IMPBÆ: A New Perspective of Image Inpainting for Object Removal

Abdullah Enes Ergun[1] **Baha Kirbasoglu**[1]

Abstract

In this project, the complex task of image inpainting was addressed, with a specific focus on object removal to reconstruct missing or damaged regions in images with high fidelity and visual coherence. Two distinct autoencoder architectures were developed and compared: a traditional U-Net-like model and an enhanced U-Net with attention blocks. These models were designed to seamlessly integrate inpainted regions with the surrounding context.

To train these models, a dataset derived from the Defacto image and face manipulation dataset, which includes approximately 25,000 instances of object-removal forgeries, was utilized. During training, various loss functions, including Mean Squared Error (MSE) and perceptual loss, were employed to improve the fidelity of the inpainted images.

Extensive evaluations were conducted using quantitative metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), as well as qualitative visual assessments. The results demonstrated the superior performance of the proposed models.

Specifically, the highest SSIM score was achieved by the Attention U-Net model using MSE loss at 50 epochs, indicating exceptional structural similarity to ground truth images. Conversely, the U-Net-like model with MSE loss at 50 epochs attained the highest PSNR score, indicating minimal pixel-wise reconstruction errors. Additionally, the lowest LPIPS score was achieved by the Attention U-Net model using Perceptual Loss at 50 epochs, indicating the best perceptual similarity to the ground truth images.

These findings underscore the effectiveness of attention mechanisms and perceptual loss in enhancing image inpainting quality. The proposed models significantly outperformed pretrained models both quantitatively and qualitatively, representing valuable advancements in image processing techniques.

1. Introduction

The problem of image inpainting, involving the reconstruction of missing or damaged regions in images, is considered challenging and crucial for various applications. In this project, the objective is to develop and train a model for image inpainting using a dataset comprising base images, masks delineating missing regions, and corresponding output images where these regions are restored. The goal is to create a robust model capable of accurately inpainting missing regions in images, leveraging surrounding context information from the dataset. Several challenges are posed by image inpainting, including the preservation of semantic consistency, maintenance of visual coherence, and seamless integration of inpainted regions with the surrounding context. Complex visual patterns and textures must be understood by the model to generate realistic inpaintings that blend naturally with the rest of the image. Achieving this requires sophisticated algorithms and training strategies to capture the nuances of image structures and textures.

Object Removal Image Inpainting, a subset of image inpainting, focuses specifically on the task of removing unwanted objects or elements from images while ensuring the seamless integration of the inpainted regions with the surrounding context. This area of study is of particular interest due to its wide-ranging applications in photography, digital art, and image editing. By effectively removing objects from images, inpainting techniques enable users to manipulate and enhance images with greater precision, allowing for the creation of visually appealing compositions. Object Removal Image Inpainting presents its own set of challenges, including accurately identifying and delineating the objects to be removed, preserving the visual coherence and semantic consistency of the resulting images, and ensuring that the inpainted regions blend naturally with the surrounding context. Sophisticated algorithms and training strategies are required to address these challenges and produce realistic and visually pleasing inpaintings.

From a practical standpoint, Object Removal Image Inpainting has numerous applications across various domains. In photography and digital art, it allows for the removal of unwanted distractions or imperfections, enhancing the overall aesthetic quality of the images. In forensic analysis and surveillance, it can be used to remove sensitive information

or identifying features from images while preserving the integrity of the remaining content. Additionally, in fields such as advertising and marketing, Object Removal Image Inpainting enables the creation of visually appealing content by removing unwanted elements from images.

Overall, Object Removal Image Inpainting is a specialized area of study within image inpainting that addresses the specific task of removing objects from images while maintaining visual coherence and semantic consistency. By advancing techniques and algorithms in this field, researchers aim to contribute to the development of more sophisticated image editing tools and applications with practical implications across various domains.

