Improving Image Inpainting through Contextual Attention in Deep Learning

Ayoub Charef
LAMIGEP, EMSI Moroccan School of Engineering, Marrakesh, Morocco
a.charef@emsi.ma (corresponding author)

Ahmed Ouqour
CRSI of the School of High Economic, Commercial, and Engineering Studies (HEEC) of Marrakech, Morocco
ouqour.ahmed@eheec.ac.ma

Received: 25 March 2024 | Revised: 11 April 2024 | Accepted: 19 April 2024

 DOI: https://doi.org/10.48084/etasr.7347

ABSTRACT

Image processing is vital in modern technology, offering a diverse range of techniques for manipulating digital images to extract valuable information or enhance visual quality. Among these techniques, image inpainting stands out, involving the reconstruction or restoration of missing or damaged regions within the images. This study explores advances in image inpainting and presents a novel approach that integrates coarse-to-fine inpainting and attention-based inpainting techniques. The proposed method in this paper leverages deep learning methods to enhance the quality and efficiency of image inpainting, achieving robust and high-quality results that balance structural integrity and contextual coherence. A comprehensive evaluation and comparison with existing methods showed that the proposed approach has superior performance in maintaining structural integrity and contextual coherence within images.

Keywords-coarse-to-fine inpainting; attention-based inpainting; image processing; contextual attention

I. INTRODUCTION

Image processing plays a crucial role in modern technology, encompassing a wide array of techniques aimed at manipulating digital images to extract valuable information or enhance visual quality. Among the various applications of image processing, an area of particular interest is image inpainting, which involves the reconstruction or restoration of missing or damaged regions within an image. Inpainting techniques have garnered significant attention due to their relevance in various domains, including image restoration, object removal, and data augmentation. Inpainting aims to seamlessly fill in gaps in an image while preserving its overall structure, coherence, and visual fidelity. Previous studies have proposed several algorithms and methods to address the challenges associated with image inpainting, leveraging advances in deep learning and neural network architectures.

This study explores the field of image inpainting [1], presenting novel approaches and methods to enhance the quality and efficiency of inpainting techniques. This study also presents a novel approach by combining innovative strategies such as coarse-to-fine inpainting and attention-based inpainting to achieve robust and high-quality results that effectively balance structural integrity and contextual coherence within images [2]. Through a comprehensive evaluation and comparative analysis with the existing methods, the proposed method demonstrates superior performance in maintaining structural integrity and contextual coherence within images, contributing to the advancement of image-processing techniques and their practical applications in various domains.

II. RELATED WORK

Image inpainting has garnered significant attention in computer vision and image processing, offering solutions for tasks such as object removal and image restoration [3]. Traditional methods, often diffusion-based or match-based, struggle with large missing regions due to their limited ability to learn high-level semantic information [4]. On the contrary, deep learning-based approaches, such as the context encoder, leverage powerful representation capabilities to produce semantically plausible and sharp content over larger regions, typically through GAN-based conditional generation architectures [5]. DSNet addressed the limitations of traditional methods by dynamically distinguishing corrupted from valid regions, enhancing realism and detail in image inpainting. DSNet's utilization of innovative modules, such as Validness Migratable Convolution (VMC) and Regional Composite Normalization (RCN), enables adaptive feature selection and normalization styles, leading to superior performance over existing methods on public datasets.

DeepGIN [6] presents a deep-generative inpainting network capable of handling various masked image types. Employing SPD ResNet blocks, the MSSA mechanism, and the BP technique, DeepGIN leverages distant features for

reconstruction, exhibiting superior performance on datasets like FFHQ and Oxford Buildings, both quantitatively and qualitatively. In [2], a comprehensive review of recent advancements in image inpainting was presented, covering sequential-based, CNN-based, and GAN-based methods along with available datasets, offering insights into their performance across various image distortions. This study serves as a valuable reference for researchers, facilitates method comparison, and provides a comprehensive overview of the current state of the field. Region Normalization (RN) [7] was introduced as a region-wise spatial alternative to Feature Normalization (FN) in image inpainting networks to address mean and variance shifts caused by corrupted regions. RN enhances training by normalizing pixels based on input masks, resulting in improved quantitative and qualitative performance compared to existing methods. Furthermore, it was generalized to other inpainting networks, demonstrating consistent performance enhancements [8].

