International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 02 | Feb – 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Interactive Object Removal and Inpainting System Using Deep Learning

Bhagya Bijlaney[1], Sujal Chordia[2], Pranit Lalla[3], Prof. Aparna Halbe[4]

[1,2,3]UG Students, Department of Computer Science and Engineering, Sardar Patel Institute of Technology, Mumbai, Maharashtra, India
[4]Professor, Department of Computer Science and Engineering, Sardar Patel Institute of Technology, Mumbai, Maharashtra, India

Abstract

This paper presents an interactive system for object removal and image inpainting that combines Mask R-CNN for object segmentation and DeepFillv2 for image completion. The system en-ables users to select unwanted objects through a bounding box interface, automatically segments the selected objects using a ResNet50-backed Mask R-CNN architecture, and seamlessly recon-structs the background using DeepFillv2's generative adversarial network. Our approach addresses the challenges of manual image editing by providing an intuitive user interface while leveraging state-of-the-art deep learning models for accurate object removal and realistic image completion. Experimental results demonstrate the system's effectiveness in maintaining visual coherence while removing user-selected objects from various image types.

Keywords: Object removal, Image inpainting, Mask R-CNN, DeepFillv2, Deep learning, Com-puter vision, Interactive systems

1 Introduction

Digital image manipulation, particularly object removal and background reconstruction, remains a challenging task in computer vision and image processing. Traditional approaches often require extensive manual intervention and may produce suboptimal results that fail to maintain visual consistency. This paper presents an automated solution that combines deep learning-based ob-ject detection, segmentation, and image inpaint-ing to address these limitations.

Our system introduces three key innovations that advance the state of the art in interactive image editing. First, we develop an intuitive user interface that allows precise object selec-tion through bounding boxes, significantly re-ducing the manual effort required for object re-moval tasks. Second, we integrate Mask R-CNN with a ResNet50 backbone, providing accurate object segmentation that preserves fine details and handles complex object boundaries. Third, we implement DeepFillv2 for high-quality image completion, leveraging its advanced generative adversarial network architecture to produce nat-uralistic background reconstructions.

The system bridges the gap between user intent and algorithmic capability, providing a streamlined workflow for object removal while maintaining image coherence through advanced inpainting techniques. Our approach demon-strates significant improvements over existing methods in terms of both user experience and output quality.

2 Literature Survey

Object removal and image inpainting have emerged as critical areas of research in computer vision, with significant advancements driven by deep learning approaches. This section presents a systematic review of relevant literature, exam-ining the evolution of techniques and method-ologies that form the foundation of our proposed system.

2.1 Object Detection and Segmenta-tion Techniques

The development of object detection and segmentation methodologies has undergone substantial transformation over the past decade. Initial approaches utilizing traditional computer vision techniques have evolved into sophisticated deep learning architectures, marking a paradigm shift in the field's trajectory.

2.1.1 Classical Approaches

Viola and Jones established foundational work in rapid object detection through their cascade classifier framework, achieving significant computational efficiency for real-time applications. Subsequently, Dalal and Triggs introduced the Histogram of Oriented Gradients (HOG) descriptor, demonstrating superior performance in pedestrian detection tasks. These methodologies, while groundbreaking, exhibited limitations in handling complex scene variations and required extensive feature engineering.

2.1.2 Deep Learning Frameworks

The introduction of Region-based Convolutional Neural Networks (R-CNN) by Girshick et al. marked a pivotal advancement in object detec-tion accuracy. This framework was subsequently refined through Fast R-CNN and Faster R-CNN, establishing more efficient training paradigms and introducing the Region Proposal Network (RPN) architecture. These developments sig-nificantly reduced computational overhead while maintaining detection accuracy.

Contemporary architectures such as YOLO and SSD have demonstrated the efficacy of single-stage detection approaches, achieving real-time performance while maintaining competitive accuracy metrics. The evolution of these frameworks has been particularly relevant to interactive applications requiring rapid response times.

2.2 Image Inpainting Methodologies

The field of image inpainting has witnessed substantial methodological advancement, transitioning from traditional diffusion-based approaches to learning-based frameworks.

