame is as important as the quality of individual frames.

The formula for FID is given by:

$$FID = \|\mu_{\text{real}} - \mu_{\text{generated}}\|^2 + \text{Tr}(\Sigma_{\text{real}} + \Sigma_{\text{generated}} - 2(\Sigma_{\text{real}}\Sigma_{\text{generated}})^{1/2})$$

where μ_{real} and Σ_{real} are the mean and covariance matrix of feature vectors from the real video, and $\mu_{\text{generated}}$ and $\Sigma_{\text{generated}}$ are the mean and covariance matrix of feature vectors from the generated video.

Video Frechet Inception Distance (VFID): VFID adapts the concept of FID for video. It extends the evaluation to assess not only the quality of individual frames but also the temporal consistency between frames in the generated video compared to the real video.

The formula for VFID is an extension of FID, considering both the frame-wise features and the temporal dynamics:

$$VFID = FID_{\text{frames}} + \alpha \cdot FID_{\text{temporal}}$$

where FID_{frames} is the FID calculated for individual frames, FID_{temporal} is the FID calculated for temporal dynamics, and α is a weighting factor to balance their contributions.

Learned Perceptual Image Patch Similarity (LPIPS): Learned Perceptual Image Patch Similarity (LPIPS), initially developed for assessing the perceptual similarity between images, has been adapted and extended

for use in evaluating videos. This metric is particularly valuable in fields like video generation, video compression, and video editing, where it's important to assess how perceptually similar a generated or altered video is to a reference video. LPIPS measures the distance in VGGNet feature space as a “perceptual loss” for image regression problems.

In video inpainting evaluation, each metric offers a distinct perspective on the quality and effectiveness of video restoration. Peak Signal-to-Noise Ratio (PSNR) is a basic, technical measure that compares videos pixel-by-pixel, focusing on the accuracy of signal reconstruction but not on human visual perception. In contrast, the Structural Similarity Index (SSIM) aligns more closely with human vision by assessing the luminance, contrast, and structural similarity between the original and inpainted videos, making it better suited for evaluating the perceptual integrity of inpainted areas. Video Frechet Inception Distance (VFID) takes a broader approach by evaluating both the visual quality of individual frames and the temporal consistency across a video sequence, thus ensuring a seamless flow in the inpainted video. Lastly, Learned Perceptual Image Patch Similarity (LPIPS) extends the perceptual evaluation to a more nuanced level, focusing on the perceptual similarity of video patches, which is particularly relevant for tasks like video generation and editing where the subtle nuances of human perception play a significant role. Each metric, therefore, caters to different aspects of video quality, from basic signal accuracy to complex perceptual fidelity.

5.1. Challenges and Limitations in Evaluating Video Inpainting Techniques

Subjective Nature of Quality Assessment: The assessment of video quality is often subjective, depending on human perception. Metrics like PSNR and SSIM may not fully capture the aesthetic and contextual quality perceived by viewers. This subjectivity can lead to discrepancies between quantitative metric scores and actual perceived quality.

Complexity of Dynamic Content: Video inpainting involves dealing with dynamic content and complex backgrounds, which are difficult to quantify with standard metrics. Metrics may not adequately account for the challenges in handling diverse and complex video content, leading to an incomplete evaluation.

Over-Reliance on Quantitative Metrics: There is often an over-reliance on quantitative metrics like PSNR and SSIM, which might not fully encompass the intricacies of video inpainting. This reliance can overshadow other important aspects such as user experience, realism, and context-awareness.

We show these limitations with varying user scores for different models in Section 7 and talk about how subjective video inpainting is and the need for a different metric to properly evaluate models.

6. Datasets

Dataset Name	Focus	Content Description	Resolution
DAVIS	VOS	Diverse objects	Full HD
YouTube-VOS	VOS	Diverse objects	Various
DEVIL	Inpainting	Camera Motion	Various

Table 1: Summary of Datasets for Video Inpainting Research

YoutubeVOS [56] and DAVIS [57] are two extensively utilized datasets in the field of video inpainting, each originally designed for specific tasks unrelated to video inpainting. YoutubeVOS, primarily developed for Video Object Segmentation, offers a diverse range of videos featuring various objects and scenarios, making it a valuable resource for understanding object motion and consistency in videos. DAVIS, on the other hand, was designed for video object segmentation and tracking, providing high-quality, full-resolution video sequences with annotated objects, which are crucial for training models to understand object movement and boundary information.

