combined

rounak Gera

August 2024

1 Introduction

Image inpainting is a process of reconstructing the missing regions of an image requiring a strong understanding of image structure and semantic understanding of the image for eg: If you have an image with some burned areas, you can correct it using image inpainting. It aims at the realistic filling of missing parts of an image. It has a vast number of applications in image restoration, object removal and replacement etc.

It has multiple approaches traditionally it was majorly done by using nearby neighbouring pixels to estimate the value of the current pixel whch struggle alot with large masks and difficult textures. But with the upcoming deep learning techniques, it is completely revolutionised with more plausible, context-aware and high-resolution filling of missing regions. Earlier CNNs were used for such computer vision-based problems then they were succeeded by GAN's which provided alot more better outputs due to their generator discriminator architecture but soon with the success of transformers in natural language processing, it has also shown its ability to produce excellent results in computer vision tasks, especially in problems where contextual awareness is needed like in image inpainting problems.

But it still faces two major issues which are a small receptive field and non-realistic and inappropriate filling of missing parts which are solved by two methods i.e. LAMA and MAT where LAMA provides an approach to increase the receptive field by using FFC(Fast Fourier Convolutions) based on converting the inputs to the frequency domain and MAT proposed a transformers and partial attention based method capable of generating high-resolution outputs for large masks.

In this paper, we combine the methods of LAMA and MAT using a third combiner model based on UNET architecture and a denoiser model to remove noise generated during inpainting. We show an approach based on a mix of transfer learning and ensemble learning to achieve high efficiency without high computational requirements. But as it is known that LAMA and MAT both are quite heavy we use a step-by-step approach to perform inference of our ensemble model using a GPU of 16GB RAM. Our contributions can be summarized as

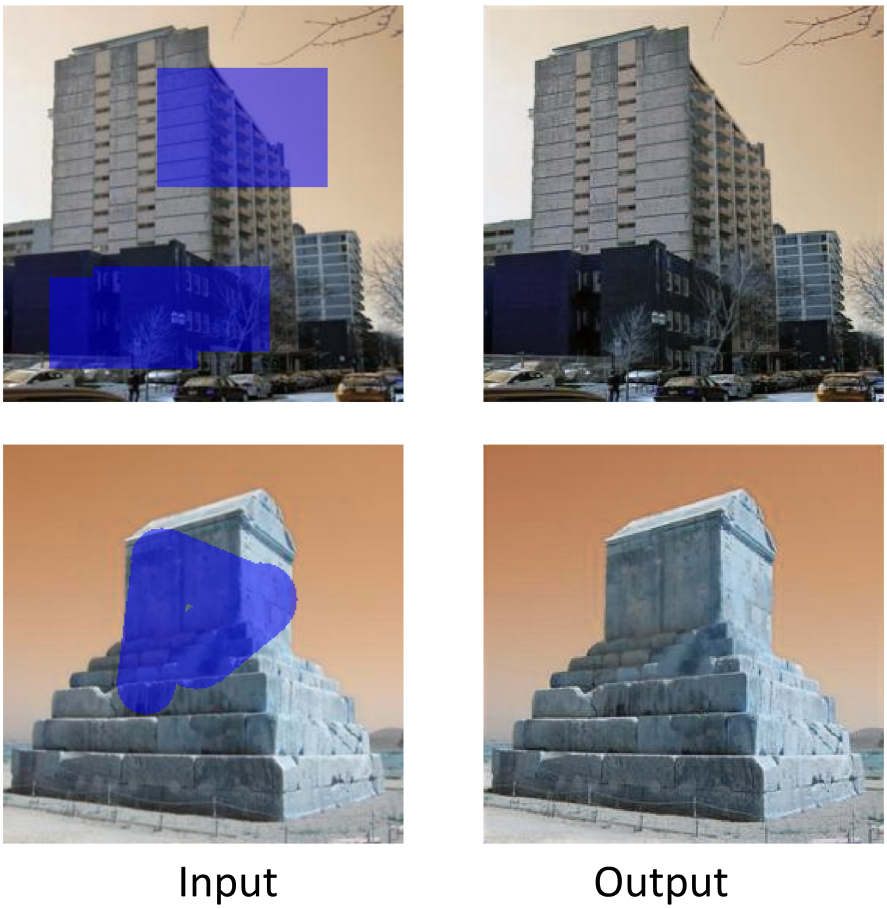

Figure 1: This figure shows the robustness of our approach in filling complex patterns and textures in images as shown in these images taken from Places365 dataset