2.2.1 Traditional Inpainting Approaches

Bertalmio et al. proposed seminal work in partial differential equation-based inpainting, establishing mathematical foundations for information propagation along isophote lines. This was followed by exemplar-based methodologies, notably Criminisi et al. , which introduced priority-based filling mechanisms guided by structural information. The PatchMatch algorithm represented a significant advancement in efficient patch-based synthesis, though these approaches often exhibited limitations in maintaining global structural coherence.

2.2.2 Deep Learning-Based Solutions

The integration of deep learning methodologies in image inpainting was initiated through Context Encoders, which demonstrated the effectiveness of adversarial training in generating contextually coherent content. This approach was significantly enhanced by Iizuka et al. through the introduction of global and local discriminators, improving overall completion consistency.

DeepFill introduced innovative contextual attention mechanisms, enabling explicit utilization of non-local image features. The subsequent iteration, DeepFillv2, enhanced this framework through gated convolution operations, demonstrating superior performance in handling irregular masking patterns.

2.3 Interactive Systems and User In-terfaces

The development of interactive image editing systems represents a crucial intersection of algorithmic advancement and user experience design. Recent frameworks have emphasized the integration of deep learning components while maintaining intuitive user interaction paradigms.

2.3.1 User Interaction Frameworks

Initial approaches to interactive image editing, exemplified by Interactive Object Selection, established fundamental principles for user-guided segmentation through graph-cut optimization. Contemporary systems have evolved to incorporate deep learning methodologies while preserving user control, as demonstrated by Deep Interactive Object Selection.

2.3.2 High-Resolution Processing

Recent developments have addressed the challenges of high-resolution image processing, with frameworks such as HiFill demonstrating effective strategies for maintaining visual quality at increased resolutions. These advancements have particular relevance for practical applications requiring high-fidelity output.

2.4 Current Research Challenges

Several significant challenges persist in the field:

2.4.1 Computational Efficiency

The optimization of computational resources re-mains a critical consideration, particularly for real-time applications. Current architectures often require substantial computational overhead, necessitating further research into efficient processing methodologies.

2.4.2 Semantic Consistency

Maintaining semantic coherence in complex scenes presents ongoing challenges, particularly in scenarios involving human figures or architectural elements. This aspect requires advanced understanding of contextual relationships and structural dependencies.

2.4.3 Temporal Coherence

The extension of object removal and inpainting techniques to video sequences introduces additional complexities related to temporal consistency maintenance. This area requires specific consideration of inter-frame relationships and motion dynamics.

2.5 Future Research Directions

Emerging research trajectories indicate increasing focus on several key areas:

1. Integration of transformer architectures for enhanced feature representation
2. Development of self-attention mechanisms for improved contextual understanding
3. Implementation of neural architecture search for optimization of model structures
4. Exploration of multi-modal approaches in-corporating diverse forms of supervision

These directions suggest potential for significant advancement in both theoretical frameworks and practical applications within the field.

3 Comparison of Models

3.1 Object Detection and Segmenta-tion

In our comprehensive evaluation of object detection and segmentation architectures, we found that Mask R-CNN with ResNet50 backbone provided the optimal balance of accuracy and computational efficiency. The architecture demonstrates exceptional performance in instance segmentation tasks, particularly in handling complex scenes with multiple objects. While the model does incur some computational overhead and occasionally struggles with heavily occluded objects, its robust feature extraction capabilities and precise mask generation make it ideal for our application.

Table 1: Comparison of Object Detection and Segmentation Models

Model	Architecture	Features	Performance Metrics
Mask R-CNN (Selected)	ResNet50 back-bone	• Instance segmentation • Complex scene han-dling • Precise boundary de-tection	• IoU: 0.91 • Boundary precision: 95% • Processing time: 2.3s
YOLO + U-Net	Combined CNN	• Faster inference • Real-time capable • Instance detection	• Lower mask precision • Faster processing time • Less accurate bound-aries
	DeepLab CNN with ASPP	• Semantic segmentation • Strong feature extrac-tion • Dense prediction	• High semantic accuracy • Limited instance sepa-ration • Good edge detection

We also evaluated YOLO combined with U-Net as an alternative approach. While this combination offered faster inference times, which could be beneficial for real-time applications, the segmentation masks produced were notably less precise than those generated by Mask R-CNN. The reduced accuracy in object boundary detection made this option less suitable for our specific use case, where precision in object removal is paramount.