Despite their initial purposes, both datasets have been adapted for use in video inpainting tasks. This adaptation is primarily due to their comprehensive and varied content, which helps in training inpainting models to handle different scenarios and object movements. However, the lack of datasets specifically designed for video inpainting led to the introduction of the DEVIL [58] dataset. This dataset is tailored specifically for video inpainting, encompassing a wider range of challenges specifically encountered in this task, such as dealing with different types of occlusions and complex object interactions.

The absence of a strong, dedicated benchmark in video inpainting has historically hindered the field’s progress. This is because generic datasets like YoutubeVOS and DAVIS, while useful, do not fully address the unique challenges of video inpainting, such as temporal consistency and complex dynamic backgrounds. The creation of DEVIL marks a significant step towards overcoming these challenges, providing a more focused and relevant bench-

Model	Input	Time	Memory	User Rating
FGT [45] [ECCV’22]	80f	241s	2.6 GB	7.2
FGVC [24] [ECCV’20]	15f	50s	1.4GB	6.5
E2FGVI [42] [CVPR’22]	70f	18s	8.0GB	6.8
FuseFormer [10] [ICCV’21]	50f	39s	10.5GB	6.9
CopyPaste [17] [ICCV’19]	70f	44s	2.6GB	6.7

Table 2: Comparing inference hardware requirements of selected models and the number of frames needed as input to the model.

mark for advancing the state of the art in video inpainting. These datasets are summarized in Table 1.

7. Quality Evaluation and Comparison

We run all experiments using a Tesla T4 GPU. We evaluate the following models: Flow-Guided Transformer for Video Inpainting (FGT) [45], Flow-edge Guided Video Completion (FGVC) [24], End-to-End Framework for Flow-Guided Video Inpainting (E2FGVI) [42], FuseFormer [10] and Copy-Paste networks [17]. These methods range from 2019 to more recent methods and showcase how the outputs of these models have evolved over time. First, we compare quantitatively these methods in Table 2. In particular, we look at the inference details of these methods with a view to them being deployed. From these selected models, E2FGVI seems the most realistically deployable model accounting for speed and memory usage. We also report average scores by asking 20 human annotators to give a score on 1 to 10 for each model on a set of videos. The average score is reported under User Rating in the Table.

From a qualitative perspective, we evaluate on samples from both the evaluation set of datasets the model has been trained on and in-the-wild videos taken from public sources ¹. First, we look at a video containing flamingoes and methods that remove one of them. This can be seen in Figure 1.

Next, we use a video of people hiking and plot the frame wise results of removal in Figure 2.

Next, we use a video of person playing tennis and plot the frame wise results of removal in Figure 3.

