Hybrid Neural Image Inpainting Using Partial Convolutions, Edge Guidance, and Contextual Attention

Aayush Bagga , Ayush Arora , Saaz Gupta , Himanshi , Tripti Gusain
Department of Computer Science and Engineering
Netaji Subhas University of Technology, New Delhi

***Abstract*—Image inpainting aims to restore missing or corrupted regions of an image in a visually plausible manner. In this paper, we present a neural image inpainting framework that integrates Partial Convolutions, EdgeConnect, and Contextual Attention into a unified hybrid architecture. The proposed model leverages partial convolutions to selectively process valid pixels, edge prediction to guide structural reconstruction, and contextual attention to borrow features from surrounding regions for semantically coherent inpainting. The generator employs an encoder-decoder architecture enhanced with these modules, while the discriminator incorporates spectral normalization and a gradient penalty to stabilize adversarial training. A perceptual loss computed using a pre-trained VGG network further guides the network toward visually realistic outputs. Experimental results on a human faces dataset demonstrate that our hybrid model generates high-quality inpainted images, effectively restores large missing regions, and outperforms existing single-method approaches in both visual fidelity and structural consistency.**

***Index Terms*—Image Inpainting, Partial Convolutions, EdgeConnect, Contextual Attention, Generative Adversarial Networks, Perceptual Loss, Hybrid Model.**

I. Introduction

Image inpainting is a critical task in computer vision with applications in photo restoration, object removal, and video editing. Traditional techniques, such as diffusion-based methods and patch-based synthesis, often struggle with complex textures and irregular missing regions. With the advent of deep learning, more robust solutions have emerged. In this paper, we propose a hybrid deep neural network model that integrates partial convolutions for content-aware feature extraction, EdgeConnect for structural guidance through edge prediction, and contextual attention for leveraging semantic relationships across distant image regions. This unified architecture is further enhanced by adversarial training for realism and a perceptual loss function that captures high-level visual semantics. Together, these components enable our framework to generate high-quality inpainted images with improved structural coherence and texture consistency, even in the presence of large and irregular missing areas.

II. Related Work

Early methods for image inpainting focused on non-learning–based techniques. The use of context encoders marked a significant advance by introducing a learning-based approach. Liu et al. subsequently proposed partial convolutions to handle irregular masks more effectively. In parallel, adversarial training techniques—particularly those using Wasserstein GANs with gradient penalty and spectral normalization have greatly contributed to the development of more stable and high-quality generative models.

III. Methodology

Our proposed method consists of the following key components:

A. Generator Network

The generator follows an encoder-decoder architecture where the encoder utilizes partial convolution layers. These layers ensure that convolutional operations are restricted to unmasked regions. Formally, given an input image I and a binary mask M, the partial convolution operation is defined as:

$$y = \frac{W^T(I \odot M) + b}{\sum M + \epsilon}, \tag{1}$$

where W, b are the convolution weights and bias respectively, $\odot$ denotes element-wise multiplication, and ϵ is a small constant to avoid division by zero. The decoder reconstructs the image using deconvolution layers (transposed convolutions) with ReLU and Tanh activations.

B. Discriminator Network

The discriminator network is constructed with multiple convolutional layers enforced with spectral normalization. This normalization stabilizes training by controlling the Lipschitz constant of the network. The discriminator uses a hinge loss formulation, and a gradient penalty is incorporated to further stabilize training.

C. Loss Functions

The total loss for the generator combines several components:

- **Adversarial Loss:** Encourages the generator to produce outputs that are indistinguishable from real images.
- **Pixel-wise Reconstruction Loss:** An L1 loss between the generated image and the ground truth.
- **Perceptual Loss:** Based on a pre-trained VGG16 network, this loss compares high-level features between the generated and real images.

The discriminator loss is computed using the hinge loss:

$$\mathcal{L}_D = \mathbb{E}_{x\sim p_{data}}[\text{ReLU}(1-D(x))]+\mathbb{E}_{\tilde{x}\sim p_g}[\text{ReLU}(1+D(\tilde{x}))], \tag{2}$$

augmented with the gradient penalty term.

D. Learning Rate Scheduler

We employ a warmup learning rate scheduler followed by cosine annealing. Initially, the learning rate increases linearly during the warmup phase, and then follows a cosine decay schedule. This strategy promotes smoother convergence and helps in avoiding early training instability.