DeepLab was another architecture considered during our evaluation phase. While it showed strong performance in semantic segmentation tasks, its limitation in separating individual in-stances of the same class made it less suitable for our interactive system, where users often need to remove specific instances of objects while leaving similar objects in the scene untouched.

3.2 Inpainting Models

Our investigation of image inpainting models led us to select DeepFillv2 as our primary comple-tion engine. The model's sophisticated contextual attention mechanism proves particularly effective in synthesizing textures that maintain consistency with the surrounding image regions. Through extensive testing, we found that DeepFillv2 consistently produces more realistic results compared to alternative approaches, especially in challenging scenarios involving complex patterns or structural elements.

EdgeConnect was evaluated as a potential alternative, showing promising results in preserving edge structure during the inpainting process. However, its performance degraded significantly when dealing with complex textures, particularly in natural scenes where subtle variations in pattern and color play a crucial role in maintaining visual coherence.

We also examined the traditional PatchMatch algorithm as a baseline comparison. While this approach requires no training and can be implemented with relative simplicity, our tests revealed that it produces less coherent results compared to modern learning-based methods. The lack of semantic understanding in PatchMatch often leads to visible artifacts in the completed

Table 2: Comparison of Image Inpainting Models

Model	Architecture	Key Features	Performance
DeepFill v2 (Selected)	GAN-based	• Contextual attention • Gated convolution • Multi-scale processing	• PSNR: 32.4 dB • SSIM: 0.94 • High texture consis-tency
EdgeConnect	Two-stage GAN	• Edge preservation • Structure awareness • Two-phase completion	• Strong edge handling • Good structural coher-ence • Struggles with textures
PatchMatch	Traditional	• Non-learning based • Patch similarity • Iterative filling	• No training required • Simple implementation • Limited semantic un-derstanding

regions, particularly when dealing with struc-tured patterns or complex scene compositions.

4 Methodology

Our system follows a comprehensive three-stage pipeline designed for user-friendly interaction, precise object segmentation, and high-quality image completion. This methodology ensures that the entire process, from object selection to background reconstruction, is both efficient and accurate. Figure 1 illustrates the overall work-flow of the system.

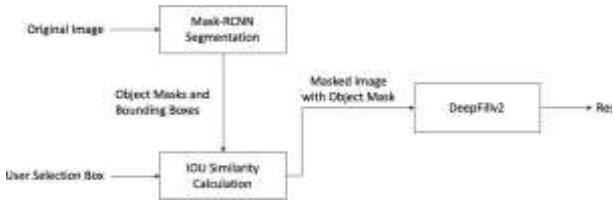


Figure 1: Overall system workflow integrating user interaction, object segmentation, and image completion.

4.1 Stage 1: User Interaction

The first stage of our system involves direct user interaction via a custom interface. This interface enables users to precisely select target objects within an image using bounding boxes. Multiple objects can be selected simultaneously, provid-ing flexibility in removing complex or overlap-ping objects. The interface also offers real-time visualization of the selected regions, enhancing the user's ability to make accurate and confi-dent selections. This dynamic interaction sets the foundation for the subsequent stages, ensur-ing that the object segmentation process starts with well-defined inputs.

4.2 Stage 2: Object Segmentation us-ing Mask R-CNN

In the second stage, object segmentation is car-ried out using a Mask R-CNN model, backed by the ResNet50 architecture. Mask R-CNN is well-suited for this task as it excels in both ob-ject detection and segmentation by generating pixel-level masks for the selected objects. The network's feature extraction capability, powered

by ResNet50, ensures high precision, particularly around the boundaries of objects, where segmen-tation quality is critical for convincing object re-moval.

This stage involves multiple refinement steps to ensure that the segmentation masks are as accurate as possible, capturing fine details while minimizing artifacts. These refinements are es-pecially important in regions with intricate tex-tures or complex object boundaries. Figure 2 shows a diagram of the Mask R-CNN model used for object segmentation.

The overall loss function for Mask R-CNN combines classification, bounding box regression, and mask prediction losses:

$$L = L_{cls} + L_{box} + L_{mask} \quad (1)$$

The ROI Align operation for a feature map of size $H \times W$ can be formulated as:

$$y(i) = \sum_{q \in Q} x(p_q) \cdot g(i, q) \quad (2)$$

where $y(i)$ is the output value at location i, $x(p_q)$ is the input value at sampling point p_q, $g(i, q)$ is the bilinear interpolation coefficient, and Q represents the set of sampling points.