(i) Making an ensemble model that combines both LAMA and MAT along with making use of denoisers to improve the results.
(ii) Step-by-step approach to train the ensemble model in a GPU of 16 GB RAM.

1.1 Motivation

Image inpainting has emerged as one of the most crucial tasks in field of computer vision, as it aims to restore the corrupted or missing region of an image by plausible filling while keeping the structure and meaning of image intact as the original one having a wide range of application ranging from object removal, object insertion, image editing, image restoration and many others. The current deep learning based methods have evolved alot to tackle the problem from simple CNN's to GAN's and now to transformers and diffusion models. But still there are many challenges being faced by these models like:-
1. Difficulty in restoring complex textures and patterns
2. Over-smoothing of restored parts
3. Filling by inappropriate content changing the original meaning of image

But all these challenges are primarily due to two causes small receptive field which make these models unable to capture image wide patterns and textures and lack of understanding of image.

These challenges motivated us to provide an approach which tackles these in a efficient way without use of much computational power. And since all these challenges are tackled by different papers i.e. LAMA and MAT so we aimed to make a combined approach to make the best use of both methods using ensemble learning. By the way of our research we aim to provide a robust method for image inpainting.

1.2 Literature Survey

In this sub section, we will discuss about recent researches about image inpainting.

1.3 Research Questions

1.4 Paper organisation

The paper is structured as following sections:-

Section 1: Introduction provides a brief introduction about image inpainting, traditional approaches and their problems, along with recent advancements in image inpainting methods including evloution from CNN's to GAN's and then to transformers, Then we specify the motivation of our research, then we explore some of the prevalent research questions and finally we explain the paper's organizational structure.

Section 2: Preliminaries provide a introduction to all the neccesary concepts needed by the reader to know for proper undertanding of paper, including preliminaries about ensembling, bagging, boosting, transfer learning, denoising models, UNET architecture , LAMA model and MAT model(which we used as base for our approach)

Section 3: Proposed Approach explains in detail about the dataset including samples from dataset and other properties of dataset , detailed explanation about how we used ensembling in our research, then we explain about the role of transfer learning and way we used it, then we explain about the combiner model which is used to combine the output of both LAMA and MAT model to provide the best results, then we explain about the denoising model, its working and where, why and how we used it to improve the results of our approach, then we tell about the overall workflow of our approach, then we explain about the loss functions we used and why we used them for purpose of calculating the overall loss while training , then we explain about the way we prepared the data so it can be fed to our model and at last we gave a overall step by step procedure of what all steps we took.

Section 4: Experimentation in this we have discussed about the various experiments like choosing the correct inpainting models for using a base of ensembling, than choosing the ensembler model and experimentation with the overall architecture.

Section 5: Comparison in this we shared all the results and comparisons with other papers and approaches like MISF, MxT and so on to show the relative performance and positioning of our model , including plotted graphs for better understanding of comparisions.

Section 6: Summary in which we provided conclusion of our approach and about the scope of future improvements in our research.

2 Literature survey

this section contains brief information about the literature survey we had done for our paper as shown in table1.

3 Preliminaries

3.1 Ensembling

It is a technique of combining two or more models in order to get better results and improve the efficiency of model by combining the power of multiple models. It is primarily of two types i.e. bagging and boosting. It combines the model with the goal of reducing either bias or variance or both to improve the generalization capabilities of the final model.