Foreground-aware image inpainting systems separate structure inference from content completion and achieve superior performance, particularly when holes overlap with foreground objects [9]. By predicting foreground contours before inpainting, these models demonstrate substantial improvements in inpainting quality, especially in challenging scenarios such as distracting object removal. Generative image inpainting systems offer solutions for handling free-form masks and guidance, leveraging gated convolutions trained on extensive unlabeled image data [10]. These systems overcome the limitations of vanilla and partial convolutions, offering dynamic feature selection across layers and facilitating effective inpainting for diverse mask shapes. PEN-Net [11], a deep generative model tailored for high-quality image inpainting, ensures both visual and semantic coherence. Incorporating a pyramid-context encoder for region affinity learning and attention transfer, along with a multiscale decoder and adversarial training loss, PEN-Net yields realistic and coherent inpainting results with superior performance across various datasets. AOT-GAN [12] is an enhanced model for high-resolution image inpainting that addresses the challenges of distorted structures and blurry textures through aggregated contextual transformation blocks and a tailored mask-prediction task for the discriminator. AOT-GAN outperformed state-of-the-art approaches on challenging benchmarks such as Places2, demonstrating promising results in practical applications such as logo removal and face editing.

PD-GAN [13] introduced probabilistic diversity for image inpainting, generating multiple diverse and visually realistic inpainting results through deep feature modulation and spatially probabilistic diversity normalization. This model achieved effective results on benchmark datasets such as CelebA-HQ, Places2, and Paris Street View, balancing realism and diversity, and enhancing the quality of inpainted images. In [14], an enhanced image inpainting method was presented, which leveraged a novel encoder and context loss function to achieve clearer and more realistic inpainting results compared to state-of-the-art algorithms across various image categories. This method demonstrated superior performance by combining a generative network with a fusion model of squeeze-and-excitation networks and a discriminative network based on the squeeze-and-excitation residual network, along with a joint context-awareness loss training method. Recent advances in data-driven image inpainting methods have revolutionized fundamental image editing tasks but struggle with high-resolution inputs that exceed 1K due to memory constraints. To address this, a Contextual Residual Aggregation (CRA) [15] mechanism enables high-frequency residual generation for missing content, facilitating sharp and detailed inpainting results up to 8K resolution while maintaining real-time performance on GPUs. In [16], a comprehensive review of deep learning-based image inpainting methods highlighted their significance in computer vision and image processing. Categorizing methods by inpainting strategies, network structures, and loss functions, this review provides insights on open-source codes, datasets, evaluation metrics, and real-world applications, offering valuable references for researchers and discussing future directions in the field. In [17], a review was carried out on inpainting methods for wireless communication applications.

Semantic inpainting that incorporates the AOT block, akin to ASPP in semantic segmentation but tailored for low-level tasks, offers a promising avenue for advancement. However, despite its novel adaptation, the method falls short of fully addressing the challenges of completing large missing regions within complex scenes. This limitation arises from the reliance on patch-based approaches prone to diffusion-related blurs and struggles with synthesizing content in regions lacking similar patches in known contexts. Moreover, while ASPP has proven efficacy in high-level recognition for semantic segmentation, its direct application in inpainting models may not sufficiently meet the specific demands of low-level inpainting tasks. As such, while this method provides valuable insight, further refinement and exploration of alternative methodologies are needed to overcome these limitations and achieve semantic coherence in completing large missing regions.

III. THE PROPOSED INPAINTING METHOD

The proposed approach uses both coarse-to-fine and attention-based inpainting techniques. Coarse-to-fine inpainting provides a systematic strategy to fill in missing regions of an image by first generating a rough approximation of the missing content and then gradually refining it to capture finer details [18]. This method enables the efficient completion of large missing areas while preserving the overall structure and coherence of the image. Additionally, by integrating attention-based inpainting mechanisms, the method aims to enhance the inpainting process by selectively focusing on relevant image features and context. Attention mechanisms allow the model to prioritize important regions during inpainting, resulting in more accurate and visually pleasing reconstructions. By combining these two approaches, this study aims to achieve robust and high-quality inpainting results that address both structural integrity and contextual coherence within image inpainting.