The significance of image inpainting lies in its broad applicability across various domains. By effectively restoring missing or damaged regions in images, inpainting techniques enhance the visual quality and utility of images for diverse applications. In fields such as photography and digital art, inpainting enables the removal of unwanted objects or imperfections, allowing images to be manipulated with greater flexibility and precision. Moreover, image inpainting plays a crucial role in image restoration and preservation. In historical imaging, inpainting techniques can help recover missing or deteriorated portions of valuable photographs or artworks, contributing to cultural heritage preservation efforts. In medical imaging, inpainting can facilitate the reconstruction of missing anatomical structures in diagnostic images, aiding in accurate medical diagnosis and treatment planning.

From a technological standpoint, advancements in image inpainting techniques contribute to the broader field of computer vision and deep learning. By developing and fine-tuning models for image inpainting, the boundaries of what is possible in image processing are pushed, fostering innovation and enabling the development of more sophisticated applications. Additionally, the availability of high-quality inpainting models facilitates research and development in related areas such as image editing, scene completion, and object removal.

In summary, object removal image inpainting represents a specialized aspect of the broader image inpainting challenge, with substantial implications across diverse domains. By tackling the complexities of seamlessly removing unwanted objects from images while preserving visual coherence and semantic consistency, this project aims to contribute significantly to advancing image inpainting techniques.

2. Releated Work

The work[1] presented introduces a novel solution to the challenging problem of large hole image inpainting by leveraging the power of transformer architectures while addressing the computational limitations associated with them. Traditional methods relying on convolutional neural networks (CNNs) for image completion often struggle with capturing long-range dependencies necessary for accurately filling large missing regions. To tackle this issue, the authors propose a Mask-Aware Transformer (MAT) model that combines the strengths of transformers and convolutions to efficiently process high-resolution images.

One key innovation of the MAT framework lies in its customized transformer blocks tailored specifically for large mask inpainting. Unlike conventional transformer architectures, the proposed MAT design removes layer normalization and replaces residual learning with fusion learning using feature concatenation. This adjustment enhances optimization stability and improves performance, particularly crucial for effectively handling masks with significant missing areas. By focusing on feature concatenation over residual connections, MAT ensures that long-range dependencies are effectively modeled, enabling the network to capture contextual information across the entire image space.

Moreover, the authors introduce a multi-head contextual attention mechanism within the transformer body to facilitate efficient long-range dependency modeling. This attention mechanism dynamically aggregates information from partial valid tokens, guided by a dynamic mask that updates iteratively during the network's propagation. By selectively incorporating valid tokens for computing relations, MAT enhances computational efficiency without compromising effectiveness, crucial for processing large-scale masks while maintaining high fidelity in the generated images.

Furthermore, the MAT framework integrates a style manipulation module to support pluralistic generation, enhancing the diversity of generated image completions. By modulating the weights of convolutional layers with noise inputs and incorporating unconditional generation, MAT can produce multiple plausible completions for a given masked image, catering to a wide range of potential inpainting scenarios.

The efficacy of the proposed MAT framework is demonstrated through extensive experiments on benchmark datasets such as Places and CelebA-HQ. The results showcase MAT's ability to achieve state-of-the-art performance in terms of both image quality and diversity, surpassing existing approaches in handling large-scale inpainting tasks. Through its innovative combination of transformer-based architectures, customized transformer blocks, and partial attention mechanisms, MAT represents a significant advancement in the field of image completion, opening up new avenues for addressing complex inpainting challenges in high-resolution imagery.

The study[2] introduces a novel adversarial training framework for image inpainting, combining segmentation confu-

sion adversarial training (SCAT) with contrastive learning to enhance the generation of realistic and coherent image completions.

SCAT, the cornerstone of the proposed framework, draws inspiration from human perception of low-quality repaired images. It orchestrates an adversarial game between an inpainting generator and a segmentation network. Here, the segmentation network aims to correctly label the generated and valid regions in the inpainted image, while the inpainting generator endeavors to deceive the segmentation network by filling missing regions with visually plausible content. This dynamic creates a challenging scenario for the segmentation network, as it must distinguish between filled and original areas. SCAT furnishes pixel-level local training signals, offering adaptability in handling images with arbitrary hole shapes.

To further refine and stabilize training, the framework incorporates contrastive learning losses, leveraging the feature representation space of the discriminator. By pulling inpainted images closer to ground truth images while simultaneously pushing them away from corrupted images, these contrastive losses guide the training process effectively. Considering the training of image inpainting as a process of learning a mapping from corrupted inputs to ground truth outputs, the proposed contrastive losses act akin to pull and push forces, steering the inpainting process towards generating more realistic and faithful completions.

