# IMPBA: Image Mask Prediction and Background-Aware Inpainting for Object Removal

Baha Kırbaşoğlu
*Graduate Program in Computer Engineering*
*Istanbul Technical University*
Istanbul, Turkey
kirbasoglu25@itu.edu.tr

***Abstract*—Image inpainting is a fundamental problem in computer vision, aiming to restore missing or corrupted regions of an image in a visually coherent and semantically meaningful manner. This study focuses on the challenging task of object removal, where the goal is to seamlessly reconstruct occluded regions while preserving surrounding structural and perceptual information. To address this problem, two autoencoder-based deep learning architectures were implemented and systematically compared: a conventional U-Net-like model and an enhanced U-Net architecture augmented with attention mechanisms.**

**The models were trained on a large-scale subset of the DeFacto image and face manipulation dataset, consisting of approximately 25,000 object-removal samples, providing diverse visual contexts and manipulation scenarios. To improve reconstruction quality and perceptual realism, a combination of pixel-wise loss functions and perceptual loss functions was employed during the training process.**

**Model performance was evaluated using widely adopted quantitative metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), in addition to qualitative visual comparisons. Experimental results indicate that attention-based architectures consistently outperform the baseline U-Net model, particularly in preserving structural integrity and perceptual consistency in complex inpainting scenarios. Furthermore, the integration of perceptual loss functions contributes to more visually realistic reconstructions. Overall, the proposed approaches achieve superior performance compared to baseline models, demonstrating the effectiveness of attention mechanisms and perceptual learning for object removal-based image inpainting tasks.**

***Index Terms*—Image inpainting, object removal, U-Net, attention mechanisms, perceptual loss**

## I. Introduction

Image inpainting is a fundamental and long-standing problem in computer vision, which aims to reconstruct missing, damaged, or occluded regions in images by exploiting the contextual information from surrounding areas. The primary objective of image inpainting is to generate visually plausible results that are indistinguishable from the original content. This task is inherently challenging, as successful inpainting requires the preservation of semantic consistency, global structural integrity, and local texture continuity within the reconstructed regions. Traditional inpainting methods based on diffusion or patch-based techniques often struggle when handling large missing regions or complex semantic structures.

With the rapid advancement of deep learning, data-driven approaches have significantly improved the performance of image inpainting systems. Convolutional neural networks (CNNs), particularly encoder–decoder architectures, have demonstrated remarkable success in learning complex image representations and generating realistic reconstructions. Despite these advancements, producing high-quality inpainted results for complex scenes with diverse textures and structures remains a challenging problem. Models must effectively capture both low-level visual details and high-level semantic information to ensure seamless integration of reconstructed regions with the surrounding context. However, many recent approaches rely on computationally expensive architectures that limit their practical applicability.

Object removal image inpainting represents a specialized and more challenging subtask of image inpainting, where the goal is to remove undesired objects from an image while maintaining visual coherence and realism. Unlike generic missing-region reconstruction, object removal requires the model to infer meaningful content that aligns with the global scene structure after eliminating salient foreground elements. This demands a strong understanding of both local textures and long-range dependencies within the image, making naive reconstruction strategies insufficient.

From an application perspective, object removal image inpainting has gained significant importance across a wide range of domains, including photography, digital image editing, content creation, and image forensics. In practical scenarios, such as photo retouching or visual content manipulation, users expect inpainted results to be visually seamless and semantically accurate. Moreover, robust inpainting techniques play a critical role in digital forensics by enabling the detection and analysis of manipulated image content, thereby contributing to the integrity and reliability of visual media.

In this study, deep learning–based autoencoder architectures are employed to address the object removal image inpainting problem. Specifically, a conventional U-Net-like model and an enhanced U-Net architecture incorporating attention mechanisms are implemented and evaluated. Attention mechanisms are introduced to improve the model's ability to focus on relevant contextual regions and capture long-range dependencies during reconstruction. The primary objective of this work is to systematically analyze the impact of attention mechanisms and

different loss functions on inpainting performance, with an emphasis on achieving visually coherent, structurally consistent, and perceptually realistic image reconstructions. The proposed approach aims to achieve this while maintaining architectural simplicity and training efficiency.

## II. Related Work

Recent advances in image inpainting have increasingly focused on modeling long-range dependencies to handle large missing regions effectively. Traditional convolutional neural network (CNN)-based approaches often struggle to capture global contextual information, leading to visually inconsistent results. To address this limitation, transformer-based architectures have been introduced into image inpainting tasks.