To further elaborate on the proposed method, it is crucial to highlight the integration and synergistic effects of the coarse-to-fine and attention-based inpainting techniques. The coarse-to-fine method starts with a broad reconstruction, laying down a foundational layer that captures the essential shapes and forms of the missing regions. This step is critical for ensuring

that the subsequent finer details are placed within an appropriate structural context, thereby maintaining the image's overall integrity. Following this, the attention-based inpainting technique comes into play, employing sophisticated algorithms to analyze the surrounding pixels and textures. This analysis enables the model to make informed decisions about the details and emphasize the missing areas, based on the context provided by the intact parts of the image. The attention mechanism acts as a discerning artist, carefully choosing which strokes to add next, ensuring that each addition harmonizes with the existing elements. This approach is particularly effective for handling complex inpainting tasks, such as those involving intricate textures or dynamic backgrounds. By intelligently directing focus and resources, it can replicate detailed and contextually appropriate content that blends seamlessly with the rest of the image.

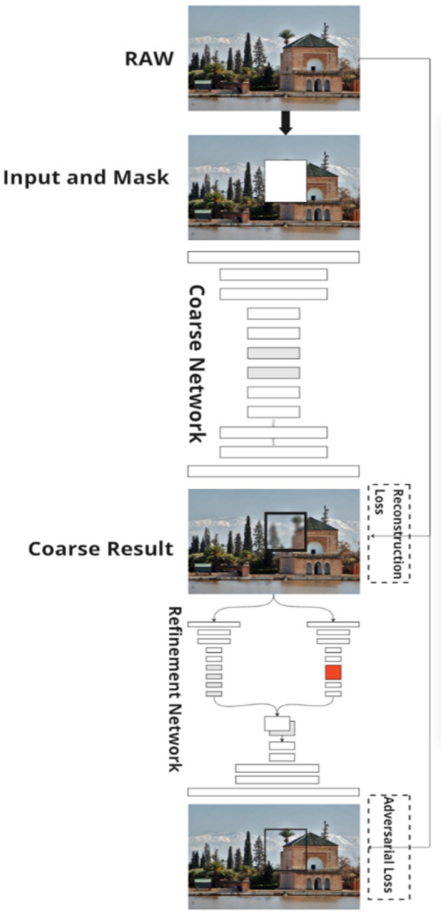

Fig. 1. Visual representation of the coarse-to-fine inpainting method.

A. *Coarse-to-Fine Inpainting*

The coarse-to-fine inpainting network architecture employs a multistage approach to effectively fill in missing regions within an image while preserving its overall structure and coherence. At the core of this architecture lies a hierarchical framework consisting of two main components: the coarse network and the refinement network. Initially, the input to the coarse-to-fine inpainting network includes the raw image and a corresponding binary mask indicating the areas to be inpainted. The raw image serves as the foundation, while the binary mask delineates the regions requiring restoration. The coarse network operates as the initial stage in the inpainting process. It takes the raw image along with the binary mask as input and generates a coarse approximation of the missing regions. This coarse output provides a foundational reconstruction, capturing the general shapes and structures of the inpainting areas.

Following the coarse inpainting stage, the output is refined through the refinement network. This network is responsible for fine-tuning the coarse result and enhancing its details, textures, and overall quality. By leveraging higher-resolution features and contextual information, the refinement network ensures that the inpainting regions seamlessly blend with the surrounding content, achieving a visually convincing reconstruction. Throughout the coarse-to-fine inpainting process, the network iteratively refines its predictions, gradually improving the inpainting output with each iteration. This hierarchical approach allows for the efficient completion of large missing areas while progressively incorporating finer details and nuances into the final reconstruction. Figure 1 illustrates the stepwise progression of the coarse-to-fine inpainting process, highlighting the input stages (raw image and binary mask), the output of the coarse network (coarse result), and the subsequent refinement performed by the refinement network. This architectural design facilitates the generation of high-quality inpainting images that seamlessly integrate with the original content, making it suitable for various applications in image restoration and enhancement.

B. *Attention-based Inpainting*

Attention-based inpainting introduces a mechanism inspired by human visual perception, where attention is selectively directed to relevant regions of an image during the inpainting process. This approach aims to prioritize important image features and context, enhancing the accuracy and realism of the inpainting results. At its core, attention-based inpainting uses attention mechanisms to dynamically assign weights to different parts of the input image, emphasizing informative regions while suppressing irrelevant ones. These mechanisms are typically integrated into Convolutional Neural Network (CNN) architectures, enabling the model to focus on specific areas during feature extraction and reconstruction. The attention mechanism can be mathematically formulated as follows:

$$\alpha_i = \frac{exp(f(h_i))}{\sum_{j=1}^{n} exp(f(h_j))} \quad (1)$$

where α_i represents the attention weight assigned to the i^{th} spatial location in the feature map, h_i denotes the feature vector at the i^{th} spatial location, $f(.)$ represents a learnable function that computes the relevance of each feature vector, and N is the total number of spatial locations in the feature map. This formula computes the attention weight α_i for each spatial location in the feature map based on its corresponding feature vector h_i. The function $f(.)$ captures the importance of each feature vector, which is then normalized across all spatial locations to ensure that the attention weights sum up to one. By

incorporating attention mechanisms into the inpainting process, the model can effectively prioritize relevant image features and context, resulting in more accurate and contextually coherent image inpainting. This attention-driven approach enhances the inpainting process by simulating human-like visual attention, leading to improved reconstruction quality and perceptual realism.