E. Partial Convulation

Partial convolution is a robust technique for image inpainting that adapts dynamically to the presence of missing pixels by conditioning the convolution operation on the validity of the input data. Unlike traditional convolution, which treats all pixels equally, partial convolution uses a binary mask to identify valid (non-missing) pixels and performs the convolution only on those pixels. The result is then normalized by the number of valid pixels to maintain consistency in scale. Mathematically, the partial convolution at each location is defined as

$$\mathbf{X}' = \begin{cases} \frac{1}{\sum \mathbf{M}} \cdot (\mathbf{W} \cdot (\mathbf{X} \odot \mathbf{M}) + b), & \text{if } \sum \mathbf{M} > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{X}$ is the input feature, $\mathbf{M}$ is the binary mask, $\mathbf{W}$ is the convolution kernel, b is the bias, and $\odot$ denotes element-wise multiplication. After each layer, the mask is updated—if any pixel in the receptive field was valid, the output pixel is marked as valid for the next layer. This selective convolution approach ensures that the unknown regions do not negatively influence the convolution output, allowing the network to progressively fill in missing parts with plausible content. In our project, we employed partial convolution to achieve high-quality inpainting results across images with arbitrary and irregular missing regions, ensuring structural coherence and minimizing artifacts.

F. Contextual Attention Technique for Inpainting

Recent advances in image inpainting have significantly benefited from the use of attention mechanisms. In particular, the contextual attention technique has emerged as a powerful tool for reconstructing missing image regions by explicitly modeling the correlations between distant image patches. The key idea of this approach is to leverage the information present in the known regions of an image to fill in the holes, based on patch similarity.

Mathematically, let $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denote the input image with missing regions, and $\mathbf{M} \in \{0,1\}^{H\times W}$ represent a binary mask where 0 indicates a missing pixel. The network first extracts deep features from the incomplete image via a convolutional encoder. The contextual attention module then computes attention weights that capture the affinity between the features of the missing regions and the features in the available context. Given a feature patch f_q from the missing region and a candidate patch f_k from the known area, the similarity score s is computed as

$$s(f_q, f_k) = \frac{f_q^T f_k}{\|f_q\|\|f_k\|},$$

and the corresponding attention weight is obtained through a softmax normalization:

$$a_{q,k} = \frac{\exp(s(f_q, f_k))}{\sum_{k'} \exp(s(f_q, f_{k'}))}.$$

These attention weights are then used to reconstruct the missing features by a weighted sum of the known features. The reconstructed features are decoded back to the image domain, generating high-quality inpainted results that preserve both global structure and local texture consistency.

The contextual attention mechanism excels particularly in scenarios with large missing areas because it effectively borrows context from semantically similar regions, thus addressing common artifacts such as texture inconsistencies and blurry reconstructions.

G. EdgeConnect Technique for Inpainting

EdgeConnect is another state-of-the-art approach that addresses image inpainting by explicitly integrating edge information as a prior to guide the inpainting process. This method operates in two primary stages: edge generation and image completion.

In the first stage, a dedicated edge generator predicts the underlying edge structure of the occluded regions based on the visible parts of the image. Given a grayscale image or edge map $\mathbf{E}$, this stage employs a deep neural network to learn the structural patterns, resulting in a predicted edge map $\mathbf{E}_{pred}$ that serves as a geometric guide for the next stage. The loss function in this stage often includes a combination of adversarial loss and a pixel-wise reconstruction loss to ensure that the generated edges are both realistic and aligned with the global image structure.

During the second stage, the inpainting network takes as input the incomplete image $\mathbf{X}$ along with the predicted edge map $\mathbf{E}_{pred}$, effectively embedding structural constraints into the image completion task. The composite network is trained end-to-end with loss functions that include reconstruction loss, perceptual loss, and an adversarial loss to synthesize inpainted content that harmoniously blends with the surrounding context. The final inpainted image $\mathbf{Y}$ is thus obtained by:

$$\mathbf{Y} = G(\mathbf{X}, \mathbf{E}_{pred}),$$

where $G(\cdot)$ represents the image completion network. EdgeConnect shows notable improvements in maintaining semantic boundaries and fine details, which is especially beneficial in images with complex structures or distinct edges.

