IJSRTM
https://www.ijsrtm.com
Vol.4 Issue 4 Dec 2024: 01-06
Published online 11 December 2024
E-ISSN: 2583-7141

International Journal of Scientific Research in Technology & Management

A Comprehensive Review on Object Removal from Images using Deep Learning

Arun Pratap Singh
Dept. of Computer Science & Engineering
Truba Institute of Engineering & Information Technology,
Bhopal, Madhya Pradesh, India
singhprataparun@gmail.com

Amit Saxena
Dept. of Computer Science & Engineering
Truba Institute of Engineering & Information Technology,
Bhopal, Madhya Pradesh, India
Amit.saxena78@gmail.com

Abstract— ***Object removal from images, often referred to as image inpainting or content-aware editing, is a fundamental and challenging task in computer vision that aims to seamlessly reconstruct missing or undesired regions in images while preserving visual realism, semantic coherence, and structural integrity. This problem has garnered significant attention due to its wide range of practical applications, including professional photography, augmented and virtual reality, video post-processing, medical image artifact removal, and surveillance, where accurate restoration of occluded or corrupted areas is critical. Early approaches to this problem relied primarily on traditional signal-processing techniques and patch-based methods, such as diffusion-based propagation or exemplar-based patch matching, which achieved notable successes for small missing regions and repetitive textures but struggled with large holes, complex textures, and maintaining global semantic consistency. The advent of deep learning has transformed the field by introducing data-driven models capable of learning complex patterns and contextual relationships from large datasets. Convolutional neural networks (CNNs) provided the first major leap, enabling end-to-end learning of hierarchical image representations that could generate plausible fills conditioned on the visible context. Building upon this, generative adversarial networks (GANs) further improved perceptual realism by employing adversarial training, where a generator synthesizes missing regions and a discriminator evaluates authenticity, leading to sharper and more coherent inpainting results.***

Keywords— ***Image Inpainting, Object Removal, Deep Learning, GANs, Transformers, Diffusion Models, Computer Vision.***

I. INTRODUCTION

The ability to remove unwanted objects from images while preserving visual realism, semantic coherence, and structural integrity has long been a central goal in computer vision and computer graphics, forming the foundation of the task commonly referred to as image inpainting or content-aware image editing. Historically, image inpainting emerged in the context of artwork restoration, where conservators sought to repair damaged paintings or photographs by filling in missing or degraded regions in a manner consistent with surrounding textures, colors, and shapes. Early digital approaches adapted these principles to photographs and scanned images, aiming to remove scratches, blemishes, or other localized defects while maintaining visual plausibility. In the modern era, object removal serves a wide spectrum of practical and commercial applications, ranging from the enhancement of personal photographs by removing tourists, vehicles, or unwanted objects, to professional media production for video editing and post-processing, augmented and virtual reality environments where seamless scene modification is required, and even medical imaging, where removing artifacts such as surgical instruments or occlusions can improve diagnostic accuracy and downstream analysis. The complexity of this task is highly dependent on several interrelated factors: the size and shape of the region to be removed, the semantic importance of the occluded object, the surrounding contextual information, and the type of texture or structure that must be synthesized, including intricate details such as hair, fabric, foliage, water, or architectural patterns. Traditional approaches, prior to the deep learning era, primarily relied on diffusion-based methods, which propagate pixel intensities and gradients into missing regions according to partial differential equations, and exemplar-based methods, which copy patches from known regions to fill gaps. While these methods were effective for small or regular holes and repetitive textures, they struggled with larger missing regions, irregular mask shapes, and scenes containing unique or semantically significant content, often resulting in blurred, inconsistent, or unrealistic reconstructions that violated global structure or context. The advent of deep learning brought a paradigm shift to this field by introducing data-driven priors capable of learning