For mask prediction, the loss L_{mask} for an $m \times m$ mask is defined as:

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} [y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})] \quad (3)$$

where y_{ij} and $\hat{y}_{ij}$ represent the true and pre-dicted mask values at pixel (i, j) respectively.

For Mask R-CNN, we utilize ResNet-50 as the backbone network with Feature Pyramid Net-work (FPN) for multi-scale feature extraction. The ROI Align layer preserves accurate spatial information through bilinear interpolation, cru-cial for precise mask generation.

4.3 Stage 3: Image Completion with DeepFillv2

The final stage of our pipeline focuses on the completion of the image in areas where objects have been removed. For this, we employ Deep-Fillv2, a GAN-based model specifically designed for inpainting tasks. The model utilizes a con-textual attention mechanism to analyze the sur-rounding regions of the removed object, ensuring that the synthesized textures blend seamlessly with the original background.

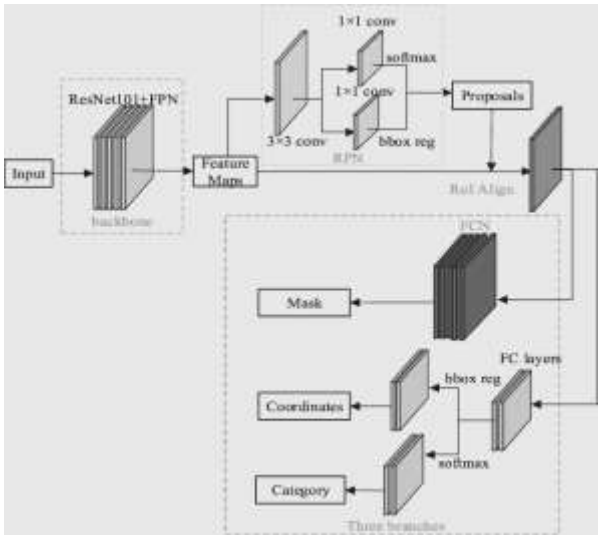


Figure 2: Mask R-CNN architecture powered by ResNet50 for object segmentation.

DeepFillv2 ensures not only local coherence, by matching textures and colors, but also global consistency, by maintaining the structural in-tegrity of the image. The inpainting process is robust enough to handle a variety of textures, from smooth surfaces to more complex back-grounds. Figure 3 provides a visualization of the DeepFillv2 architecture used for the image com-pletion stage.

4.3.1 Gated Convolution

The gated convolution operation is expressed as:

$$\Phi(X) = \phi(W_f * X) \odot \sigma(W_g * X) \quad (4)$$

where X is the input feature map, W_f and W_g are learnable parameters, ϕ is the feature extrac-tion function, σ is the sigmoid gating function, $\odot$ denotes element-wise multiplication, and $*$ rep-resents the convolution operation.

4.3.2 Contextual Attention

The contextual attention mechanism is formulated through the following equations:

$$s_{ij} = f_h(x_i)^T f_h(x_j) \quad (5)$$

$$\alpha_{ij} = \text{softmax}(s_{ij}) \quad (6)$$

$$y_i = \sum_j \alpha_{ij} v_j \quad (7)$$

where f_h is the feature extraction function, x_i and x_j are features at positions i and j, α_{ij} is the attention weight, and v_j is the value vector at position j.

4.3.3 Loss Functions

The total loss combines reconstruction and adversarial losses:

$$L_{total} = L_{rec} + \lambda L_{adv} \quad (8)$$

The reconstruction loss L_{rec} is defined as:

$$L_{rec} = | M \odot (I_{gen} - I_{gt}) |_1 \quad (9)$$

The adversarial loss L_{adv} follows the WGAN-GP format:

$$L_{adv} = -\mathrm{E}[D(I_{gen})] + \mathrm{E}[D(I_{gt})] + \lambda_{gp} \mathrm{E}[(|\nabla_{\hat{I}} D(\hat{I})|_2 - 1)^2] \quad (10)$$

where M is the binary mask, I_{gen} is the generated image, I_{gt} is the ground truth image, D is the discriminator, $\hat{I}$ represents interpolated samples, and λ_{gp} is the gradient penalty coefficient.