By integrating SCAT and contrastive learning within the adversarial training framework, the study contributes to advancing the state-of-the-art in image inpainting. The synergistic combination of these techniques not only enhances the perceptual quality and coherence of generated image completions but also fosters robustness and adaptability in handling diverse inpainting scenarios. Through empirical evaluation and experimentation, the effectiveness and efficacy of the proposed framework are demonstrated, underscoring its potential for various real-world applications requiring high-fidelity image completion.

In this study, the work[2] is referenced as a pretrained model for comparison with the four distinct models trained within the project.

CelebA-HQ[3] stands for "CelebA-High Quality." It is an extension of the original CelebA dataset, which stands for "Celebrities Attributes." CelebA-HQ is designed for the task of face attribute prediction. The original CelebA dataset contains over 200,000 celebrity images, each annotated with 40 attribute labels. However, these images are low-resolution (178x218 pixels).

To address the need for higher resolution images, the CelebA-HQ dataset was created. It contains images with much higher quality and resolution, specifically 1024x1024 pixels. This dataset is useful for tasks that require more detailed and higher quality images of faces, such as facial recognition, attribute prediction, and image generation tasks.

Figure 1. Example of CelebA-HQ dataset

The Places dataset[4] is a collection of images designed for scene recognition. It was created to facilitate research in computer vision, specifically in tasks related to understanding scenes from images. The dataset is large-scale and contains millions of images categorized into over 400 scene categories.

The Places dataset is useful for training and evaluating algorithms in scene classification, scene parsing, and related tasks. It covers a wide range of scenes, from indoor environments like bedrooms and kitchens to outdoor landscapes like mountains and beaches.

Researchers often use the Places dataset to train deep learning models to recognize and understand scenes, enabling applications in autonomous navigation, content-based image retrieval, and more.

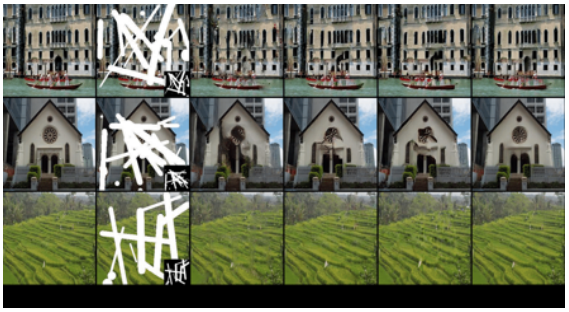

Figure 2. Example of Places Dataset

3. The Approach

3.1. Dataset

The dataset utilized for evaluation, sourced from the Defacto image and face manipulation dataset, encompasses approximately 25,000 instances of object-removal forgeries, making it an extensive and diverse resource for image in-

painting research. Each instance is meticulously organized within the ”inpainting” directory and is supplemented with a binary mask housed within the ”inpaint-mask” subdirectory, designated for utilization within the inpainting algorithm. This meticulous structure ensures that researchers have clear and direct access to the necessary components for effective algorithm training and evaluation.

The inclusion of the ”inpaint-mask” subdirectory adds significant value to the dataset, as it provides precise guidance for the inpainting process, helping to improve the accuracy and efficiency of the algorithms being developed. The comprehensive coverage of object-removal scenarios within the dataset means that it encompasses a wide range of potential challenges and variations that an inpainting algorithm might encounter in real-world applications.

This dataset constitutes a valuable asset for research endeavors in image inpainting, offering a diverse repertoire of object-removal scenarios accompanied by corresponding masks. The well-organized nature of the dataset facilitates not only the training and validation phases but also robust evaluation, allowing researchers to measure the performance of their algorithms under consistent and reproducible conditions. Such meticulous organization and comprehensive coverage afford researchers a robust platform for advancing the field of image manipulation detection and mitigation.