Li et al. [1] proposed the Mask-Aware Transformer (MAT), which combines convolutional feature extraction with transformer-based attention mechanisms to efficiently model global context for large hole image inpainting. By leveraging customized transformer blocks and contextual attention, MAT demonstrates state-of-the-art performance on high-resolution datasets such as Places and CelebA-HQ. This work highlights the effectiveness of attention mechanisms in capturing long-range dependencies for image completion.

Another line of research explores adversarial learning frameworks to enhance perceptual quality. Zuo et al. [2] introduced a generative image inpainting approach that integrates segmentation confusion adversarial training with contrastive learning. Their method improves structural consistency and visual realism by providing pixel-level supervision and robust feature-level guidance during training.

While transformer-based and adversarial approaches achieve strong performance, they often incur high computational costs and complex training procedures. In contrast, this study focuses on convolutional autoencoder-based architectures, investigating the impact of attention mechanisms and loss functions on object removal image inpainting. By evaluating standard U-Net and attention-enhanced U-Net models, this work aims to provide a computationally efficient yet effective alternative for high-quality image inpainting.

## III. The Approach

This section describes the dataset, model architectures, training strategy, and evaluation methodology employed in this study for object removal image inpainting.

### A. Dataset

The dataset used in this study is derived from the Defacto image and face manipulation dataset, which contains a large collection of object-removal inpainting samples. The dataset includes approximately 25,000 object-removal instances, each consisting of an input image, a corresponding binary mask indicating the removed region, and a ground truth image representing the desired reconstruction.

Each sample is organized into structured directories, where input images and their corresponding masks are explicitly paired. The binary masks provide precise spatial information about the missing regions, enabling controlled and consistent inpainting during training. The object-removal scenarios in the dataset vary significantly in terms of object size, location, and image content, making the dataset well-suited for evaluating the robustness of inpainting models.

The original source images are derived from the MSCOCO dataset, ensuring high visual diversity and real-world scene complexity. All images are resized to a fixed resolution of $256 \times 256$ pixels to maintain computational efficiency and consistency across experiments.

The dataset is divided into training, validation, and test sets. The training set consists of 30,000 images, evenly distributed among input, mask, and ground truth samples. The validation and test sets each contain 3,000 images, following the same structured distribution. This partitioning allows for reliable performance evaluation while avoiding data leakage.

A visual example of the dataset structure, including an input image, its corresponding mask, and the ground truth image, is shown in Fig. 1.

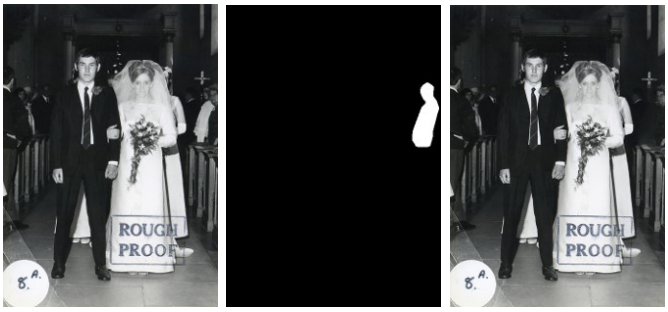


Fig. 1. Sample input image, binary mask, and ground truth image from the Defacto inpainting dataset.

### B. Model Architectures

Two autoencoder-based architectures are employed in this study: a conventional U-Net-like architecture and an enhanced U-Net architecture incorporating attention blocks.

The U-Net-like architecture follows an encoder–decoder structure with symmetric skip connections. The encoder progressively extracts hierarchical features by applying convolutional layers followed by downsampling operations, while the decoder reconstructs the image using upsampling layers and convolutional refinement. Skip connections between corresponding encoder and decoder layers preserve spatial information and improve reconstruction quality, particularly around object boundaries.

The second architecture extends the U-Net design by integrating attention mechanisms within the skip connections. These attention blocks allow the network to selectively emphasize relevant feature regions during the decoding process, enabling better contextual reasoning and reducing the influence of irrelevant background features. This mechanism is particularly beneficial for object removal tasks, where large missing regions require long-range dependency modeling.

