

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. We analyze various methodologies, evaluate their performance on standard benchmarks, and discuss future research directions in this rapidly evolving domain.

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.

The importance of video inpainting extends beyond entertainment applications, finding critical uses in medical imaging, surveillance video enhancement, and historical film restoration. As video content continues to dominate digital media consumption, the demand for robust inpainting solutions has grown exponentially.

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.

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

This formulation allows for principled uncertainty quantification and enables sampling-based inference, making it particularly suitable for complex video inpainting tasks where multiple plausible solutions may exist.

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. The challenge lies in balancing spatial quality with temporal stability, as overly aggressive temporal constraints may lead to blurry or static results.

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. The pyramid-based approach allows for coarse-to-fine refinement, ensuring that both local texture and global motion patterns are preserved.

3.3 Attention-Based Approaches

Transformer-based architectures have revolutionized video inpainting by enabling long-range dependencies and adaptive feature aggregation. Self-attention mechanisms allow the model to reference distant pixels when filling missing regions, while cross-attention facilitates the integration of temporal information across frames.

4. Evaluation Metrics

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

- Peak Signal-to-Noise Ratio (PSNR): Measures pixel-level reconstruction quality
- Structural Similarity Index (SSIM): Assesses perceptual similarity
- Temporal Consistency Score (TCS): Evaluates frame-to-frame coherence
- User Preference Scores: Human evaluation of visual quality

5. Future Directions

Future research in video inpainting is likely to focus on:

1. Real-time processing capabilities for live video applications
2. 3D-aware inpainting for volumetric video and VR content
3. Interactive inpainting tools with user guidance
4. Cross-modal inpainting using audio or text cues
5. Lightweight models for mobile and edge deployment

6. Conclusion

Video inpainting has evolved from simple patch-based methods to sophisticated deep learning approaches capable of handling complex scenarios. The integration of diffusion models has pushed the boundaries of what's possible, achieving results that were unimaginable just a few years ago. As the field continues to advance, we can expect to see even more impressive developments in both quality and efficiency.

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.
- [4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in NeurIPS, 2020, pp. 6840-6851.
- [5] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in NeurIPS, 2017, pp. 5099-5108.