¹<https://mixkit.co/free-stock-video/>

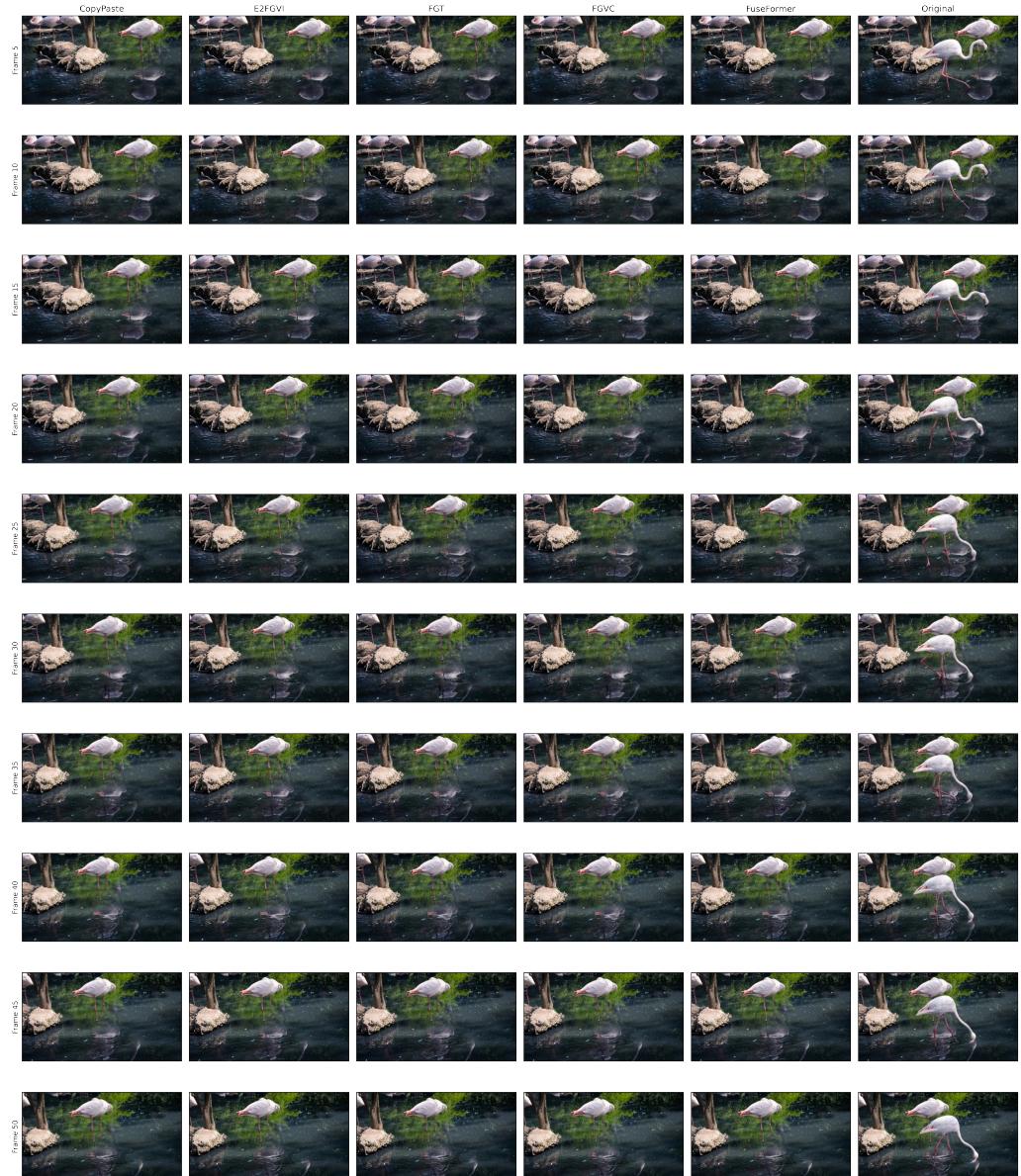


Figure 1: Comparison of different video inpainting methods on a flamingo video. Different frames are taken from the output and shown along with the original frame (ground truth).



Figure 2: Comparison of different video inpainting methods on a hiking video. Different frames are taken from the output and shown along with the original frame (ground truth).



Figure 3: Comparison of different video inpainting methods on a tennis video. Different frames are taken from the output and shown along with the original frame (ground truth).

8. Applications and Future Directions

8.1. Practical Applications of Video Inpainting

- **Video Editing:** Video inpainting is a pivotal tool in professional video editing, used for removing unwanted elements like microphones, wires, or unintentional background objects. This technology has found applications in various sectors including film production, news broadcasting, and content creation for digital platforms. It enhances the visual appeal and professionalism of videos by enabling editors to maintain continuity and visual coherence in their narratives.
- **Restoration:** Inpainting plays a critical role in the restoration of old or damaged footage. This includes repairing scratches, dust, and other defects in archival footage. Notable projects in film restoration have demonstrated how inpainting helps preserve historical and cultural content, providing clearer and more accessible versions of historically significant films and documentaries.
- **Visual Effects:** In the realm of visual effects, inpainting serves as a tool for creating seamless composites and backgrounds. It facilitates the blending of various elements in a scene, aiding in storytelling and the creation of fantastical environments. The use of inpainting in major film productions for generating realistic and visually stunning effects has become increasingly prevalent.

8.2. Current Trends and Emerging Research Directions

- **Diffusion Models:** Recent trends [48, 9] show the integration of diffusion models in video inpainting. These models, based on advanced machine learning algorithms, predict and recreate complex scenes with remarkable accuracy, surpassing traditional techniques in both quality and efficiency.
- **Real-Time Inpainting:** The pursuit of real-time video inpainting [59, 60, 61] aims to enable live modifications of video streams. This advancement is particularly significant for live broadcasting and virtual reality applications, where the ability to alter or enhance video content instantaneously is invaluable.

