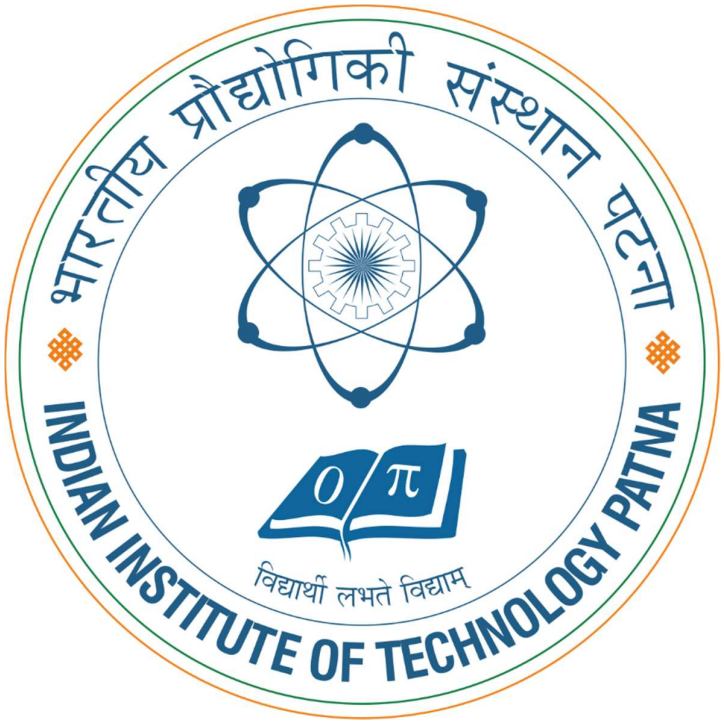
MAT: Mask-Aware Transformer for Large Hole Image Inpainting



BTP-1 Report 2024-25

By

Meet Patel [2101EE44]

Supervisor-Dr. Rajib K Jha

Table of Contents:

Objective:-

The primary objective of this project is to address the problem of image inpainting, particularly focusing on the challenging scenario of large hole inpainting. Image inpainting, also known as image completion, is a fundamental task in computer vision where the goal is to fill missing regions of an image with visually plausible content. This task has numerous applications, such as image editing, photo restoration, image re-targeting, and object removal. We will use a Mask-Aware Transformer (MAT) that combines the strengths of transformers and convolutions to efficiently process high-resolution images for large hole inpainting. The MAT framework combines the effectiveness of convolutional processes with transformers to describe long-range dependencies and produce outputs that are both diverse and of high quality. The proposed model introduces several key components, such as multi-head contextual attention (MCA) and a style manipulation module, aimed at improving the performance, stability, and diversity of the generated images.

Motivation:-

There are some limitations in the previous methods in handling large missing areas in images. Traditional approaches, such as those based on fully convolutional neural networks (CNNs), struggle to accurately model long-range dependencies, which are crucial for reconstructing large missing regions in complex images. The effective receptive field of CNNs grows slowly, which makes it challenging for them to establish semantic correspondences between distant areas. To address this limitation, prior research has attempted to integrate attention modules into CNNs to improve their ability to capture long-range relationships. However, due to the quadratic computational complexity of attention mechanisms, these methods have been applied only to small-scale feature maps, leading to coarse and suboptimal predictions for large holes.

Transformers, with their inherent ability to model non-local interactions, are better suited for capturing contextual information across an entire image. However, due to computational constraints,

existing transformer-based methods have only been able to generate low-resolution predictions, which compromises the quality of the final output, especially for large-scale masks. The need for a solution that can effectively handle high-resolution image inpainting led to the development of MAT. By customizing transformer blocks and introducing multi-head contextual attention, MAT aims to strike a balance between efficiency and effectiveness in modeling long-range dependencies for high-quality image reconstruction. The integration of a dynamic mask further allows MAT to selectively focus on valid tokens, making it computationally efficient while still capable of producing realistic and diverse outputs.

Design Methodology:-

The proposed MAT architecture consists of a convolutional head to extract tokens, followed by a transformer body with five stages of transformer blocks that model long-range dependencies using Multi-Head Contextual Attention (MCA). The output tokens are then processed by a convolutional tail to reconstruct the spatial resolution back to the input size. To further refine the details, a Conv-U-Net is used, enhancing high-frequency texture information. Lastly, a style manipulation module modulates convolution weights, enabling diverse and plausible output predictions.