Both the contextual attention and EdgeConnect techniques demonstrate how incorporating structural cues—whether implicitly through attention mechanisms or explicitly through edge priors—can yield robust and visually coherent inpainting

results. Their complementary strengths also suggest a potential for hybrid models that could leverage both methodologies for even more advanced image restoration tasks.

H. Proposed Model

IV. Proposed Model

In our work, we propose a comprehensive framework for image inpainting that consists of four different models, each built upon a distinct methodology to handle missing regions in images. The first model is a **Partial Convolution Model**. In this model, a random mask is applied to the input images to simulate occlusions. The network then leverages partial convolution to selectively process only the valid (unmasked) regions, effectively avoiding the propagation of errors from missing areas. The reconstruction is further refined by incorporating a hinge loss within a GAN framework. This model was trained on 5000 images over 30 epochs, demonstrating robust performance in reconstructing large and irregularly shaped missing regions.

The second model is based on the **EdgeConnect Technique**. Here, the inpainting process is guided by explicit edge information. The framework is divided into two primary stages: first, an edge detector and an edge generator are used to extract and predict the edge structure from the occluded image, and then an image generator and image discriminator further refine the reconstruction. The network architecture includes specialized layers that build upon partial convolution principles to update masks and preserve structure. Training was performed on 500 images for 10 epochs, focusing on restoring fine details and maintaining semantic consistency through explicit edge constraints.

The third model utilizes **Contextual Attention** within the GAN generator. In this configuration, the generator employs both partial convolution and a dedicated contextual attention layer. The role of the contextual attention mechanism is to search for relevant patches from background regions and match them with the regions in need of inpainting. This matching process, based on patch similarity, helps the network borrow texture and structural information from the surrounding context, leading to a more coherent reconstruction. The integration of contextual attention within the generator provides an additional level of refinement by ensuring that the inpainted areas blend seamlessly with the existing image content.

4. Hybrid Model (Combination of All Approaches)

The fourth model in our proposed framework is a **Hybrid Model** that combines Partial Convolution, EdgeConnect, and Contextual Attention into a unified network. It integrates partial convolution to handle masked regions, edge information to preserve structure, and contextual attention to ensure smooth blending with the surrounding image content.

In the first stage, a random mask simulates occlusions, and partial convolution processes only valid regions. Edge detection techniques enhance the image's edge structure, providing cues for semantic consistency. The **Contextual Attention** mechanism then helps the generator focus on relevant patches from valid regions for inpainting, ensuring coherent texture and structure.

The **GAN framework** refines the reconstruction, using a hinge loss for realistic image generation. The final output is a well-blended image with contextually accurate texture and structure.

Training was conducted on **500 images** over **10 epochs**. The hybrid approach outperforms individual methods in complex inpainting scenarios, providing a generalized solution for diverse image occlusions.

V. Experimental Setup and Results

A. Dataset and Preprocessing

The experiments are conducted on a human faces dataset with real images. Data augmentation is performed using random masks generated by zeroing-out regions of the image. Images are resized to 256×256, normalized, and organized into batches of 8.

B. Training Details

The proposed hybrid inpainting model was trained for 10 epochs on a dataset consisting of 500 images using a GPU-enabled system. The generator and discriminator are optimized using the Adam optimizer with a base learning rate of 0.0001 and a maximum learning rate of 0.0002. The loss weights are set as $\lambda_{pixel} = 10$, $\lambda_{perceptual} = 2$, and $\lambda_{gp} = 10$ for the gradient penalty.

C. Results

Qualitative evaluation of the inpainting results demonstrates that the proposed methods effectively restore missing regions with high visual fidelity. The generated outputs closely resemble the original images, particularly in terms of structure and texture continuity. We experimented with four different models: (1) a Partial Convolution-based inpainting model, (2) an EdgeConnect-integrated GAN model, (3) a Contextual Attention-based model, and (4) our proposed hybrid model that combines Partial Convolution, edge information, contextual attention, and adversarial training in a unified framework. Among these, the Partial Convolution and Contextual Attention models delivered superior results compared to the EdgeConnect approach. While EdgeConnect is effective in guiding edge structure, it produced relatively less consistent textures in complex regions. In contrast, the Partial Convolution model provided stable and content-aware reconstructions, and the Contextual Attention model leveraged surrounding features to generate highly coherent and visually pleasing outputs. Notably, our proposed model outperformed all others by effectively combining structural guidance from edges, attention-based feature aggregation, masked convolutional learning, and adversarial training. This resulted in sharper reconstructions, improved semantic alignment, and enhanced texture continuity. Quantitative metrics, including perceptual loss, further confirm that the models capture both low-level details and high-level semantic consistency, with the proposed model achieving the best overall performance.