- **Increasing Automation:** The trend [48, 11] towards more automated inpainting processes is making this technology increasingly accessible. Automation reduces the need for manual intervention, thereby streamlining the video editing process and making it more efficient for users with varying levels of expertise.

8.3. Open Challenges and Potential Future Developments

- **Handling Complex Dynamics:** A significant challenge in video inpainting is dealing with intricate and rapid movements within videos. Future research may focus on developing more sophisticated algorithms that can understand and replicate complex dynamics, thus enhancing the realism and fluidity of inpainted content.
- **Enhanced Resolution and Quality:** As video standards continue to evolve towards higher resolutions like 4K and 8K, there is a growing need for inpainting techniques that can maintain detail and clarity at these levels. This development is crucial for ensuring that inpainting technology remains relevant and effective in the face of rapidly advancing display technologies.
- **Interactivity and User Control:** Future developments in video inpainting might include more interactive tools, providing users with greater control over the inpainting process. This would allow for more personalized and creative applications, catering to a wider range of use cases and artistic visions.

9. Conclusion

In conclusion, this paper has provided a thorough examination of the current state of video inpainting techniques, a vital niche in the realms of computer vision and artificial intelligence. Video inpainting, which involves the restoration or completion of missing or corrupted segments in video sequences, has significantly advanced through the integration of deep learning methodologies. Our comprehensive study breaks down key techniques, their foundational theories, and their real-world applications, illuminating a landscape that, despite its rapid evolution, remains complex and challenging. Central to our analysis was an exhaustive comparative study focusing on visual quality and computational efficiency. By employing a human-centric approach for assessing visual quality, involving a panel of annotators, we

gained valuable qualitative insights that enhance traditional quantitative assessments. This human validation aspect brought a unique perspective to our evaluation of various video inpainting methods. In tandem, we conducted a rigorous examination of the computational demands of these techniques, comparing inference times and memory requirements across a uniform hardware setup. Our findings highlight the crucial balance between quality and efficiency, especially pertinent in practical scenarios where resource limitations are a common constraint. This paper not only clarifies the intricate landscape of video inpainting techniques but also sets a direction for future research in this dynamic and rapidly evolving field. By merging human-centric evaluations with computational resource analysis, our study contributes to a deeper understanding of video inpainting and lays the groundwork for future innovations that are both high in quality and efficient in execution.

10. Acknowledgements

This work was partially supported by “Qing Lan Project” in Jiangsu universities, NSFC under No. U1804159 and RDF with No. RDF-22-01-020.

References

- [1] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 3551–3558.
- [2] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: A review, *IEEE transactions on pattern analysis and machine intelligence* (2022).
- [3] S. N. Gowda, M. Rohrbach, L. Sevilla-Lara, Smart frame selection for action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 1451–1459.
- [4] S. N. Gowda, P. Eustratiadis, T. Hospedales, L. Sevilla-Lara, Alba: Reinforcement learning for video object segmentation, *arXiv preprint arXiv:2005.13039* (2020).
- [5] R. Yao, G. Lin, S. Xia, J. Zhao, Y. Zhou, Video object segmentation and tracking: A survey, *ACM Transactions on Intelligent Systems and Technology (TIST)* 11 (4) (2020) 1–47.

- [6] Y.-T. Hu, J.-B. Huang, A. G. Schwing, Videomatch: Matching based video object segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 54–70.
- [7] X. Wang, K. C. Chan, K. Yu, C. Dong, C. Change Loy, Edvr: Video restoration with enhanced deformable convolutional networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, pp. 0–0.
- [8] S. Lee, D. Cho, J. Kim, T. H. Kim, Restore from restored: Video restoration with pseudo clean video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3537–3546.
- [9] V. Voleti, A. Jolicoeur-Martineau, C. Pal, Mcvd-masked conditional video diffusion for prediction, generation, and interpolation, Advances in Neural Information Processing Systems 35 (2022) 23371–23385.
- [10] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, H. Li, Fuseformer: Fusing fine-grained information in transformers for video inpainting, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 14040–14049.
- [11] R. Zhang, W. Li, P. Wang, C. Guan, J. Fang, Y. Song, J. Yu, B. Chen, W. Xu, R. Yang, Autoremover: Automatic object removal for autonomous driving videos, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 12853–12861.
- [12] S. Yun, S. J. Oh, B. Heo, D. Han, J. Kim, Videomix: Rethinking data augmentation for video classification, arXiv preprint arXiv:2012.03457 (2020).
- [13] S. N. Gowda, M. Rohrbach, F. Keller, L. Sevilla-Lara, Learn2augment: learning to composite videos for data augmentation in action recognition, in: European conference on computer vision, Springer, 2022, pp. 242–259.
- [14] S. N. Gowda, D. Pandey, S. N. Gowda, From pixels to portraits: A comprehensive survey of talking head generation techniques and applications, arXiv preprint arXiv:2308.16041 (2023).