S. No.	Title
1.	HINT: High-quality Inpainting Transformer with Mask-Aware Encoding and Enhanced Attention (
2.	Learnable Prompt for Few-Shot Semantic Segmentation in Remote Sensing Domain(open earth ma
3.	Towards Context-Stable and Visual-Consistent Image Inpainting (IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024)
4.	A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpaint
5.	Text Image Inpainting via Global Structure-Guided Diffusion Models (AAAI Conference on Artificial Intelligence (AAAI) in 2024)
6.	GraphFill: Deep Image Inpainting using Graphs (IEEE/CVF Winter Conference on Applications o

Table 1: table showing top 6 papers of the image inpainting literature survey we had done showing authors, title, dataset, domain work done, limitations and USP

3.2 Bagging

It also called bootstrap aggregating is a type of ensemble learning where multiple models are trained in parallel and average of their outputs is taken to return the final output. In case of regression problems usually the final output is obtained by calculating average of all models output while in case of classification problems the mode or most occurring output is decided to be the final output. In this all the models are trained on different subset of data. It helps in reducing variance. Its popular example is random forest which combines multiple decision trees. In this models are ran in parallel order. It can be expressed as

$$\hat{y} = 1/M \sum_{m=1}^{M} f_m(x)$$

where f_b will be the b th model and B refer to total number of models being used and this formula is used to have average prediction from all models.

3.3 Boosting

It is a type of ensemble learning where multiple models mostly weak learners are ran sequentially with each model learning from mistakes of previous models. It have three main algorithms i.e. Adaboost(adaptive boosting), gradient boosting and XGBoost. It helps in reducing both bias and variance. In this models are ran in sequential order i.e. one learning on errors of previous one.

$$\hat{y} = \sum_{m=1}^{M} \alpha_m f_m(x)$$

where f_t refer to t th model and α_t refer to weight of respective model.

3.4 Stacking

It is a type of ensemble learning which is quite different from bagging and boosting as it allows for usage of different base models with each having ability to tackle certain aspect of problem like in our case LAMA and MAT both models focus on tackling specific aspects of image inpainting so by combining them by stacking helps in merging both's power.

3.5 Transfer learning

It is a technique where pretrained model on one task is fine tuned to be used on any other custom task so as to avoid all the time and cost of training the model on custom case. We can express it as

$$y = f_{pretrained} + O_{custom}$$

where $f_{pretrained}$ refer to the pretrained model and O_{custom} refer to the finetuned parameters for our custom case.

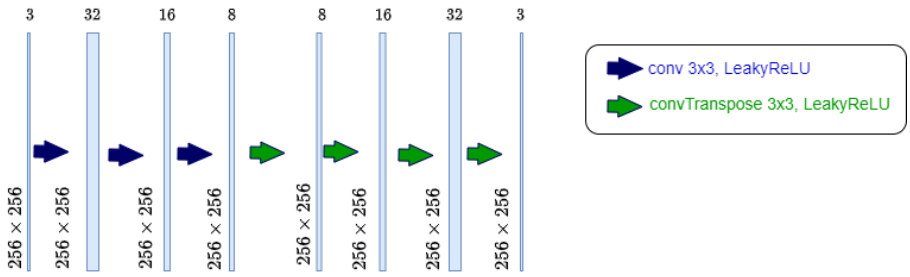

Figure 2: the architecture of denoising model used in our approach.

3.6 Denoising models

These are models which are used to obtain a clean image from a noisy image generally denoising autoencoders(DAE) are used for such task which consist of an encoder-decoder architecture. Let x_{noisy} represents the noisy image and x_{clean} refer to desired clean image and denoising model learns a mapping such that $f(x_{noisy}) = x_{clean}$ by minimizing the following loss function

$$L_{denoise} = ||x_{clean} - f(x_{noisy})||$$

Figure 2, shows the architecture of denoising model used in our approach.