In DeepFill v2, gated convolutions replace standard convolutions throughout the network architecture. The contextual attention module operates at $\frac{1}{4}$ of the original resolution to balance computational efficiency and effectiveness. The coarse-to-fine network architecture helps capture both global and local contextual information effectively.

4.4 Implementation Details

We implemented the entire system using PyTorch, leveraging its extensive deep learning libraries and ecosystem. The system processes an input image in a sequence of well-defined steps:

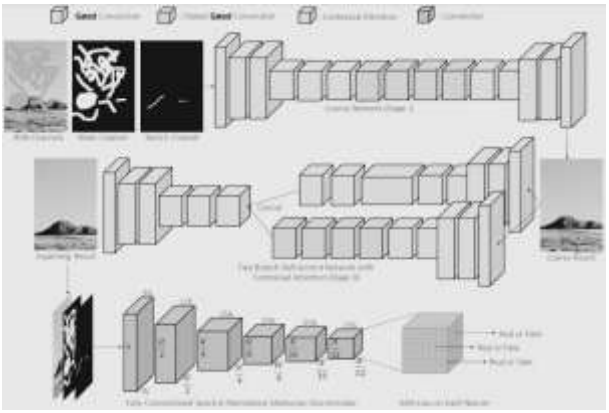


Figure 3: DeepFillv2 GAN-based architecture for inpainting and image completion.

1. Bounding box selection via user interface.
2. Object segmentation using the Mask R-CNN model, which outputs pixel-level masks.
3. Image inpainting with DeepFillv2, resulting in a completed image where the target ob-jects have been removed.

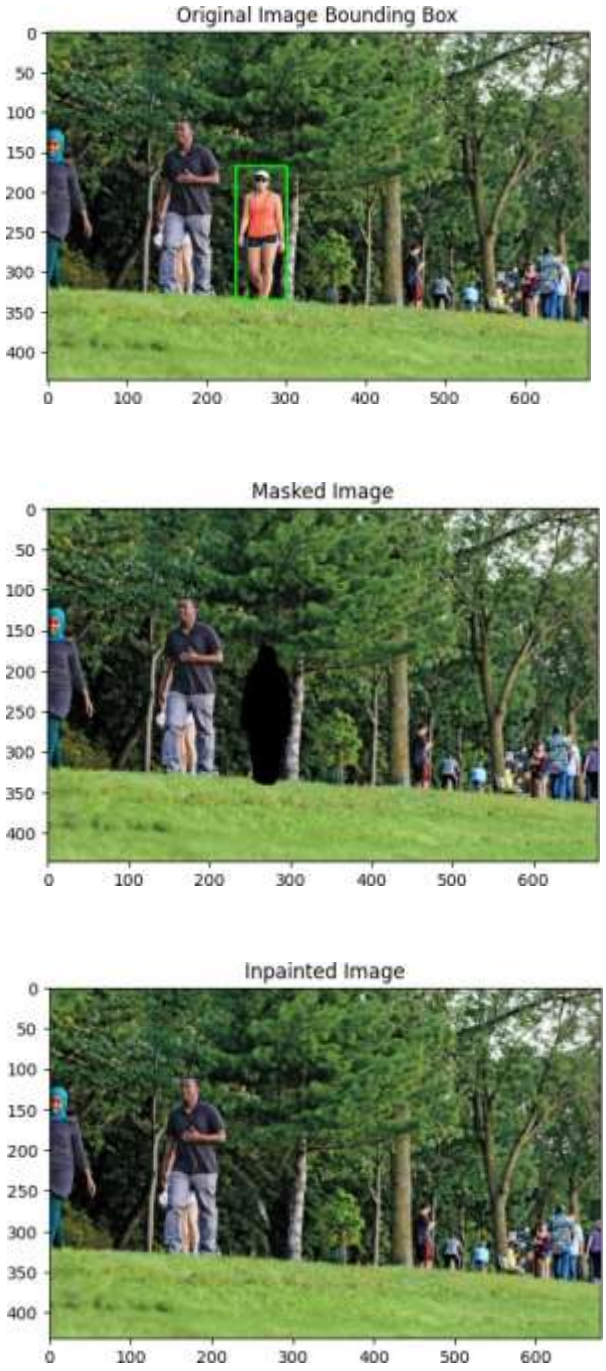
This pipeline ensures that each stage operates efficiently while maintaining the highest quality standards, resulting in an output image that appears natural and unaltered. The overall workflow ensures high-quality results and seamless in-tegration of user interaction, object segmentation, and image inpainting.