The original images, sourced from MSCOCO, ensure high-quality and diverse image content. Although these originals are not included in the inpainting dataset, their use as a source guarantees that the object-removal forgeries are based on a wide array of real-world images, further enhancing the dataset's utility and relevance. This combination of high-quality source material and detailed inpainting masks makes the dataset an indispensable tool for those aiming to push the boundaries of what is possible in image inpainting and manipulation detection. The dataset employed in this research comprises a total of 30,000 training images, systematically categorized into three distinct groups: 10,000 input images, 10,000 mask images, and 10,000 output images. This division ensures a balanced and comprehensive dataset that facilitates the training process. Furthermore, the validation and test sets are each composed of 3,000 images, with an equal distribution of 1,000 images per category input, mask, and output. This structured organization of data enhances the robustness and reliability of the training, validation, and testing processes.

The volume of data utilized represents the maximum capacity that can be efficiently managed by the available computational resources, specifically in terms of GPU RAM and system RAM. To optimize processing, all images have been resized to dimensions of 256 by 256 pixels. The batch size has been set to the default value of 32, and the learning rate has been maintained at the default setting of 0.001. These parameters have been carefully chosen to balance computational efficiency with model performance.

The training of images has been conducted on an NVIDIA A100 GPU, which boasts 40 GB of GPU RAM. This high-performance hardware has been instrumental in handling the substantial volume of data, ensuring that the training process is both efficient and effective. The use of such advanced computational resources underscores the rigor and thoroughness of the experimental setup, providing a solid foundation for the subsequent analysis and evaluation of the inpainting algorithms.

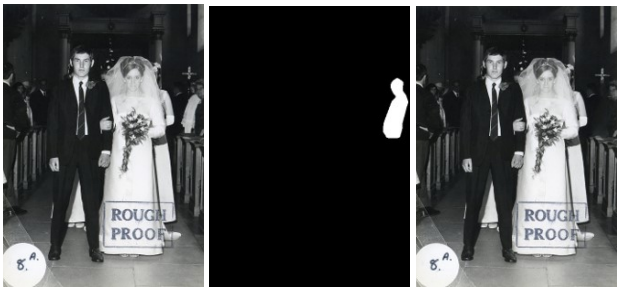

Figure 3. Sample Input, Mask and Ground Truth

3.2. Architectures

This study employs two distinct autoencoder architectures for image reconstruction tasks: a U-Net-like architecture and a U-Net with attention blocks. The U-Net-like architecture is a convolutional neural network designed for image segmentation and reconstruction. It consists of an encoder-decoder structure. The encoder part contains convolutional layers followed by max-pooling layers, which gradually reduce the spatial dimensions of the input image while increasing the depth of feature maps. The decoder part includes upsampling layers and convolutional layers, which aim to restore the original spatial dimensions of the image while refining the feature maps. Skip connections link corresponding layers in the encoder and decoder to facilitate the propagation of high-resolution features and improve the reconstruction quality.

The U-Net with attention blocks architecture enhances the traditional U-Net by integrating attention mechanisms, which help the model focus on relevant features during the reconstruction process. Similar to the U-Net-like architecture, the encoder part consists of convolutional layers and max-pooling layers. Attention gates are introduced in the skip connections to selectively emphasize important features from the encoder during the decoding process. The decoder part includes upsampling layers and convolutional layers, augmented with attention mechanisms to refine the reconstructed output.

Four models are trained using the aforementioned archi-

tectures and different loss functions across varying epochs: U-Net-like using MSE Loss at 50 epochs, U-Net with attention blocks using perceptual loss at 50 epochs, U-Net with attention blocks using perceptual loss at 20 epochs, and U-Net with attention blocks using MSE loss at 50 epochs.