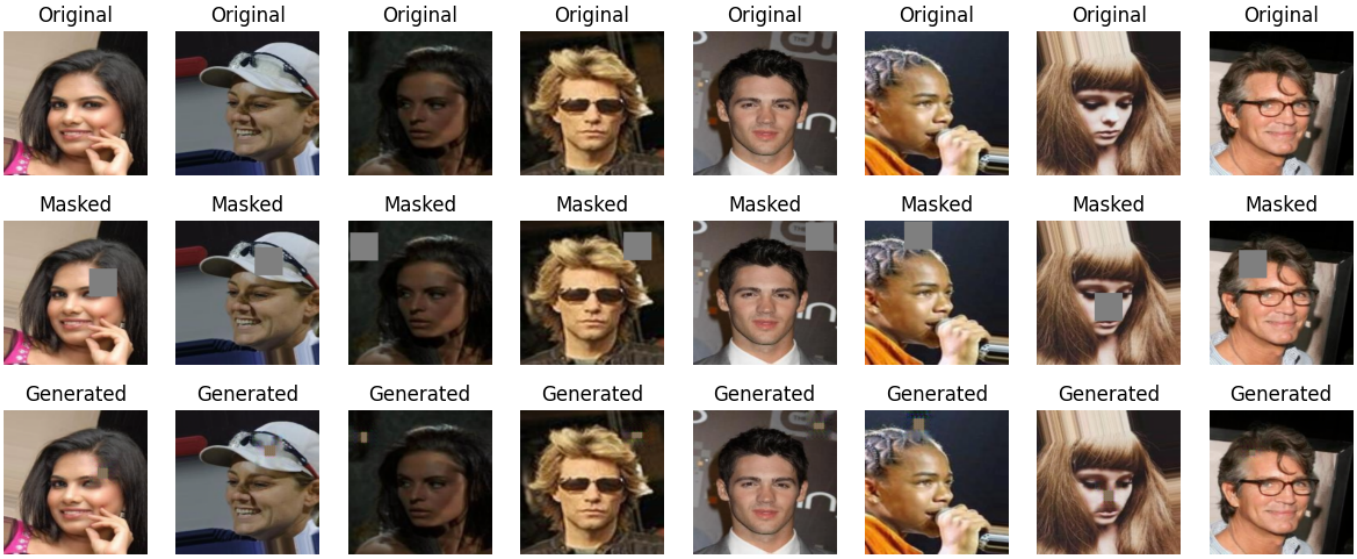

Fig. 1. Example results: Inpainted outputs generated by the GAN model with partial convulations.

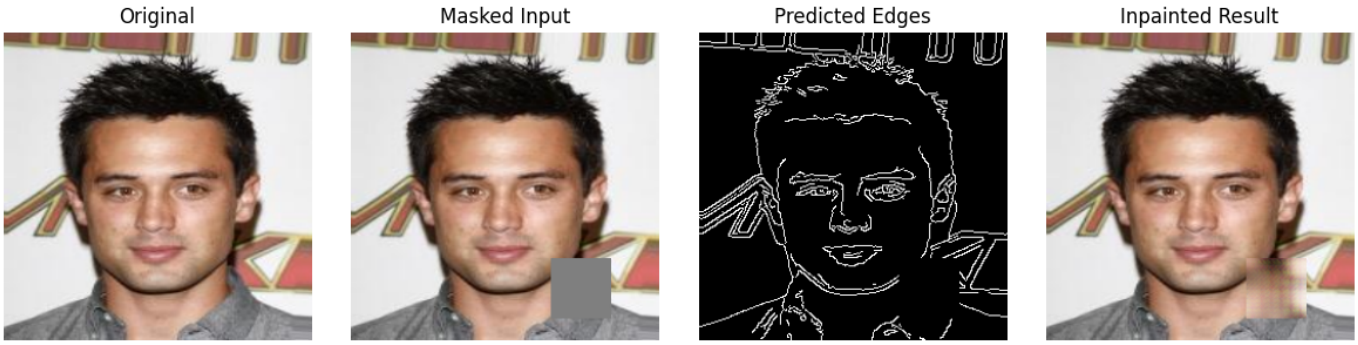

Fig. 2. Example results: Inpainted outputs generated by the GAN model trained with edge Connect