Convolutional Head:-

The convolutional head takes in the incomplete image and its mask, producing feature maps reduced to 1/8 of the original resolution. It consists of four convolutional layers—one for adjusting the input dimension and three for downsampling. This design is used for two main reasons: it incorporates local inductive priors for improved representation and optimizability, and it facilitates fast downsampling, reducing computational complexity and memory cost.

Transformer Block:-

The transformer body in the MAT architecture is responsible for building long-range correspondences through token processing. It consists of five stages of adjusted transformer blocks that use an efficient attention mechanism guided by a dynamic mask.

The adjusted transformer block aims to enhance training stability, particularly when dealing with large hole masks. It removes Layer Normalization (LN) and instead of residual learning, employs fusion learning using feature concatenation.

We concatenate the input and output of attention and use a fully connected (FC) layer:

$$X'_{k,l} = FC\left(\left[MCA\left(X_{k,l-1}\right), X_{k,l-1}\right]\right),$$

$$X_{k,l} = MLP\left(X'_{k,l}\right),$$

where $X_{k,l}$ is the output of the MLP module of the ℓ-th block in the k-th stage.

This modification helps to stabilize the training process and avoids issues like gradient explosion, which can occur due to the high ratio of invalid tokens. Additionally, 3x3 convolutions are used to

provide positional information, removing the need for positional embeddings. This approach relies solely on feature similarity, promoting long-range interactions.

Multi-Head Contextual Attention:-

To manage a large number of tokens and handle low fidelity of given tokens, Multi-Head Contextual Attention (MCA) is introduced. MCA uses shifted windows and a dynamic mask to perform non-local interactions efficiently. Only valid tokens are considered in the attention calculation, and the dynamic mask is updated after each attention pass.

The output is computed as the weighted sum of valid tokens, which is formulated as

$$Att(Q,K,V) = Softmax\left(\frac{QK^T + M'}{\sqrt{d_k}}\right)V,$$

Where Q,K,V are the Query,key,value metrices and $\frac{1}{\sqrt{d_k}}$ is the scaling factor.The mask M' is expressed as:

$$M'_{i,j} = \begin{cases} 0, & if\ token\ j\ is\ valid \\ -\Gamma, & if\ token\ j\ is\ invalid \end{cases}$$

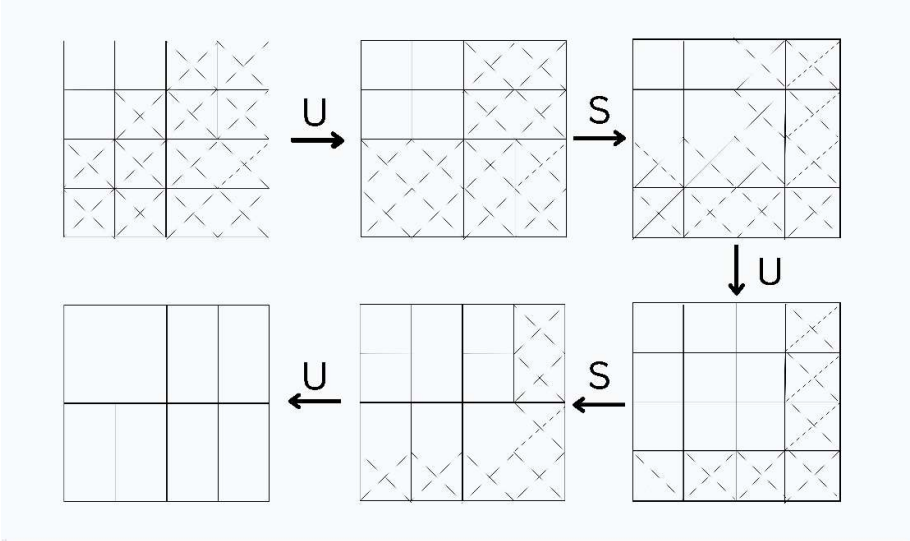
Where Γ is a large positive integer.

During the process, tokens in a window are marked as valid if there is at least one valid token present, helping the model progressively expand the effective attention area.

Mask Updating strategy:-

The mask (M′) points out whether a token is valid or invalid, which is initialized by the input mask and automatically updated during propagation. The updating follows a rule that all tokens in a window are updated to be valid after attention as long as there is at least one valid token before. If all tokens in a window are invalid, they remain invalid after attention. After several times of window shift and attention, the mask is updated to be fully valid.