C. Contextual Attention-based Inpainting

Contextual attention-based inpainting introduces a novel approach that harnesses the power of contextual attention to enhance the inpainting process. Unlike traditional inpainting methods that focus solely on local information, contextual attention considers both local and global context, allowing the model to better understand the surrounding context of missing regions. This approach enables the inpainting model to effectively utilize contextual information from non-missing regions to guide the reconstruction of missing areas, resulting in more coherent and visually pleasing image inpainting. By dynamically adjusting the attention mechanism based on contextual cues, this method achieves improved accuracy and realism in inpainting, making it a promising technique for various image restoration tasks.

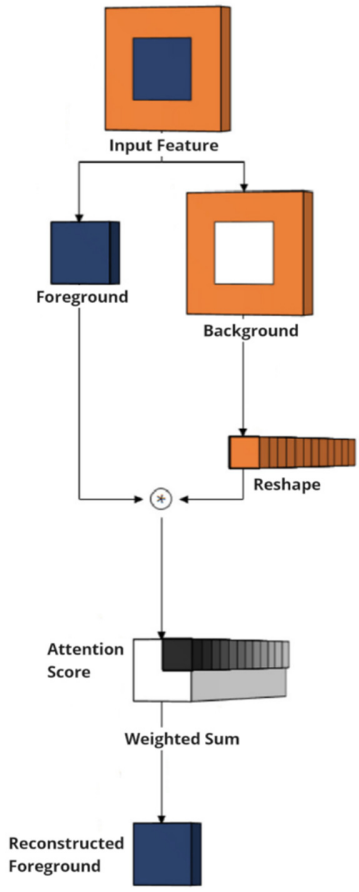

Fig. 2. The contextual attention layer.

Figure 2 shows the stepwise progression of the contextual attention-based inpainting process, highlighting the key stages involved in utilizing contextual attention to inpainting missing regions within an image.

- Input Feature: The process begins with the input features, which comprise the feature representations extracted from the input image. These features serve as the basis for the inpainting procedure, containing essential information about the image content.
- Foreground and Background Separation: The input feature undergoes foreground and background separation, where the model distinguishes between the foreground objects and the background context. This separation enables the inpainting model to focus its attention selectively on the foreground regions while considering the contextual information provided by the background.
- Reshape Operation: Following the foreground and background separation, the feature representations are reshaped to facilitate further processing. This reshaping operation prepares the feature maps for subsequent computations, ensuring compatibility with the attention mechanism.
- Attention Score Computation: The reshaped feature maps are then used to calculate the attention scores, which determine the relevance of each spatial location within the feature maps. The attention mechanism dynamically assigns weights to different regions based on their contextual significance, emphasizing informative regions while suppressing irrelevant ones.
- Weighted Sum Calculation: Once the attention scores are computed, they are applied to the reshaped feature maps to obtain a weighted sum of the feature vectors. This weighted sum aggregates information from multiple spatial locations, with higher weights assigned to regions deemed more relevant by the attention mechanism.
- Reconstructed Foreground: Finally, the weighted sum of feature vectors is used to reconstruct the foreground regions within the image. By integrating contextual attention, the model effectively incorporates information from both local and global contexts, resulting in coherent and visually appealing inpainting reconstructions.

IV. COMPARISON AND DISCUSSION

The proposed image inpainting method was evaluated against several existing methods, including GConv, EdgeConnect, RN, and DSNet. GConv harnesses graph-based CNNs to capture spatial dependencies and semantic information, providing a robust framework for inpainting tasks. EdgeConnect focuses on leveraging edge information to guide inpainting, ensuring accurate reconstruction with detailed edge features. RN employs residual learning to effectively handle complex textures and structures, contributing to high-quality inpainting results. DSNet integrates deep supervision to facilitate multiscale feature learning, enhancing inpainting performance across various resolutions. Each of these approaches offers unique insights and contributions to the field of image inpainting, addressing different aspects of the inpainting challenge, such as spatial context modeling, texture

synthesis, and detail preservation. This comparative analysis aimed to provide a comprehensive understanding of the strengths and weaknesses of these methods relative to the proposed approach, facilitate informed decision-making, and advance state-of-the-art image inpainting.