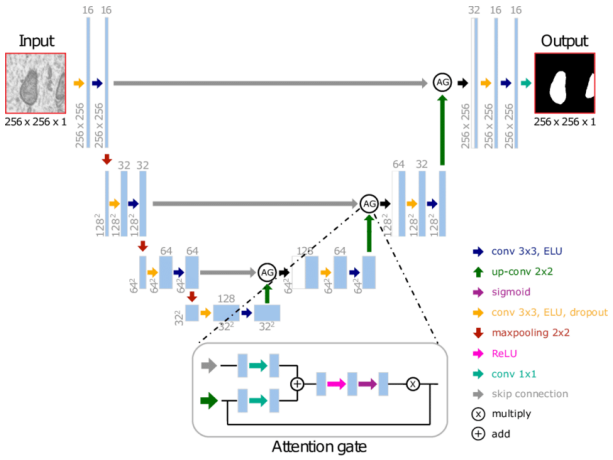


Fig. 2. Architecture of the U-Net model with attention blocks used for image inpainting.

An overview of the U-Net with attention blocks architecture is illustrated in Fig. 2.

Four model configurations are trained and evaluated: a U-Net-like model using Mean Squared Error (MSE) loss, an attention-based U-Net trained with perceptual loss, and two additional attention-based variants trained with different loss functions and epoch counts. This setup enables a systematic analysis of the impact of attention mechanisms and loss functions on inpainting performance.

### C. Training Strategy

All models are trained using a batch size of 32 and a learning rate of 0.001. The training process includes standard data preprocessing steps such as normalization and dataset shuffling. Model optimization is performed using backpropagation with gradient-based updates.

To enhance perceptual quality, perceptual loss is employed for selected models. This loss is computed by extracting feature representations from a pre-trained VGG19 network and measuring the discrepancy between reconstructed and ground truth images in the feature space. Specifically, activations from intermediate convolutional layers are used to encourage semantic and structural consistency beyond pixel-wise accuracy.

Training is conducted on an NVIDIA A100 GPU with 40 GB of memory, enabling efficient processing of the large dataset and stable convergence of deep architectures.

### D. Evaluation Metrics

Model performance is evaluated using a combination of quantitative and qualitative assessment methods. Quantitative evaluation includes Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). PSNR measures pixel-level reconstruction fidelity, SSIM assesses structural similarity, and LPIPS evaluates perceptual similarity based on deep feature representations.

In addition to numerical metrics, qualitative evaluation is performed through visual inspection of inpainted results. This assessment focuses on realism, structural coherence, and the seamless integration of reconstructed regions with the surrounding context. The combination of objective metrics and visual analysis provides a comprehensive understanding of model performance.

## IV. Experimental Results

### A. Quantitative and Qualitative Results

The performance of the proposed models is evaluated using both quantitative metrics and qualitative visual comparisons. A qualitative comparison of representative inpainting results is presented in Fig. 3, illustrating the visual differences between model outputs and the ground truth images.


Fig. 3. Qualitative comparison of inpainting results produced by different models.

Quantitative evaluation is conducted using PSNR, SSIM, and LPIPS metrics. The results demonstrate that different models excel under different evaluation criteria. The Attention U-Net model trained with Mean Squared Error (MSE) loss for 50 epochs achieves the highest average SSIM score of 0.98, indicating superior structural similarity to the ground truth images. This suggests that attention mechanisms effectively preserve spatial structure during the inpainting process.

In terms of PSNR, the U-Net-like model trained with MSE loss for 50 epochs attains the highest score of 37 dB. Since PSNR primarily measures pixel-wise reconstruction accuracy, this result indicates that the U-Net-like architecture is particularly effective in minimizing overall reconstruction error.

When evaluated using LPIPS, which measures perceptual similarity aligned with human visual perception, the Attention U-Net model trained with perceptual loss for 50 epochs achieves the lowest score of 0.0215. Lower LPIPS values correspond to higher perceptual similarity, highlighting the effectiveness of perceptual loss in producing visually realistic inpainted results.

The discrepancy between LPIPS and traditional metrics such as PSNR and SSIM emphasizes the importance of perceptual evaluation. While PSNR and SSIM provide valuable insights into reconstruction fidelity and structural consistency, they do not always correlate with perceived image quality. LPIPS, by leveraging deep feature representations, captures perceptual differences that are more consistent with human judgment.

Qualitative results further support these findings. Although some models achieve higher PSNR or SSIM values, visual inspection reveals that models optimized with perceptual loss often generate more natural textures and seamless transitions in the inpainted regions. In several examples, outputs with slightly lower PSNR scores are visually preferred due to improved perceptual coherence.