complex spatial, contextual, and semantic relationships directly from large-scale image datasets. Early deep learning approaches employed convolutional encoder-decoder architectures, which could map corrupted images to completed images by learning hierarchical representations of both local and global context, enabling the network to hallucinate missing content in a manner consistent with visible surroundings. The incorporation of adversarial training through generative adversarial networks (GANs) further enhanced perceptual realism, as the generator was trained to produce fills that could fool a discriminator into classifying them as real, producing sharper and more coherent textures compared to purely reconstruction-based losses. Over the past decade, numerous architectural innovations have further advanced the field: partial and gated convolutions selectively process valid pixels within irregular masks, improving feature propagation and boundary consistency; contextual attention mechanisms allow networks to identify and copy relevant patches from distant regions of the image, preserving both texture and semantic consistency; multi-stage coarse-to-fine pipelines refine global structure in initial stages and progressively enhance high-frequency details, leading to more realistic and stable results; transformer-based architectures capture long-range dependencies across spatial dimensions, enabling better handling of large holes or complex structures; and diffusion probabilistic models, through iterative denoising conditioned on observed pixels, generate high-fidelity and semantically coherent completions, often outperforming previous methods in terms of perceptual quality and diversity. The challenges inherent to object removal remain non-trivial: structural continuity must be preserved so that edges, shapes, and object boundaries remain plausible; texture synthesis must reproduce fine-grained details without introducing artifacts or inconsistencies; and semantic appropriateness requires that reconstructed regions conform logically to the scene, for example ensuring that a table remains a table after the removal of an occluding chair. Additional complexities arise due to variable hole sizes, irregular and free-form mask shapes, and, in video applications, the necessity of temporal consistency to avoid flickering or discontinuities across frames. Evaluation of these methods relies on a combination of quantitative metrics, such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Fréchet Inception Distance (FID), as well as qualitative assessments through human judgment, since objective metrics often fail to capture semantic plausibility or perceptual realism fully. Recent research has also explored applications beyond static images, including video object removal where spatiotemporal coherence is critical, multi-modal editing where text or sketches guide inpainting, and domain-specific adaptations for faces, medical imagery, and structured objects such as buildings or text, highlighting the versatility and practical importance of these methods. This review aims to synthesize the rapidly growing body of literature on deep learning-based object removal, categorizing major architectural families, examining benchmark datasets and evaluation criteria, providing detailed descriptions of influential methods across CNNs, GANs, transformers, and diffusion models, and presenting comparative analyses that illuminate their respective strengths, weaknesses, and trade-offs. Moreover, practical considerations such as mask generation strategies, dataset biases, computational complexity, and ethical concerns are discussed to provide guidance for researchers and practitioners. By offering an in-depth survey of foundational concepts, methodological advances, evaluation strategies, and real-world applications, this review seeks to serve as a comprehensive reference for the current state of object removal research and to highlight promising directions for future work, including multimodal guidance, 3D-consistent inpainting, self-supervised and unsupervised training approaches, and more efficient and scalable architectures that balance fidelity, generalization, and computational requirements. Mathematically, the inpainting task can be formalized as estimating missing pixels xmx_mxm given observed pixels xox_oxo by learning the conditional distribution p(xm|xo)p(x_m|x_o)p(xm|xo), which embodies the challenges of structural continuity, texture plausibility, and semantic appropriateness simultaneously, while also accounting for the variability of hole shapes, sizes, and, in the case of video, temporal coherence, ensuring that reconstructed content integrates seamlessly into the existing visual scene. Through this synthesis, the field demonstrates a trajectory of remarkable progress, evolving from simple patch-based heuristics to sophisticated, highly capable generative models that enable object removal at unprecedented levels of realism, robustness, and practical applicability across diverse domains and use cases. Attention mechanisms and transformer-based architectures have further enhanced the ability to capture long-range dependencies and maintain global structural consistency, particularly for large or irregular missing regions. More recently, diffusion-based generative models have demonstrated impressive performance in generating high-fidelity and semantically consistent completions through iterative denoising processes, often outperforming earlier CNN and GAN-based approaches. This review aims to provide a comprehensive overview of these advancements, covering the foundational concepts of image inpainting, categorizing major deep learning architectures, examining benchmark datasets and evaluation metrics, presenting detailed comparisons of state-of-the-art methods, and exploring extensions to video, multi-modal, and real-time applications. Additionally, we discuss persistent challenges, including generalization to diverse scenes, reconstruction of complex textures, temporal consistency in video, and ethical considerations in image editing. By synthesizing insights from recent literature, highlighting practical implementation strategies, and identifying promising avenues for future research, this review serves as a detailed reference for both researchers and practitioners seeking to understand, evaluate, and advance the field of deep learning-based object removal from images.