Figure 3 presents visual assessments performed on 512×512 images, comparing the performance of the proposed with other existing methods. Through these visual comparisons, the effectiveness of the proposed method can be evaluated in terms of its ability to restore missing or damaged regions in images while preserving visual quality and semantic coherence.

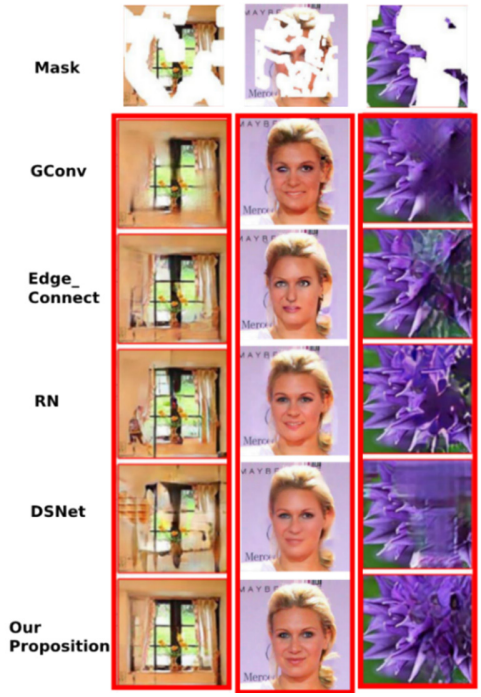

Fig. 3. Comparative visual assessment on 512×512 images.

Table I presents a comparative analysis of various metrics for different image inpainting methods. The metrics evaluated include Frechet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and L1 distance. In these metrics, a down arrow indicates that lower values mean better performance, while an up arrow indicates that higher values mean better performance.

TABLE I. QUANTITATIVE COMPARISONS BETWEEN THE PROPOSED AND OTHER EXISTING APPROACHES

Metric	Mask	GConv	EC	RN	DSNet	Proposed
FID (↓)	10-20%	0.847	0.777	0.809	0.559	0.474
	40-50%	6.384	4.244	6.488	2.842	2.721
PSNR (↑)	10-20%	31.23	31.15	33.17	32.28	33.46
	40-50%	23.37	23.57	25.02	24.60	25.15
SSIM (↑)	10-20%	0.9538	0.9470	0.9642	0.9582	0.9658
	40-50%	0.8173	0.8025	0.8466	0.8334	0.8490
L1(%) (↓)	10-20%	0.67	0.79	0.55	0.61	0.53
	40-50%	2.71	3.04	2.29	2.41	2.23

For FID, which measures the similarity between generated and real images (lower values are better), the proposed method consistently outperformed the others across different mask sizes (10-20% and 40-50%), indicating that it generated more realistic images compared to the alternatives. Regarding PSNR, which evaluates image quality based on noise levels (higher values are better), the proposed approach also demonstrates superior performance, especially at larger mask sizes (40-50%), suggesting that it produced images with higher fidelity and less noise distortion. Similarly, for SSIM, a metric that evaluates structural similarity (higher values are better), the proposed method achieved the highest scores across both mask sizes, indicating that it preserves image structures more effectively compared to the other methods. Finally, the L1 distance metric, which measures the pixel-wise difference between the inpainted and ground truth images (lower values are better), shows that the proposed method consistently achieved the lowest L1 distances, indicating superior accuracy in reconstructing missing regions. In general, the results demonstrate the effectiveness of the proposed approach in producing high-quality and visually plausible image inpainting, surpassing existing techniques across various evaluation metrics and mask sizes.

V. CONCLUSION

The proposed image-inpainting method demonstrated superior performance across various evaluation metrics compared to existing techniques, including GConv, EC, RN, and DSNet. The results of experimentation and analysis showed that the proposed method consistently produced more realistic, visually pleasing, and structurally accurate image inpainting, particularly in scenarios with larger mask sizes. The FID, PSNR, SSIM, and L1 distance metrics confirmed the effectiveness of the proposed approach in generating high-quality image inpaintings with minimal distortion and noise. These findings highlight the potential of the proposed method to significantly enhance image inpainting, offering promising solutions for a wide range of applications in computer vision, image editing, and beyond. Further research and development in this direction could lead to even more advanced and robust inpainting techniques, opening up new possibilities for creative image manipulation and restoration.