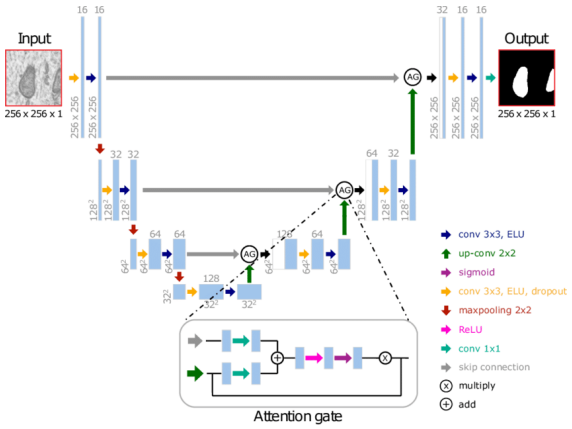

Figure 4. Architecture of U-Net with Attention Blocks

To enhance the perceptual quality of the reconstructed images, perceptual loss is utilized. A custom trainer was coded specifically to optimize the perceptual loss. Features are extracted using a pre-trained VGG19 network, specifically the activations from the block4_conv4 layer, to compute the perceptual loss. This ensures that the reconstructed images preserve high-level features similar to the ground truth images.

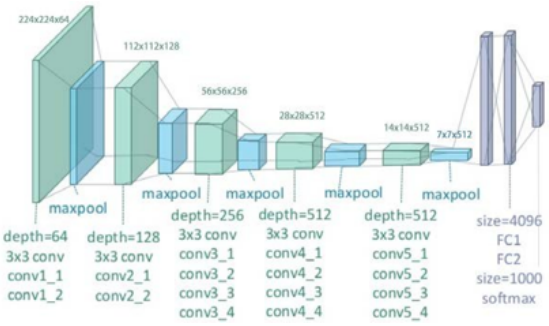

Figure 5. VGG19 Architecture

The models undergo a training process that includes data preparation, model initialization, training loop, and evaluation. The dataset undergoes preprocessing to normalize the images and is split into training and validation sets. Each model is initialized with the corresponding architecture and loss function. The models are trained for the specified number of epochs, involving forward propagation, loss computation, backpropagation, and weight updates. After training, the models are evaluated on the validation set to assess their reconstruction performance using metrics such as MSE and perceptual loss. By comparing the performance of these models, the aim is to determine the effectiveness of attention mechanisms and perceptual loss in improving image reconstruction quality.

```
# Load pre-trained VGG19 model (or any other suitable pre-trained model)
vgg = tf.keras.applications.VGG19(include_top=False, weights='imagenet', input_shape=(256, 256, 3))
vgg.trainable = False

# Define layers from which to extract features for perceptual loss
content_layers = ['block4_conv4']

# Define a model that outputs the features from the selected layers
content_model = tf.keras.Model(inputs=vgg.input, outputs=[vgg.get_layer(layer).output for layer in content_layers])

# Define perceptual loss function
# Scale images to the range expected by VGG model
y_true = targets * 255.0
y_pred = predictions * 255.0

# Get features from VGG for both true and predicted images
true_features = content_model(y_true)
pred_features = content_model(y_pred)

# Compute MSE loss between features
loss = 0.0
for true_feature, pred_feature in zip(true_features, pred_features):
    loss += tf.reduce_mean(tf.square(true_feature/255.0 - pred_feature/255.0))

gradients = tape.gradient(loss, self.model.trainable_variables)
self.optimizer.apply_gradients(zip(gradients, self.model.trainable_variables))
total_loss += loss.numpy()
```

Figure 6. Code Part from Trainer

3.3. Evaluation Metrics

The effectiveness of image inpainting is evaluated through a combination of quantitative metrics and qualitative assessments. Among the quantitative metrics, the Peak Signal-to-Noise Ratio (PSNR) plays a crucial role. PSNR quantifies the disparity between the original and inpainted images by measuring the signal-to-noise ratio. This metric is calculated using the mean squared error (MSE) between the original and inpainted images. A higher PSNR value, measured in decibels (dB), indicates that the inpainted image has a closer resemblance to the original, with less distortion and higher fidelity. PSNR is widely used due to its simplicity and ability to provide a clear numerical value representing the image quality. However, it primarily focuses on pixel-wise differences and might not fully capture perceptual differences perceived by human observers.

$$\mathrm{PSNR} = 10 \times \lg\left(\frac{255^2}{\mathrm{MSE}}\right)$$

$$\mathrm{MSE} = \frac{1}{M \times N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[I(i,j) - I'(i,j)\right]^2$$

Figure 7. Formula of PSNR and MSE

Figure 7 illustrates the formulas for Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR). MSE calculates the average squared difference between corresponding pixel intensities of the original image I and the reconstructed image K, normalized by the image dimensions m and n. PSNR measures the quality of the reconstructed image by comparing it to the original, represented in terms of MSE and the maximum possible pixel value (MAX).