- [15] T. K. Shih, N. C. Tang, J.-N. Hwang, Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity, *IEEE transactions on circuits and systems for video technology* 19 (3) (2009) 347–360.
- [16] A. Koochari, M. Soryani, Exemplar-based video inpainting with large patches, *Journal of Zhejiang University Science C* 11 (2010) 270–277.
- [17] S. Lee, S. W. Oh, D. Won, S. J. Kim, Copy-and-paste networks for deep video inpainting, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4413–4421.
- [18] S. N. Gowda, C. Yuan, Colornet: Investigating the importance of color spaces for image classification, in: *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision*, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14, Springer, 2019, pp. 581–596.
- [19] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, P. Pérez, Video inpainting of complex scenes, *Siam journal on imaging sciences* 7 (4) (2014) 1993–2019.
- [20] T. T. Le, A. Almansa, Y. Gousseau, S. Masnou, Motion-consistent video inpainting, in: *2017 IEEE international conference on image processing (ICIP)*, IEEE, 2017, pp. 2094–2098.
- [21] J. Cai, C. Li, X. Tao, C. Yuan, Y.-W. Tai, Devit: Deformed vision transformers in video inpainting, in: *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 779–789.
- [22] Y. Ding, C. Wang, H. Huang, J. Liu, J. Wang, L. Wang, Frame-recurrent video inpainting by robust optical flow inference, *arXiv preprint arXiv:1905.02882* (2019).
- [23] Y. Wu, R. Gao, J. Park, Q. Chen, Future video synthesis with object motion prediction, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5539–5548.
- [24] C. Gao, A. Saraf, J.-B. Huang, J. Kopf, Flow-edge guided video completion, in: *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, Springer, 2020, pp. 713–729.

- [25] M. Ebdelli, C. Guillemot, O. Le Meur, Examplar-based video inpainting with motion-compensated neighbor embedding, in: 2012 19th IEEE International Conference on Image Processing, IEEE, 2012, pp. 1737–1740.
- [26] R. March, G. Riey, Euler equations and trace properties of minimizers of a functional for motion compensated inpainting, *Inverse Problems and Imaging* 16 (4) (2022) 703–737.
- [27] B. Gu, Y. Yu, H. Fan, L. Zhang, Flow-guided diffusion for video inpainting, arXiv preprint arXiv:2311.15368 (2023).
- [28] J. Wang, Z. Yang, Z. Huo, W. Chen, Local and nonlocal flow-guided video inpainting, *Multimedia Tools and Applications* (2023) 1–20.
- [29] A. Criminisi, P. Pérez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, *IEEE Transactions on image processing* 13 (9) (2004) 1200–1212.
- [30] K. A. Patwardhan, G. Sapiro, M. Bertalmio, Video inpainting of occluding and occluded objects, in: IEEE International Conference on Image Processing 2005, Vol. 2, IEEE, 2005, pp. II–69.
- [31] K. A. Patwardhan, G. Sapiro, M. Bertalmío, Video inpainting under constrained camera motion, *IEEE Transactions on Image Processing* 16 (2) (2007) 545–553.
- [32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, *Communications of the ACM* 63 (11) (2020) 139–144.
- [33] B. K. Horn, B. G. Schunck, Determining optical flow, *Artificial intelligence* 17 (1-3) (1981) 185–203.
- [34] R. Xu, X. Li, B. Zhou, C. C. Loy, Deep flow-guided video inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3723–3732.
- [35] H. Zhang, L. Mai, N. Xu, Z. Wang, J. Collomosse, H. Jin, An internal learning approach to video inpainting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2720–2729.