Building on this foundation, the untapped potential in specific domains, such as manga translation and the application of inpainting to integrate textual translations directly onto images, should be recognized. This area represents an exciting research area, as it blends the challenges of linguistic accuracy with visual artistry. Addressing these challenges not only expands the scope of image inpainting but also enriches the interaction between text and imagery, opening up innovative pathways for storytelling and content creation. As moving forward, exploring these possibilities will form a key aspect of future work, aiming to seamlessly bridge the gap between visual and textual content.

REFERENCES

[1] Z. Qin, Q. Zeng, Y. Zong, and F. Xu, "Image inpainting based on deep learning: A review," *Displays*, vol. 69, Sep. 2021, Art. no. 102028, https://doi.org/10.1016/j.displa.2021.102028.

[2] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image Inpainting: A Review," *Neural Processing Letters*, vol. 51, no. 2, pp. 2007–2028, Apr. 2020, https://doi.org/10.1007/s11063-019-10163-0.

[3] R. H. Mwawado, B. J. Maiseli, and M. A. Dida, "Robust Edge Detection Method for Segmentation of Diabetic Foot Ulcer Images," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 6034–6040, Aug. 2020, https://doi.org/10.48084/etasr.3495.

[4] A. Pajot, E. de Bezenac, and P. Gallinari, "Unsupervised Adversarial Image Inpainting." arXiv, Dec. 18, 2019, https://doi.org/10.48550/arXiv.1912.12164.

[5] L. Zhao *et al.*, "UCTGAN: Diverse Image Inpainting Based on Unsupervised Cross-Space Translation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5740–5749, https://doi.org/10.1109/CVPR42600.2020.00578.

[6] C. T. Li, W. C. Siu, Z. S. Liu, L. W. Wang, and D. P. K. Lun, "DeepGIN: Deep Generative Inpainting Network for Extreme Image Inpainting," in *Computer Vision – ECCV 2020 Workshops*, Glasgow, UK, Aug. 2020, pp. 5–22, https://doi.org/10.1007/978-3-030-66823-5_1.

[7] T. Yu *et al.*, "Region Normalization for Image Inpainting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12733–12740, Apr. 2020, https://doi.org/10.1609/aaai.v34i07.6967.

[8] A. Charef, Z. Jarir, and M. Quafafou, "Smart System for Emergency Traffic Recommendations: Urban Ambulance Mobility," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, pp. 32–45, 2022.

[9] W. Xiong *et al.*, "Foreground-Aware Image Inpainting," presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 5833–5841, https://doi.org/10.1109/CVPR.2019.00599.

[10] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-Form Image Inpainting With Gated Convolution," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 4470–4479, https://doi.org/10.1109/ICCV.2019.00457.

[11] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1486–1494, https://doi.org/10.1109/CVPR.2019.00158.

[12] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Aggregated Contextual Transformations for High-Resolution Image Inpainting," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 7, pp. 3266–3280, Mar. 2023, https://doi.org/10.1109/TVCG.2022.3156949.

[13] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "PD-GAN: Probabilistic Diverse GAN for Image Inpainting," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 9367–9376, https://doi.org/10.1109/CVPR46437.2021.00925.

[14] Y. Chen *et al.*, "The improved image inpainting algorithm via encoder and similarity constraint," *The Visual Computer*, vol. 37, no. 7, pp. 1691–1705, Jul. 2021, https://doi.org/10.1007/s00371-020-01932-3.

[15] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 7505–7514, https://doi.org/10.1109/CVPR42600.2020.00753.

[16] D. J. B. Rojas, B. J. T. Fernandes, and S. M. M. Fernandes, "A Review on Image Inpainting Techniques and Datasets," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Porto de Galinhas, Brazil, Nov. 2020, pp. 240–247, https://doi.org/10.1109/SIBGRAPI51738.2020.00040.

[17] V. Yatnalli, B. G. Shivaleelavathi, and K. L. Sudha, "Review of Inpainting Algorithms for Wireless Communication Application," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5790–5795, Jun. 2020, https://doi.org/10.48084/etasr.3547.

[18] J. Li, F. He, L. Zhang, B. Du, and D. Tao, "Progressive Reconstruction of Visual Structure for Image Inpainting," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 5961–5970, https://doi.org/10.1109/ICCV.2019.00606.