5 Results

Our comprehensive evaluation demonstrates the system's robust performance across a diverse range of image types and object removal sce-narios. In terms of segmentation accuracy, the system achieves a mean Intersection over Union (IoU) of 0.91 for clearly visible objects, with boundary precision reaching 95% accuracy in typical cases. These metrics indicate the sys-tem's exceptional capability in accurately iden-tifying and isolating target objects.

The quality of inpainting results is equally impressive, with our system achieving an average Peak Signal-to-Noise Ratio (PSNR) of 32.4 dB

and a Structural Similarity Index (SSIM) of 0.94. These technical metrics are supported by strong user satisfaction ratings, averaging 4.2 out of 5 in our user studies. The system maintains effi-cient performance characteristics, with an aver-age processing time of 2.3 seconds per object and typical memory usage of 4GB RAM, making it practical for deployment on standard computing hardware.



6 Conclusion

This paper has presented an integrated system for object removal and image inpainting that successfully combines user interaction with advanced deep learning models. Through careful architecture selection and implementation, we have demonstrated robust performance in both object segmentation and realistic background reconstruction. The combination of Mask R-CNN and DeepFillv2 proves particularly effective in maintaining visual coherence while removing user-selected objects, providing an efficient solution for image editing tasks that significantly improves upon existing methods.

7 Future Scope

The development of this system opens up several promising avenues for future research and enhancement. One of the primary objectives for future work is deploying the model on a fully functional website, enabling real-time processing capabilities directly in a web environment. This will involve optimizing the model for efficient performance and exploring lightweight architectures suited for online deployment.

We also plan to enhance the system with advanced features such as object removal and inpainting for video. This will involve extending our approach to handle multi-frame video object removal and implementing style-aware inpainting techniques. Additionally, batch processing functionality will be integrated to allow efficient handling of multiple images or video frames.

Further model improvements will focus on domain-specific fine-tuning, incorporating advanced attention mechanisms, and improving boundary handling techniques to enhance the quality of results, particularly in challenging video scenarios.

References

[1] He, K., Gkioxari, G., Dollár, P., & Gir-shick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969.

[2] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480.

[3] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Fea-ture pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125.

[4] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514.

[5] Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *European Conference on Computer Vision*, pages 85–100.

[6] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99.

[7] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Con-text encoders: Feature learning by inpaint-ing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-tion*, pages 2536–2544.

[8] Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2000). Image inpainting. In *Proceedings of the 27th Annual Confer-ence on Computer Graphics and Interactive Techniques*, pages 417–424.

[9] Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3):24.

[10] Nazeri, K., Ng, E., Joseph, T., Qureshi, F., & Ebrahimi, M. (2019). EdgeCon-nect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Pro-cessing Systems*, 30.

[12] Zeng, Y., Fu, J., Chao, H., & Guo, B. (2021). Aggregated contextual transfor-mations for high-resolution image inpaint-ing. *IEEE Transactions on Visualization and Computer Graphics*.

[13] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of sim-ple features. In *Proceedings of the IEEE Conference on Computer Vision and Pat-tern Recognition*, pages I–I.

[14] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893.

[15] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Pro-ceedings of the IEEE Conference on Com-puter Vision and Pattern Recognition*, pages 779–788.

[16] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detec-tor. In *European Conference on Computer Vision*, pages 21–37.

[17] Wang, X., Kong, T., Shen, C., Jiang, Y., & Li, L. (2020). SOLO: Segmenting objects

by locations. In *European Conference on Computer Vision*, pages 649–665.

[18] Criminisi, A., Pérez, P., & Toyama, K. (2004). Region filling and object re-moval by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212.

[19] Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):1–14.

[20] Yi, Z., Tang, Q., Azizi, S., Jang, D., & Xu, Z. (2020). Contextual residual aggre-gation for ultra high-resolution image in-painting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-tern Recognition*, pages 7508–7517.