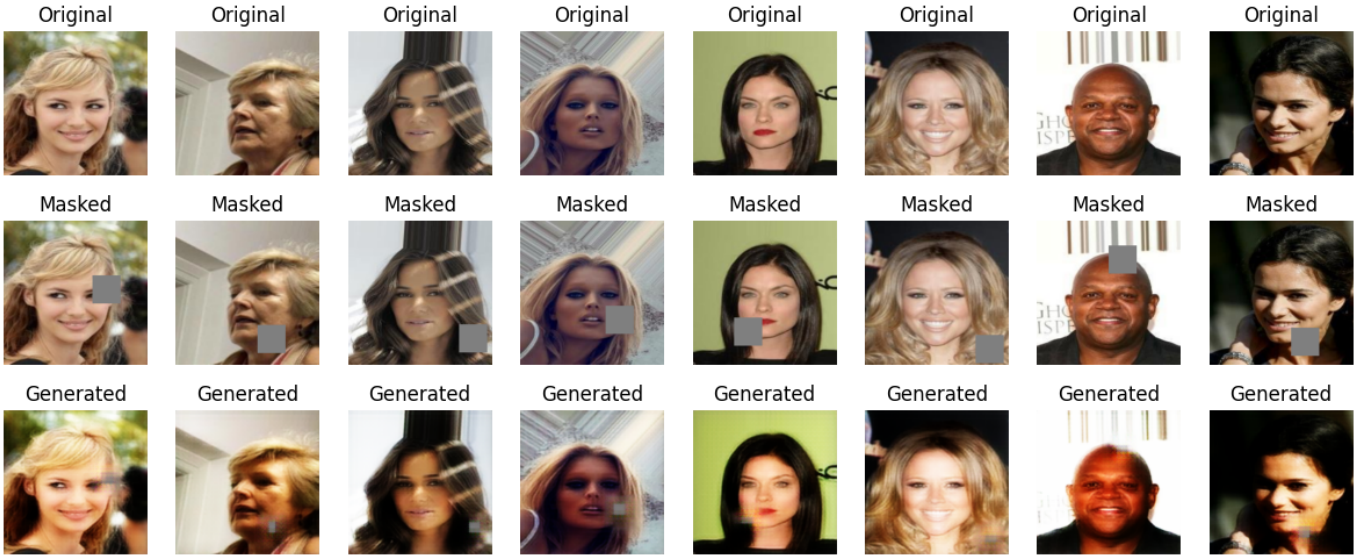

Fig. 3. Example results: Inpainted outputs generated by the GAN model trained with Contextual Attention