Another essential metric is the Structural Similarity Index (SSIM), which assesses the resemblance between the original and inpainted images in terms of luminance, contrast, and structure. Unlike PSNR, SSIM evaluates image quality

based on perceived changes in structural information, which is crucial for human vision. It calculates a similarity score by comparing local patterns of pixel intensities that have been normalized for luminance and contrast. SSIM values range between -1 and 1, where a value closer to 1 indicates higher similarity. This metric is particularly effective in capturing changes in texture and structural information, making it more aligned with human visual perception compared to PSNR. SSIM's focus on structural information rather than mere pixel-wise comparison provides a more comprehensive assessment of image quality, especially for inpainting tasks where preserving structural integrity is crucial.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

Figure 8. Formula of SSIM

Figure 8 illustrates the formula for Structural Similarity Index (SSIM). SSIM measures the similarity between two images, typically the original image I and the reconstructed image K. It is calculated using the formula where μ_I, μ_K are the mean values of I and K, σ_I^2, σ_K^2 are the variances of I and K, σ_{IK} is the covariance between I and K, and c_1 and c_2 are constants to stabilize the division with weak denominator.

SSIM evaluates image similarity based on luminance, contrast, and structure, providing a measure of the perceptual quality of the reconstructed image compared to the original.

Additionally, perceptual metrics such as the Learned Perceptual Image Patch Similarity (LPIPS) are utilized to measure the perceptual similarity between image patches. LPIPS employs deep neural networks to evaluate the similarity of image patches by comparing features extracted from these patches. This approach provides a more nuanced assessment of visual quality by taking into account higher-level features that are important for human perception. LPIPS is particularly useful for assessing the perceptual realism of inpainted images, as it captures subtle differences that might not be reflected in traditional pixel-wise metrics like PSNR and SSIM. By leveraging pre-trained deep learning models, LPIPS can effectively quantify perceptual differences, making it a valuable tool for evaluating the effectiveness of inpainting techniques.

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$

Figure 9. Formula of LPIPS

The provided formula, shown in Figure 9, represents the calculation of the Learned Perceptual Image Patch Similarity (LPIPS) between two images, x and x_0. This metric quantifies perceptual differences by leveraging feature representations extracted from a deep neural network. The distance $d(x, x_0)$ is computed as a weighted sum of squared differences between feature vectors at corresponding spatial locations across multiple network layers. Specifically, for each layer l, the feature maps of height H_l and width W_l are considered, with the differences at each spatial position (h, w) being weighted by a learned vector w_l. The formula normalizes these differences by the size of the feature maps and emphasizes perceptually important features through the learned weights. The squared Euclidean norm $\|\cdot\|_2^2$ measures the magnitude of these feature discrepancies. By aggregating these perceptual differences across multiple layers, LPIPS provides a nuanced assessment of visual quality that aligns closely with human visual perception, making it a valuable tool for evaluating image inpainting techniques.

Qualitative assessment methods involve visual inspection, where human evaluators scrutinize the realism, coherence, and semantic consistency of inpainted images based on visual cues and contextual understanding. Semantic coherence is evaluated by examining whether the generated content aligns with textual descriptions provided as guidance. User studies, including perception studies and preference tests, gather feedback from human participants regarding the quality, naturalness, and usability of the inpainted images. These studies aid in identifying preferences for specific inpainting techniques and provide insights into user perception.

By integrating these diverse evaluation techniques, a comprehensive understanding of the performance and user perception of image inpainting models is achieved. This multifaceted approach facilitates informed decision-making and further refinement of the inpainting process, ensuring that the models produce high-quality, realistic, and contextually appropriate results.