There common approaches include denoising autoencoders(The one we used in our approach), CNN based denoisers, etc. While these days its primarily done by deep learning models but earlier there were many traditional approaches for doing this although not much effective like non local mean, wavelet baed method etc. The efficiencyof denoising models/approaches is measured by metrics like PSNR(Peak Signal to Noise ratio) and SSIM.

3.7 LAMA

It is a state of the art inpainting architecture which uses FFC(Fast Fourier Convolutions) with convolutional layers to solve the problem of low receptive field of inpainting architectures due to which they are not able to recognise global patterns and face difficulty in complex structures. In its architecture it FFC to capture global patterns by converting the image from spatial to frequency domain and it uses batch normalisation layers, with ReLU activations. For loss function it uses a combinations of high receptive field perceptual loss and adverserial loss to properly measure the loss. The loss function is

$$\mathcal{L}_{final} = \kappa L_{adv} + \alpha \mathcal{L}_{HRFPL} + \beta \mathcal{L}_{DiscPL} + \gamma R_1$$

where L_{adv} is adverserial loss, $\mathcal{L}_{HRFPL}$ is high receptive field perceptual loss and $\mathcal{L}_{DiscPL}$ is discriminator based perceptual loss.

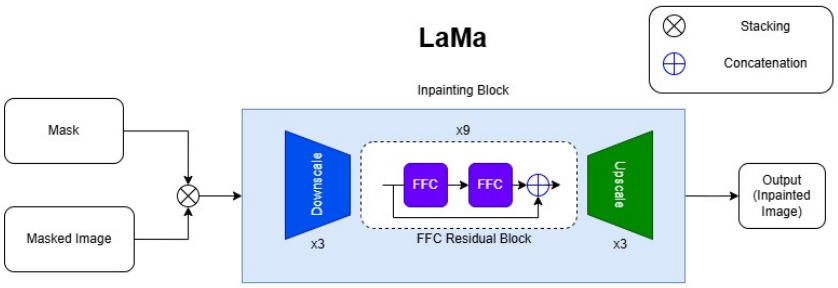

Figure 3: LaMa architecture showing stacking of masked image and mask, passing it to inpainting block where its first downsampled then passed through a FFC residual block and then again upsampled to return a high resolution inpainted image

Additionally it generates masks during the training phase i.e. it gets a real image and super imposes it with randomly generated mask and than pass it to model to train it.

The figure 3, shows the architecture of LAMA.

3.8 MAT

It is a state of the art inpainting architecture which tackles the challenge of large masks and high resolution images using a mix of transformers, convolution layers and multi head contextual attention mechanism. Also it tries to increase the fidelity by carefully designing the architecture.Also it introduces a a style manipulation module(SMM) for tackling the large masks. For loss function it uses a combination of perceptual and adversarial loss.

$$\mathcal{L} = \mathcal{L}_G + \gamma R_1 + \lambda \mathcal{L}_p$$

where $\mathcal{L}_G$ is generator loss and $\mathcal{L}_p$ refer to the perceptual loss.

While training it adopts a strategy of updating masks for preventing the attention mechanism from failing.

Figure 4, shows the architecture of MAT.

3.9 U-Net

It is a convolutional neural network(CNN) primarily developed for image segmentation consisting of a encoder, bottleneck and decoder with residual connections. In this the encoder transforms the image to a latent representation to capture information and later it is reconstructed by decoder block.Let $E(x)$ be the encoder block with x image input and $D(z)$ be decoder block with z as latent representation of image formed by encoder block, then UNET can be

