Video Inpainting and Restoration: A Comprehensive Survey

Abstract

Video inpainting and restoration are critical processes in post-production workflows, aimed at repairing missing or corrupted video segments to improve visual quality and continuity. This survey provides a comprehensive overview of recent advances in video inpainting techniques, with particular focus on diffusion-based approaches that have revolutionized the field in recent years.

1. Introduction

Video inpainting has emerged as a fundamental challenge in computer vision and multimedia processing. The task involves filling missing regions in video sequences while maintaining temporal consistency and visual realism. Traditional approaches relied on optical flow and patch-based synthesis, but recent developments in deep learning, particularly diffusion models, have opened new possibilities for high-quality video restoration.

2. Diffusion Models for Video Inpainting

Diffusion models employed for video inpainting are typically based on a probabilistic framework, where the process involves both forward and reverse diffusion steps. The forward process gradually adds noise to the video frames, while the reverse process aims to reconstruct the original data by removing this noise [4], [32].

The mathematical foundation of these models can be expressed in terms of a forward diffusion process $q(x_t|x_{t-1})$ and a reverse process $p_\theta(x_{t-1}|x_t)$, where θ represents model parameters optimized during training.

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}\right)$$

3. Key Techniques and Innovations

3.1 Temporal Consistency

Maintaining temporal consistency across frames is crucial for video inpainting. Recent approaches employ attention mechanisms and recurrent networks to ensure that inpainted regions remain coherent throughout the video sequence.

3.2 Multi-Scale Processing

Multi-scale processing techniques have shown significant improvements in handling both small and large missing regions. By operating at multiple resolutions, these methods can capture both fine details and global structure.

4. Evaluation Metrics

The performance of video inpainting methods is typically evaluated using several metrics:

- Peak Signal-to-Noise Ratio (PSNR)
- Structural Similarity Index (SSIM)
- Temporal Consistency Score (TCS)
- User Preference Scores

5. Future Directions

Future research in video inpainting is likely to focus on:

1. Real-time processing capabilities
2. 3D-aware inpainting for volumetric video
3. Interactive inpainting tools

4. Cross-modal inpainting using audio cues

References

[1] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in CVPR, 2017, pp. 6882-6890.

[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in CVPR, 2019, pp. 5505-5514.

[3] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in ECCV, 2016, pp. 483-499.