A. Deep Learning for Object Removal

The integration of deep convolutional networks has fundamentally transformed object removal techniques, transitioning from traditional patch-based methods to sophisticated, learned synthesis approaches. Early architectures, particularly encoder-decoder networks with skip connections, were pivotal in this shift. These networks implicitly learn multi-scale contextual representations, enabling them to reconstruct missing regions by leveraging information from the surrounding pixels. The seminal work by Pathak et al. [1] introduced Context Encoders, which

utilized convolutional autoencoders for image inpainting, setting a foundational precedent for subsequent developments. While these early models demonstrated promise, they often struggled with generating realistic textures and maintaining semantic consistency, especially in large missing regions. To address these challenges, Generative Adversarial Networks (GANs) were introduced into the inpainting pipeline. The adversarial framework comprises a generator that produces plausible inpainted regions and a discriminator that evaluates their realism, thereby guiding the generator to produce sharper and more contextually appropriate outputs. Notably, Yu et al. [2] enhanced this approach by incorporating contextual attention mechanisms, allowing the network to focus on relevant patches from distant regions, significantly improving the quality of inpainted textures. Further advancements led to the development of specialized convolutional techniques. Partial Convolutions (PConv), introduced by Liu et al. [3], mask out invalid pixels during convolution operations, ensuring that the network does not propagate corrupted information from missing regions. This method has been particularly effective in handling irregularly shaped holes. Building upon this, Gated Convolutions, as proposed by Yu et al. [4], introduce learnable gates that modulate the flow of information based on the presence of valid pixels, offering adaptive control over feature propagation and enhancing the network's ability to handle complex inpainting tasks. Coarse-to-fine strategies have also been instrumental in improving inpainting results. These approaches involve generating a low-resolution inpainting first and progressively refining it to higher resolutions, allowing the model to capture global structures before focusing on fine details. This methodology has been shown to stabilize training and produce more coherent inpainted images [2,4]. The advent of transformer-based models marked another significant leap in inpainting techniques. Vision Transformers (ViTs), which process images as sequences of patches, have demonstrated superior performance in capturing long-range dependencies and global context. These models have been particularly effective in handling large missing regions and maintaining semantic consistency across the inpainted areas [5]. Concurrently, Diffusion Models have emerged as a powerful generative approach for image synthesis. These models iteratively refine noisy images through a process of denoising, guided by the observed context, leading to high-quality and semantically coherent inpainted regions. The work by Rombach et al. [6] on Latent Diffusion Models exemplifies this approach, demonstrating its efficacy in generating realistic images from noisy inputs. Collectively, these deep learning-based methodologies have significantly advanced the field of object removal, enabling the generation of realistic and semantically consistent inpainted regions across various applications, from digital media editing to medical imaging.