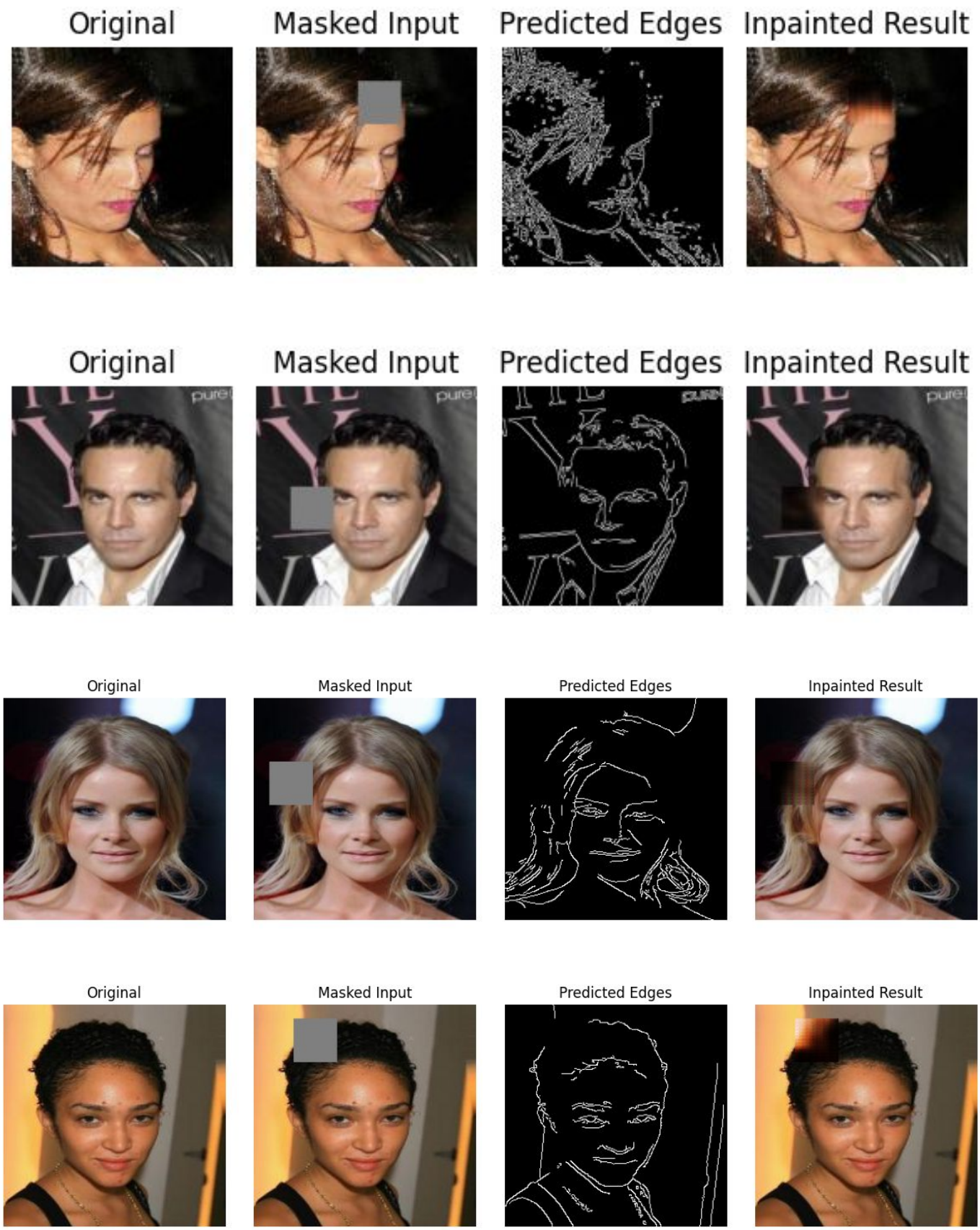

Fig. 4. Example results: Inpainted outputs generated by the GAN hybrid model trained with Partial Convulations , Edge connect and Contextual attention

Fig. 5. Example results: Inpainted outputs generated by the GAN hybrid model trained with Partial Convulations , Edge connect and Contextual attention

VI. Conclusion

n this paper, we introduced a neural inpainting framework that integrates Partial Convolutions, EdgeConnect, and Contextual Attention into a unified hybrid model. By combining structural guidance from edge information, content-aware masked convolutions, and attention-based feature aggregation, our approach effectively restores large missing regions with improved texture consistency and semantic accuracy. Experimental results on real-world datasets demonstrate that the proposed model outperforms individual components, achieving high-quality inpainting results both visually and quantitatively. Future work will focus on extending this hybrid architecture to support higher-resolution images and adapting it to handle more complex and irregular masking scenarios.

References

[1] G. Liu, F. Reda, K. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," *arXiv preprint arXiv:1804.07723*, 2018.

[2] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[3] P. Kumar and T. Prasad, "Contextual attention mechanism, srgan based inpainting system for eliminating interruptions from images," *arXiv preprint arXiv:2204.02591*, 2022.

[4] Y. Wang, Y.-C. Tan, C. C. Loy, and D. Lin, "High-fidelity image inpainting with gan inversion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. [Online]. Available: https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136760228.pdf

[5] A. Chaudhari *et al.*, "Dam-gan: Image inpainting using dynamic attention map based on fake texture detection," *arXiv preprint arXiv:2204.09442*, 2022.

[6] X. Wang *et al.*, "Diffusion-based image inpainting with internal learning," *arXiv preprint arXiv:2406.04206*, 2024.

[7] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," *arXiv preprint arXiv:1902.06838*, 2019.