Overall, both quantitative and qualitative evaluations indicate that attention mechanisms and perceptual loss significantly enhance the visual quality of image inpainting, particularly for object removal tasks involving complex scenes.

It should also be noted that each evaluation metric captures different aspects of image quality. While PSNR emphasizes pixel-level accuracy and SSIM focuses on structural similarity, LPIPS provides a perceptual assessment that better aligns with human visual judgment. Therefore, relying on a combination of these metrics allows for a more comprehensive and reliable evaluation of inpainting performance.

### *B. Discussion*

Despite the promising results, several limitations were encountered during the experimental process. One notable challenge was the constraint imposed by computational resources. Although training was conducted using an NVIDIA A100 GPU, memory limitations restricted the maximum achievable image resolution and batch size. Attempts to train models at a resolution of $512 \times 512$ pixels did not yield stable convergence, affecting output quality.

Additionally, variability in reconstruction quality was observed across different samples, particularly in scenes with complex textures or large missing regions. This highlights the inherent difficulty of object removal image inpainting and suggests the need for more advanced contextual modeling or larger-scale training.

Nevertheless, the proposed approach demonstrates several strengths. The models consistently achieve strong PSNR and SSIM scores under constrained resources and outperform pretrained baseline models across all evaluated metrics. The use of a more diverse and generic dataset, rather than commonly used datasets such as Places2 or CelebA-HQ, further strengthens the robustness and real-world applicability of the proposed method.

In summary, the experimental results confirm that incorporating attention mechanisms and perceptual loss leads to significant improvements in image inpainting quality. While resource limitations impose certain constraints, the findings provide a solid foundation for future work aimed at scaling the models and further enhancing perceptual realism.

## V. Conclusion

In this study, an autoencoder-based image inpainting framework was presented with a specific focus on object removal tasks. Two different architectures, namely a U-Net-like model and an Attention U-Net model, were implemented and evaluated using the DeFacto image and face manipulation dataset. The experimental results demonstrated that incorporating attention mechanisms improves the model's ability to preserve structural consistency and generate more visually coherent reconstructions in regions with missing content.

Quantitative evaluations using reconstruction-based metrics such as PSNR, SSIM, and LPIPS, together with qualitative visual comparisons, showed that the Attention U-Net consistently outperformed the baseline U-Net architecture, particularly in complex occlusion and large missing-region scenarios. These results indicate that attention-enhanced encoder–decoder architectures are better suited for capturing long-range dependencies and contextual relationships within images.

Overall, the findings of this study highlight the importance of attention mechanisms in improving both structural fidelity and perceptual quality in image inpainting tasks. The proposed approach demonstrates strong potential for practical object removal applications, offering reliable and visually convincing reconstructions under challenging conditions.

## VI. Future Work

Future research may focus on extending the proposed framework to higher-resolution image inpainting in order to better capture fine-grained textures and subtle visual details. Handling larger image resolutions would enable more realistic reconstructions, especially for complex natural scenes and high-frequency textures.

Additionally, exploring alternative attention mechanisms or hybrid architectures that integrate transformer-based components could further improve global contextual understanding and long-range dependency modeling. Such architectures may enhance the semantic coherence of inpainted regions, particularly in images with complex structures.

Another promising direction involves incorporating adversarial training strategies and advanced perceptual loss functions to further improve visual realism beyond pixel-level accuracy. Finally, evaluating the proposed approach on larger, more diverse datasets and real-world object removal scenarios would provide deeper insights into its generalization capability, robustness, and applicability in practical image editing and restoration tasks.

## References


[1] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mask-Aware Transformer for Large Hole Image Inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Z. Zuo, L. Zhao, A. Li, Z. Wang, Z. Zhang, J. Chen, W. Xing, and D. Lu, "Generative Image Inpainting with Segmentation Confusion Adversarial Training and Contrastive Learning," 2021. [Online]. Available: https://github.com/comzzw/Generative-Image-Inpainting

[3] CelebA-HQ Dataset, [Online]. Available: https://www.tensorflow.org/datasets/catalog/celeb_a_hq

[4] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "The Places Database: A Large-Scale Scene Recognition Dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. [Online]. Available: https://www.tensorflow.org/datasets/catalog/placesfull

[5] Defacto Dataset Team, "Defacto Inpainting Dataset," [Online]. Available: https://www.kaggle.com/datasets/defactodataset/defactoinpainting

[6] Defacto Dataset Team, "Defacto: Image and Face Manipulation Dataset," [Online]. Available: https://defactodataset.github.io/