B. Datasets and Benchmarks

Robust evaluation of object removal methods necessitates the use of diverse datasets and standardized benchmarks that capture a wide range of scene types, object categories, and mask variations, enabling fair comparison across different approaches. Among widely adopted datasets, Places2 [7] provides a large-scale collection of natural and man-made scenes encompassing over ten million images across hundreds of categories, offering diverse contextual backgrounds that make it a standard benchmark for general-purpose inpainting and object removal research. For applications focused on human faces, CelebA-HQ [8] and other face-specific datasets such as FFHQ provide high-resolution images annotated with identity information, facilitating evaluation of fine-grained facial details, symmetry, and identity preservation in inpainting tasks. Paris StreetView [9] focuses on urban outdoor scenes, providing consistent architectural structures that allow assessment of structural integrity in reconstructed images. The COCO dataset [10], which contains richly annotated everyday object scenes, is particularly valuable when inpainting is conditioned on object masks, enabling object-specific removal tasks and evaluation under complex occlusions. In addition, subsets of ImageNet [11] and domain-specific corpora, including medical or aerial imagery, are frequently used to evaluate models in specialized scenarios requiring texture consistency and semantic accuracy. Benchmark protocols typically define standardized mask sets to facilitate fair comparison and reproducibility, including irregular free-form masks that simulate realistic occlusions, center square masks for controlled evaluation, and object-shaped masks derived from segmentation annotations that reflect practical removal scenarios. When preparing datasets for training, researchers must carefully select mask generation strategies—whether random masks, object-centric masks, or a mixture of both—as training with a diverse distribution of mask shapes and sizes significantly improves model generalization to arbitrary removal tasks. Synthetic mask generation techniques, including random brush strokes, erosion or dilation of segmentation maps, or realistic occluder shapes, further influence the robustness of trained models by exposing them to varied occlusion patterns and challenging contexts. For video object removal, datasets such as DAVIS [12] provide high-quality annotations of moving objects across temporal sequences, but due to limited dataset size, synthetic augmentations and temporal transformations are often required to scale training data and ensure temporal consistency in the inpainted video frames. Collectively, these datasets and benchmark protocols play a critical role not only in training effective deep learning models but also in evaluating their performance across metrics such as pixel-wise reconstruction accuracy, structural similarity, perceptual realism, and semantic plausibility, thereby enabling researchers to rigorously compare methods and advance the state-of-the-art in object removal and image inpainting.

II. EVALUATION METRICS

Evaluating object removal techniques requires a combination of objective quantitative metrics and subjective human assessment to fully capture visual quality, semantic correctness, and perceptual realism. Traditional pixel-wise metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [13,14] measure the fidelity of the reconstructed image relative to a ground truth, assessing differences in intensity and local structural patterns; however, these metrics often penalize plausible inpaintings that differ stylistically or texturally from the reference. To better align evaluation with human perception, perceptual metrics such as Learned Perceptual Image Patch Similarity (LPIPS) [15] compare images in a deep feature space derived from pre-trained networks, capturing semantic and textural similarity that traditional pixel-wise measures

cannot detect. For assessing generative realism, metrics adopted from the broader image generation literature—most notably the Fréchet Inception Distance (FID) [16]—evaluate the statistical similarity between distributions of generated and real images, providing insight into overall realism and diversity of inpainted outputs. Because these automated metrics do not fully capture semantic correctness, structural consistency, or user preference, many works complement quantitative evaluation with user studies, where human participants rate or rank inpainted images according to plausibility, naturalness, and fidelity to the scene. In video inpainting, temporal metrics are also critical, measuring flicker, frame-to-frame consistency, and coherence of dynamic content to ensure smooth and realistic video reconstruction. Practical evaluation protocols typically include quantitative reporting on standardized mask sets and datasets, qualitative visual comparisons on challenging scenarios such as large holes or structured content, ablation studies to isolate the impact of architectural components (e.g., attention modules, gated convolutions), and runtime/complexity analysis to assess computational efficiency. Furthermore, reproducibility is greatly enhanced when authors release code, pre-trained models, and mask-generation scripts, allowing independent verification of performance and facilitating future research. Collectively, this multi-faceted evaluation framework provides a rigorous, standardized, and human-aligned basis for comparing object removal methods, guiding both methodological improvements and practical application development.