For images dominated by missing regions, the default attention strategy not only fails to borrow visible information to inpaint the holes, but also undermines the effective valid pixels. To reduce color discrepancy or blurriness, we propose to only involve valid tokens (selected by a dynamic mask) for computing relations.



Style Manipulation Module:-

The Style Manipulation Module (SMM) in the MAT architecture is designed to enable pluralistic generation, allowing the model to produce diverse inpainting outputs by manipulating the convolution weights during the reconstruction process. This is achieved by incorporating both image-conditional and noise-unconditional styles. The style manipulation works by modifying the weight normalization of the convolutional layers during reconstruction using an additional noise input.

Conv-U-Net:-

The Conv-U-Net plays a critical role in refining the high-frequency details of generated images during the inpainting process. After the initial inpainting is performed by the transformer body, the output may lack fine details, and this is where the Conv-U-Net comes in. Its architecture combines convolutional layers with skip connections, which allows it to capture both low-level and high-level features effectively. This hierarchical structure is particularly beneficial for learning textures and structures essential for producing realistic inpainted images. The Conv-U-Net maintains high resolution throughout the process, ensuring that the final output matches the original image size and quality. Additionally, its ability to localize and incorporate contextual information helps preserve spatial hierarchies and details that may be lost in deeper layers. By processing the output from the transformer body, the Conv-U-Net enhances detail, improves overall visual coherence, and contributes to the final inpainting results. This combination of the transformer's capacity for modeling long-range dependencies with the U-Net's strengths in capturing fine details results in a powerful framework for large hole image inpainting.

Designed system diagram and photographs:-

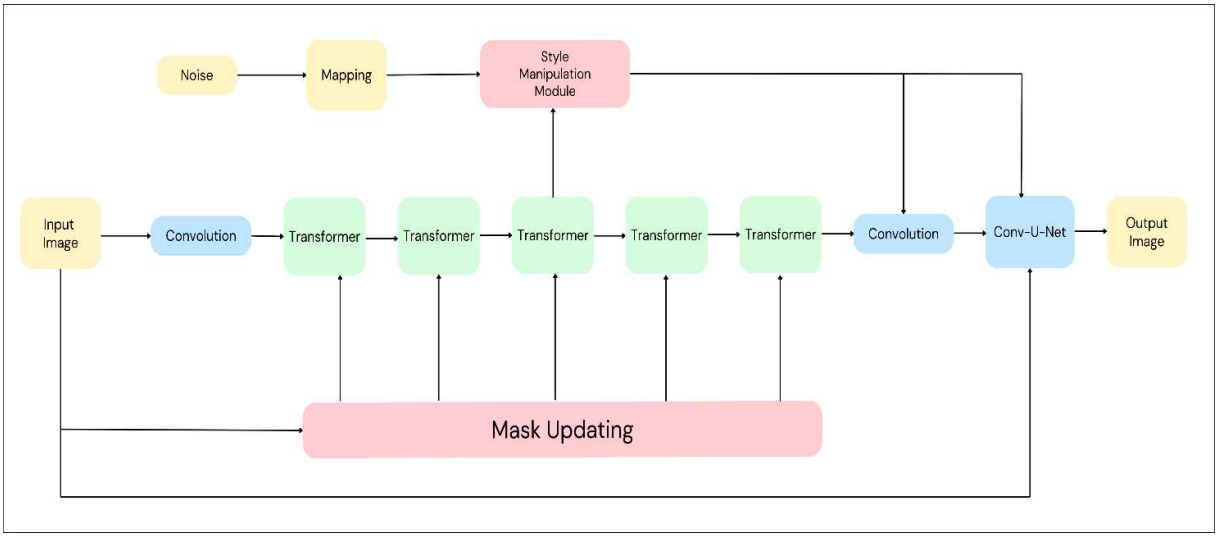


Figure-Overall flowchart

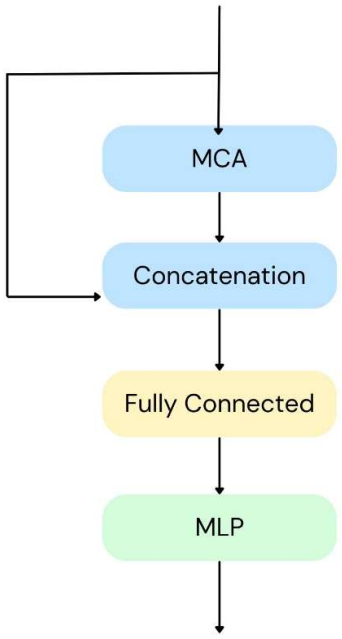
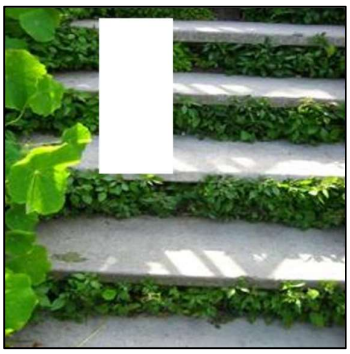