- [36] D. Lao, P. Zhu, P. Wonka, G. Sundaramoorthi, Flow-guided video inpainting with scene templates, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14599–14608.
- [37] Y. Zeng, J. Fu, H. Chao, Learning joint spatial-temporal transformations for video inpainting, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, Springer, 2020, pp. 528–543.
- [38] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, P. Pérez, Towards fast, generic video inpainting, in: Proceedings of the 10th European Conference on Visual Media Production, 2013, pp. 1–8.
- [39] C. Barnes, E. Shechtman, A. Finkelstein, D. B. Goldman, Patchmatch: A randomized correspondence algorithm for structural image editing, ACM Trans. Graph. 28 (3) (2009) 24.
- [40] D. Kim, S. Woo, J.-Y. Lee, I. S. Kweon, Deep video inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5792–5801.
- [41] R. Liu, Z. Weng, Y. Zhu, B. Li, Temporal adaptive alignment network for deep video inpainting, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 927–933.
- [42] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, M.-M. Cheng, Towards an end-to-end framework for flow-guided video inpainting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17562–17571.
- [43] J. Kang, S. W. Oh, S. J. Kim, Error compensation framework for flow-guided video inpainting, in: European Conference on Computer Vision, Springer, 2022, pp. 375–390.
- [44] K. Zhang, J. Peng, J. Fu, D. Liu, Exploiting optical flow guidance for transformer-based video inpainting, arXiv preprint arXiv:2301.10048 (2023).

- [45] K. Zhang, J. Fu, D. Liu, Flow-guided transformer for video inpainting, in: European Conference on Computer Vision, Springer, 2022, pp. 74–90.
- [46] D. Kim, S. Woo, J.-Y. Lee, I. S. Kweon, Recurrent temporal aggregation framework for deep video inpainting, *IEEE transactions on pattern analysis and machine intelligence* 42 (5) (2019) 1038–1052.
- [47] Y. Zhang, Z. Wu, Y. Yan, Pfta-net: Progressive feature alignment and temporal attention fusion networks for video inpainting, in: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, pp. 191–195.
- [48] D. Ceylan, C.-H. P. Huang, N. J. Mitra, Pix2video: Video editing using image diffusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 23206–23217.
- [49] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, A. Dittadi, Diffusion models for video prediction and infilling, arXiv preprint arXiv:2206.07696 (2022).
- [50] N. Cherel, A. Almansa, Y. Gousseau, A. Newson, Infusion: Internal diffusion for video inpainting, arXiv preprint arXiv:2311.01090 (2023).
- [51] L. Ke, Y.-W. Tai, C.-K. Tang, Occlusion-aware video object inpainting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14468–14478.
- [52] H. Wang, H. Li, B. Li, Video inpainting for largely occluded moving human, in: 2007 IEEE International Conference on Multimedia and Expo, IEEE, 2007, pp. 1719–1722.
- [53] C. Wang, X. Chen, S. Min, J. Wang, Z.-J. Zha, Structure-guided deep video inpainting, *IEEE Transactions on Circuits and Systems for Video Technology* 31 (8) (2020) 2953–2965.
- [54] S. Zhou, C. Li, K. C. Chan, C. C. Loy, Propainter: Improving propagation and transformer for video inpainting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10477–10486.

- [55] M. Granados, K. I. Kim, J. Tompkin, J. Kautz, C. Theobalt, Background inpainting for videos with dynamic objects and a free-moving camera, in: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part I 12, Springer, 2012, pp. 682–695.
- [56] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, T. Huang, Youtube-vos: Sequence-to-sequence video object segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 585–601.
- [57] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 724–732.
- [58] R. Szeto, J. J. Corso, The devil is in the details: A diagnostic evaluation benchmark for video inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21054–21063.
- [59] B. J. Matthews, C. Reichherzer, B. H. Thomas, R. T. Smith, Maskwarp: Visuo-haptic illusions in mixed reality using real-time video inpainting, in: 2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), IEEE, 2023, pp. 767–768.
- [60] T. Kim, G. J. Kim, Real-time and on-line removal of moving human figures in hand-held mobile augmented reality, *The Visual Computer* 39 (7) (2023) 2571–2582.
- [61] L. Wang, T. Tian, X. Yan, F. Ruan, G. J. Aadityaa, H. Choset, L. Li, Real-time video inpainting for rgb-d pipeline reconstruction, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2023, pp. 9543–9550.