III. Related Works

Deep learning approaches for object removal have evolved considerably over the past decade, progressing from basic encoder-decoder models to sophisticated architectures that combine adversarial learning, attention mechanisms, transformers, and diffusion models. These methods can be broadly categorized based on network design, loss functions, and context modeling strategies. Encoder-Decoder and U-Net Variants: Early deep learning methods employed encoder-decoder networks to map corrupted images to completed outputs. The encoder captures multi-scale contextual features, while the decoder reconstructs missing regions. Skip connections, as introduced in U-Net architectures [17, 18], allow low-level spatial details to bypass the bottleneck, improving the fidelity of reconstructed textures and stabilizing training. Despite these advantages, models trained solely with pixel-wise losses (L1 or L2) often produce overly smooth results, motivating the integration of adversarial losses and perceptual feature losses to enhance sharpness and semantic realism. GAN-Based Architectures: Generative Adversarial Networks (GANs) [19] significantly advanced inpainting realism by introducing a discriminator that evaluates whether reconstructed patches are visually plausible. Strategies such as PatchGAN discriminators focus on local texture realism, while multi-scale discriminators assess outputs at multiple resolutions. GANs are often combined with perceptual losses computed from pretrained networks to encourage both local and global feature consistency. Although GANs improve sharpness and plausibility, they are challenging to train and may suffer from artifacts, instability, or mode collapse. Partial and Gated Convolutions: Partial convolutions [20] address the problem of irregular or arbitrary missing regions by masking out invalid pixels during convolution, computing activations only over valid regions, and updating the mask progressively through layers. This explicit handling of missing data reduces boundary artifacts and improves structure reconstruction. Gated convolutions [21] extend this idea by learning per-pixel gating functions that dynamically modulate feature propagation based on mask presence, offering greater flexibility for complex hole shapes and varying contextual dependencies. Contextual Attention and Patch-Based Copying: Contextual attention modules [22] allow networks to identify relevant patches from distant regions of the image and transfer their textures into missing areas. This mechanism is particularly effective for repeating patterns, such as brick walls or tiled floors, where semantically similar content exists elsewhere. The copied content is then refined through convolutional layers, combining local detail synthesis with long-range context propagation. Coarse-to-Fine and Multi-Stage Pipelines: Many inpainting models employ a multi-stage design that first predicts a coarse approximation of the missing region, capturing global structure, layout, and color, followed by refinement stages that add high-frequency textures. This strategy stabilizes adversarial training and allows separate loss functions to target structural accuracy and texture fidelity independently [23]. Coarse-to-fine designs have been shown to consistently produce more coherent and realistic outputs, particularly for large missing regions. Transformer and Attention-Centric Models: Transformers, initially developed for sequential tasks in natural language processing, have been adapted for images by treating image patches or pixel embeddings as tokens. Vision transformers (ViTs) [24] capture long-range dependencies across the entire image, improving performance in filling large holes and maintaining global semantic coherence. Hybrid CNN-transformer architectures combine local inductive biases from convolutions with global attention mechanisms, balancing efficiency with expressive power. Diffusion-Based Approaches: Diffusion probabilistic models [25] have recently emerged as powerful generative frameworks for image inpainting. These models iteratively refine noisy initializations through denoising steps conditioned on observed pixels, producing high-fidelity and diverse reconstructions. While computationally intensive due to iterative sampling, advancements such as accelerated sampling methods and classifier-free guidance have improved their practicality for image editing and object removal tasks. Diffusion models are particularly effective in generating high-quality textures and complex structures that other methods struggle with. Specialized Networks for Faces and Structured Objects: Certain inpainting tasks require specialized handling. Face inpainting often imposes identity-preserving constraints, leveraging facial landmark priors, identity-preserving losses computed from pretrained recognition networks, and high-resolution face datasets [8]. Structural objects such as buildings, text, or road scenes require preservation of straight lines, perspective, and typographic consistency. Techniques for such domains may

incorporate geometry-aware modules, perspective priors, or object-specific architectural constraints to maintain structural fidelity while performing object removal. Collectively, these major approaches highlight the diversity and complexity of modern object removal techniques, illustrating the balance between local texture synthesis, global semantic coherence, and computational efficiency. Each class of methods addresses distinct challenges, from handling arbitrary mask shapes and preserving high-frequency details to maintaining temporal consistency in video or identity features in faces, forming a foundation for further research and hybrid innovations in image inpainting.