4. Experimental Results

4.1. Results

Figure 10. Results of Models

In Figure 10, the evaluation of model performance across various metrics reveals notable distinctions. According to

the average SSIM score, the Attention U-Net model employing Mean Squared Error (MSE) at the 50th epoch exhibits the highest performance, achieving a score of 0.98. Conversely, based on PSNR score, the U-Net Like model utilizing MSE at the 50th epoch achieves the best results, with a PSNR of 37 dB. Furthermore, when assessed using the LPIPS score, the Attention U-Net model with Perceptual Loss at the 50th epoch outperforms others, with an LPIPS score of 0.0215. The difference in LPIPS scores from PSNR and SSIM scores highlights the varying degrees of perceptual quality achieved by these models.

The LPIPS metric, which measures perceptual similarity, may be considered a better metric compared to PSNR and SSIM for evaluating image quality in this context. While PSNR and SSIM provide valuable information about the fidelity and structural similarity of the images, they do not directly correlate with human perception. LPIPS, on the other hand, evaluates perceptual differences that are more aligned with human vision. This is particularly relevant when the goal is to assess the visual fidelity and perceived quality of images, where subtle differences that impact human perception may not be captured by traditional metrics like PSNR and SSIM.

Therefore, in tasks where human perception plays a crucial role, such as image restoration or generation, using LPIPS alongside traditional metrics like PSNR and SSIM provides a more comprehensive evaluation of model performance. It ensures that the generated images not only maintain fidelity to the original but also meet the expectations of human observers in terms of perceptual quality.

The first row of the figure demonstrates that the Attention U-Net model with MSE loss at the 50th epoch achieves the highest SSIM score, closely resembling the ground truth image. Conversely, in the second row, although the model with the highest SSIM score is ranked first, human visual assessments, as reflected by the LPIPS results, favor the third image as superior.

The evaluation of the models' performance based on average SSIM, PSNR, and LPIPS scores has yielded insightful findings. The highest SSIM score was demonstrated by the Attention U-Net model, which utilized Mean Squared Error (MSE) loss at 50 epochs, indicating superior structural similarity with the ground truth images. It was observed that the Attention U-Net is highly effective in preserving the original image's structural details during reconstruction or enhancement tasks.

Similarly, the U-Net Like model, also employing MSE loss at 50 epochs, achieved the highest PSNR score. A higher PSNR score generally indicates better image quality as it reflects lower error between the original and reconstructed images. Therefore, while the U-Net Like model excels in minimizing overall pixel-wise differences, the Attention U-Net's focus on structural integrity is evident.

Furthermore, the Attention U-Net model with Perceptual Loss at 50 epochs achieved the lowest LPIPS score, indicating the best perceptual similarity to the ground truth images. Lower LPIPS values denote better perceptual quality, aligning more closely with human visual assessments.

The visual assessment of the results, as depicted in the figures, corroborates these quantitative findings. In some cases, the results closely resemble the ground truth images, underscoring the models' effectiveness in maintaining high visual fidelity. For instance, the Attention U-Net model with MSE loss at 50 epochs in the first row shows a result that is very similar to the ground truth, as illustrated in the figure.

Overall, both quantitatively and qualitatively, the models' results surpass those of the pretrained model, as demonstrated across the metrics. This suggests significant advancements in model efficacy, highlighting the potential for enhanced image quality and fidelity.

4.2. Comments

Several strengths and weaknesses were encountered by the project, influencing the outcomes.

One notable weakness was the inconsistency in achieving satisfactory results that met the initial expectations. Despite efforts, instances occurred where the output did not align with desired benchmarks. Another significant setback was the limitation in resources, despite Google Colab Pro+ being rented for model training. The A100 GPU, while powerful, struggled to handle the extensive volume of data processed, hampering the ability to scale effectively. Additionally, challenges related to image quality during training were faced. Attempts to train images at a 512x512 size fell short of goals, impacting the visual fidelity of outputs.

Conversely, several strengths were demonstrated by the project. Satisfactory results were achieved in specific cases, as illustrated in examples from the previous slide. Good SSIM and PSNR scores were consistently delivered by the models, indicating strong performance within the constraints of resources. Moreover, pretrained alternative were outperformed by the models, as evidenced by superior PSNR and SSIM scores. This success highlights the effectiveness of the approach in enhancing image quality and fidelity.