Figure-Adjusted Transformer Block

Outputs:-


Original Image


Mask+Image


Output Image


Original Image


Mask+Image


Output Image


Original Image


Mask+Image


Output Image


Original Image


Mask + Image


Output Image

Original Image

Mask+Image

Output Image

Original Image

Mask + Image

Output Image

Original Image

Mask + Image

Output Image

Original Image

Mask+Image

Output Image

Abalation Study:-

Models	FID	P-IDS	U-IDS
Full Model	**5.97**	**13.17**	**29.23**
Without Transformer	6.21	11.30	27.39
Without Adjusted Block	6.36	12.30	28.05
Without MCA	6.08	13.13	29.19

FID (Fréchet Inception Distance)

FID computes the distance between the feature distributions (extracted using a pre-trained Inception network) of real and generated images. It evaluates realism and diversity in generated outputs.It evaluates by comparing their distribution to the distribution of real images. Lower FID scores indicate that the generated images are closer to the real ones in terms of quality and distribution.

$$FID = ||\mu_r - \mu_g||^2 + Tr(\sum r + \sum g - 2\left(\sum r \sum g\right)^{1/2})$$

P-IDS (%) (Precision for IDS)

Precision measures the proportion of positive identifications (e.g., correctly matched pixels, regions, or features) that are actually correct. Represents the percentage of correctly matched or identified instances (e.g., objects, regions) in a dataset. Higher percentages indicate better precision in identifying or matching relevant elements in the images.

$$P - IDS = \frac{TP}{TP+FP} \times 100\ \%$$

U-IDS (%) (Recall or Utility for IDS)

U-IDS emphasizes how well the method captures all the relevant information (e.g., regions, objects, or features), reflecting the utility of the system. Measures the proportion of true positive instances detected out of all possible relevant instances in the dataset. Higher percentages indicate that the system successfully identifies more of the true positive instances.

$$U - IDS = \frac{TP}{TP+FN} \times 100\ \%$$

Conv-Transformer Architecture:-

We explore whether the long-range context relations modelled by transformers are useful for filling large holes. Replacing the transformer blocks with convolution blocks we find an obvious performance drop on all metrics, especially on P-IDS and U-IDS, indicating that the inpainted images lose some fidelity. Compared to the fully convolutional network, MAT takes advantage of distant context to reconstruct the image, showing the effectiveness of long-range interactions.

Adjusted Transformer Block:-

In this framework,a novel transformer block is used to address the instability issues commonly encountered with conventional designs, which often require reducing the learning rate of the transformer body. As shown in Table this design achieves significantly better performance, improving the FID score by 0.39 compared to other model which uses the original transformer block.

Multi-Head Contextual Attention:-

To quickly fill the missing regions with realistic contents, the model uses a multi head contextual attention (MCA). To make a deeper understanding, a model without partial aggregation from valid tokens. It is noted that FID drops by 0.1 yet other metrics do not change too much. However, it is helpful for maintaining color consistency and reducing blurriness.

Conclusion:-

The MAT framework seamlessly integrates transformers and convolutional neural networks to effectively model long-range dependencies, enabling the generation of high-resolution inpainted images. By incorporating a customized multi-head contextual attention mechanism, the framework enhances inpainting efficiency by concentrating on valid regions of the image, ensuring that the inpainting process is both precise and resource-efficient. Additionally, the inclusion of the Conv-U-Net module plays a critical role in refining high-frequency details, which is essential for producing visually realistic and coherent outputs. Together, these components establish MAT as a robust and innovative solution for image inpainting, combining the strengths of advanced attention mechanisms and convolutional architectures to deliver superior performance and high-quality results.

References:-

[1] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, Jiaya Jia, MAT: Mask-Aware Transformer for Large Hole Image Inpainting.In CVPR 2022.