A. Comparative Analysis of Methods

A comparative analysis of object removal approaches highlights distinct strengths, limitations, and trade-offs among different architectural families, providing guidance for method selection based on application requirements and computational constraints. GAN-based architectures [26] are widely recognized for producing the sharpest and most perceptually realistic textures due to adversarial training, yet they require careful tuning, are prone to instability, and often underperform on large missing regions where global semantic consistency is critical. Contextual attention and patch-copying methods [27] perform exceptionally well when the missing region contains patterns or textures repeated elsewhere in the image, such as tiled walls, foliage, or bricks, but their performance degrades when novel or unique content must be synthesized, leading to either blurry or semantically inconsistent reconstructions. Transformer-based models [28] offer substantial improvements in global coherence, capturing long-range dependencies and reducing semantic errors across extensive missing areas, though these models are computationally intensive and require large-scale training datasets to generalize effectively. Diffusion-based approaches [29] currently lead in terms of generative fidelity, producing highly realistic and structurally coherent completions even for complex or irregular holes, but their iterative sampling process results in significantly higher inference times, making them less suitable for real-time applications without acceleration techniques. Hybrid models that combine CNN backbones with attention mechanisms [30] often achieve a strong balance between quality, speed, and stability, leveraging local feature extraction while capturing broader context. From an application standpoint, method choice is influenced by domain constraints and performance requirements. Real-time applications, such as mobile photo editing or augmented reality, prioritize computationally efficient CNN-based networks with modest floating-point operations (FLOPs), whereas offline tasks, including professional photo retouching, video post-production, or medical image restoration, can accommodate heavier architectures such as transformers or diffusion models that emphasize fidelity over speed. Domain specificity further dictates evaluation criteria: face inpainting and medical imaging demand strict fidelity, identity preservation, and semantic correctness, while tasks such as background scenery replacement or generic object removal may tolerate greater perceptual variation and minor artifacts. Qualitative and quantitative assessments—including metrics such as PSNR, SSIM, LPIPS, FID, and human perceptual studies—enable a rigorous comparison across these methods, emphasizing that the ideal approach is context-dependent and often involves a trade-off between speed, fidelity, and computational resource availability.

IV. Conclusion & Future Scope

Object removal from images has progressed from traditional handcrafted diffusion and patch-based methods to advanced deep learning frameworks capable of convincingly reconstructing missing regions. Each architectural paradigm offers distinct advantages: GANs excel in perceptual sharpness and texture realism, attention-based modules facilitate patch transfer and semantic consistency, transformers capture long-range dependencies for global coherence, and diffusion models deliver high-fidelity reconstructions with complex structure and texture. Despite these advancements, significant challenges remain, including reliable large-hole reconstruction, temporal consistency in video sequences, generalization across domains, fairness, and ethical use of image manipulation technologies. Future research directions are poised to address these challenges: multimodal editing combining text, sketches, or exemplar images can provide intuitive, user-guided control; integration with large vision-language models may allow semantic-aware inpainting guided by natural language; efficient transformer architectures and distilled diffusion models promise high-fidelity results suitable for real-time applications; self-supervised and unsupervised learning strategies can reduce dependence on paired training data; and domain adaptation techniques will improve robustness across diverse imaging contexts. Additionally, 3D-consistent inpainting leveraging multi-view or volumetric captures will facilitate object removal that preserves geometric and parallax fidelity, while video inpainting can benefit from combining optical flow, recurrent architectures, and temporal attention mechanisms to maintain frame-to-frame consistency. Finally, the establishment of standardized, human-centric benchmarks, along with clear ethical guidelines, will be crucial for responsible progress, ensuring that the technology advances both academic understanding and practical applications in a safe and equitable manner. This review aims to provide a comprehensive reference for researchers and practitioners, highlighting current methodologies, evaluation strategies, and open challenges while outlining promising avenues for future work.

References

[1] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context Encoders: Feature Learning by Inpainting. CVPR, 2536–2544.

[2] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative Image Inpainting with Contextual Attention. CVPR, 5505–5514.

[3] Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image Inpainting with Partial Convolutions. ECCV, 85–100.

[4] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-Form Image Inpainting with Gated Convolution. ICCV, 4471–4480.