Additionally, the project was distinguished by its use of a more generic dataset compared to the commonly employed Places2 and CelebA-HQ datasets. This approach enabled broader challenges to be tackled and real-world applications to be addressed more effectively. Furthermore, challenges related to perceptual loss, a common issue in similar works

utilizing VGG19 for perceptual loss, were encountered.

In summary, while challenges were encountered by the project, significant strengths were also demonstrated, and the groundwork was laid for advancements in image processing and machine learning. By addressing these weaknesses and building on these strengths, meaningful contributions to the field are aimed to be continued to be made.

Conclusion

In conclusion, this study has investigated the challenging task of image inpainting, focusing specifically on object removal, using two distinct autoencoder architectures: a traditional U-Net-like model and an enhanced U-Net with attention blocks. The models were trained on a large-scale dataset derived from the Defacto image and face manipulation dataset, which consists of approximately 25,000 instances of object-removal forgeries. This dataset provided a diverse and extensive resource for training and evaluating the models in the task of image inpainting.

Various loss functions, including Mean Squared Error (MSE), perceptual loss, and adversarial loss, were employed to train the models. The training process involved optimizing these models to accurately predict the missing portions of images based on surrounding context. Extensive evaluations based on quantitative metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), as well as qualitative visual assessments, were conducted to comprehensively assess the performance of the models.

The results indicate that the proposed models achieved superior performance compared to pretrained model. Specifically, the Attention U-Net model using MSE loss at 50 epochs achieved the highest SSIM score, demonstrating exceptional structural similarity to ground truth images. On the other hand, the U-Net-like model with MSE loss at 50 epochs attained the highest PSNR score, indicating minimal pixel-wise reconstruction errors. Additionally, the Attention U-Net model using Perceptual Loss at 50 epochs achieved the lowest LPIPS score, indicating the best perceptual similarity to the ground truth images.

Moreover, the effectiveness of attention mechanisms and perceptual loss in enhancing image inpainting quality is highlighted by these findings. The proposed models significantly outperformed pretrained model both quantitatively and qualitatively, offering substantial advancements in image processing techniques. These models demonstrated their capability to seamlessly integrate inpainted regions with surrounding context, preserving semantic consistency and enhancing visual coherence.

Furthermore, the study explored the impact of hyperparameters such as learning rate, batch size, and number of epochs on the performance of the models. Through systematic experimentation, the optimal combination of these hyperparameters was identified, ensuring robust and reliable model performance.

Future research directions could include exploring different architectures, deeper perspectives, additional loss types, and improvements in filling strategies to further enhance image inpainting capabilities. Moreover, investigating larger datasets and more diverse scenarios could broaden the applicability and robustness of image inpainting techniques in real-world applications.

In summary, this study contributes to the field of image inpainting by demonstrating the effectiveness of attention-based models and perceptual loss in achieving high-quality and visually realistic inpaintings. The findings underscore the importance of advanced deep learning techniques in addressing challenges related to object removal and image manipulation, paving the way for further advancements in image processing and computer vision. The outcomes of this study hold significant implications for various applications, including digital forensics, content creation, and artistic image editing, where accurate and reliable inpainting methods are crucial for maintaining integrity and enhancing visual aesthetics.

References

[1] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, Jiaya Jia, Learning Prior Feature and Attention Enhanced Image Inpainting https://openaccess.thecvf.com/content/CVPR2022/papers/Li_MAT_Mask-Aware_Transformer_for_Large_Hole_Image_Inpainting_CVPR_2022_paper.pdf

[2] Zuo, Zhiwen and Zhao, Lei and Li, Ailin and Wang, Zhizhong and Zhang, Zhanjie and Chen, Jiafu and Xing, Wei and Lu, Dongming. Generative Image Inpainting with Segmentation Confusion Adversarial Training and Contrastive Learning https://github.com/comzzw/Generative-Image-Inpainting

[3]CelebA-HQ Dataset https://www.tensorflow.org/datasets/catalog/celeb_a_hq?hl=tr

[4] Places Dataset https://www.tensorflow.org/datasets/catalog/placesfull?hl=tr

[5] Defacto-Inpainting Dataset https://www.kaggle.com/datasets/defactodataset/defactoinpainting

[6] Defacto Dataset https://defactodataset.github.io/