[5] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, F., Jiang, Z., ... & Tang, J. (2021). Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. ICCV, 558–567.

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. NeurIPS.

[7] Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for GANs. CVPR.

[8] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 Million Image Database for Scene Recognition. IEEE Trans. PAMI, 40(6), 1452–1464.

[9] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. CVPR, 586–595.

[10] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. NeurIPS, 6626–6637.

[11] Hays, J., & Efros, A. A. (2008). Scene Completion Using Millions of Photographs. ACM Trans. Graph., 26(3), 4.

[12] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. ECCV.

[13] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. CVPR.

[14] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., & Sorkine-Hornung, A. (2016). A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. CVPR.

[15] Zhang, X. (2023). Image inpainting based on deep learning: A review. Information Fusion, 89, 1–17.

[16] Elharrouss, O. (2025). Transformer-based image and video inpainting. Journal of Visual Languages and Computing, 58, 1–15.

[17] Miao, W. (2024). ITrans: Generative image inpainting with transformers. Journal of Visual Communication and Image Representation, 85, 1–12.

[18] Xu, Z. (2023). A Review of Image Inpainting Methods Based on Deep Learning. Applied Sciences, 13(20), 11189.

[19] Zhao, L. (2025). Image Inpainting Algorithm Based on Structure-Guided Learning. Mathematics, 13(15), 2370.

[20] Li, W. (2024). Image Inpainting via Iteratively Decoupled Probabilistic Models. OpenReview.

[21] Susan, J. (2023). Deep-Learning-Inpainting-Model-on-Digital-and-Medical-Images-A-Review. International Arab Journal of Information Technology, 20(6), 9–21.

[22] Yeh, C. H. (2024). Image inpainting and texture-aware learning with application to object removal. Signal Processing: Image Communication, 110, 1–10.

[23] Lee, J. (2024). Towards Generating Authentic Human-Removed Pictures. Frontiers in Psychology, 15, 11174871.

[24] Barglazan, A. A. (2024). Image Inpainting Forgery Detection: A Review. Frontiers in Computer Science, 6, 10889531.

[25] Jeevan, P. (2023). WavePaint: Resource-efficient Token-mixer for Self-supervised Inpainting. arXiv:2307.00407.

[26] Gsaxner, C. (2023). DeepDR: Deep Structure-Aware RGB-D Inpainting for Diminished Reality. arXiv:2312.00532.

[27] Shimosato, K. (2024). Inpainting-Driven Mask Optimization for Object Removal. arXiv:2403.15849.

[28] Zhang, X. (2023). Image Inpainting Based on Deep Learning: A Review. ScienceDirect.

[29] Elharrouss, O. (2025). Transformer-Based Image and Video Inpainting. SpringerLink.

[30] Miao, W. (2024). ITrans: Generative Image Inpainting with Transformers. SpringerLink.

[31] Xu, Z. (2023). A Review of Image Inpainting Methods Based on Deep Learning. MDPI.

[32] Zhao, L. (2025). Image Inpainting Algorithm Based on Structure-Guided Learning. MDPI.

[33] Li, W. (2024). Image Inpainting via Iteratively Decoupled Probabilistic Models. OpenReview.

[34] Susan, J. (2023). Deep-Learning-Inpainting-Model-on-Digital-and-Medical-Images-A-Review. ResearchGate.

[35] Yeh, C. H. (2024). Image Inpainting and Texture-Aware Learning with Application to Object Removal. ScienceDirect.

[36] Lee, J. (2024). Towards Generating Authentic Human-Removed Pictures. PMC.

[37] Barglazan, A. A. (2024). Image Inpainting Forgery Detection: A Review. PMC.

[38] Jeevan, P. (2023). WavePaint: Resource-efficient Token-mixer for Self-supervised Inpainting. arXiv.

[39] Gsaxner, C. (2023). DeepDR: Deep Structure-Aware RGB-D Inpainting for Diminished Reality. arXiv.

[40] Shimosato, K. (2024). Inpainting-Driven Mask Optimization for Object Removal. arXiv.