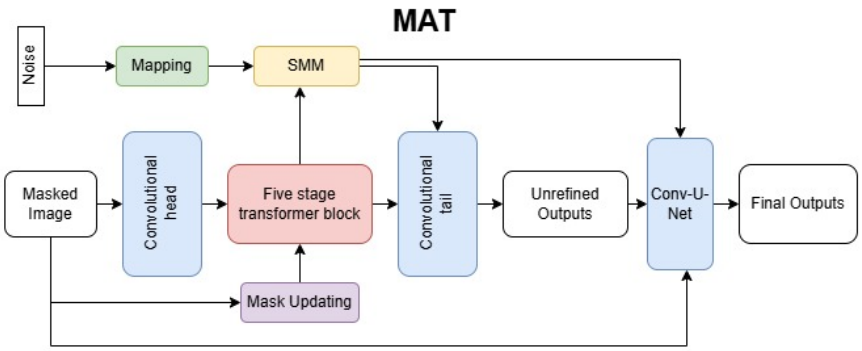

Figure 4: Showing MAT architecture where first of all the masked image is passed onto convolutional head, then transformer block along with size based mask updating for each transformer then the convolutional tail upscales to return the results, finally a Conv-U-Net is used along with SMM(style manipulation module) for returning a refined output

expressed as

$$y = D(E(x))$$

UNET usually use cross entropy loss due to their primary task of image segmentation

$$L_{crossentropy} = -\sum y - log(q(x))$$

where y refer to ground truth and q(x) refer to the prediction given by the model.

Figure 5, shows the architecture of a U-NET as used in our approach.

4 Proposed Approach

Our methods solves the 16 GB GPU RAM constraint by a step by step process where the whole process is distributed into seprate parts to make the best use of GPU along with ensembling the LAMA and MAT model and also additionally utilizing the denoising models to improve the performance and quality of outputs.

4.1 Dataset

For the dataset we have used Places365 dataset, it contains images from 365 different scenes from various places including scenes like airport, baseball field,auditorium etc. Due to its variety of scenes of varying and difficult textures, we choose it. Although its primarily use for classification tasks but we easily transformed it for our usecase.

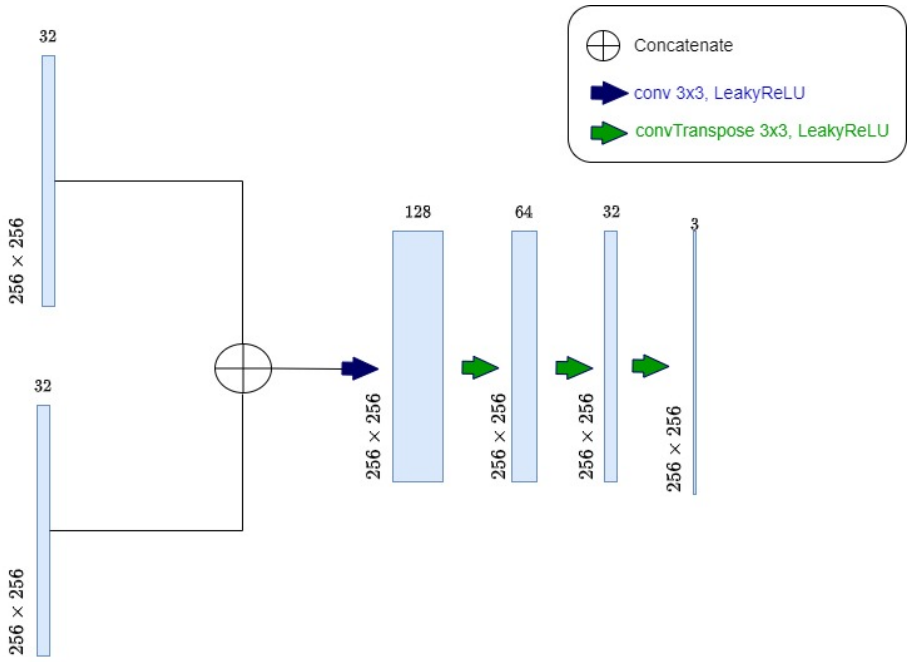

Figure 5: architecture of variation of U-NET used in our approach

Dataset	No. of Images	No. of images used
Train-standard	1.8 Million	7,300
Train-challenge	8 Million	-
Validation	36,500	36,500

Table 2: Table showing different version and splits of Places365 dataset we used

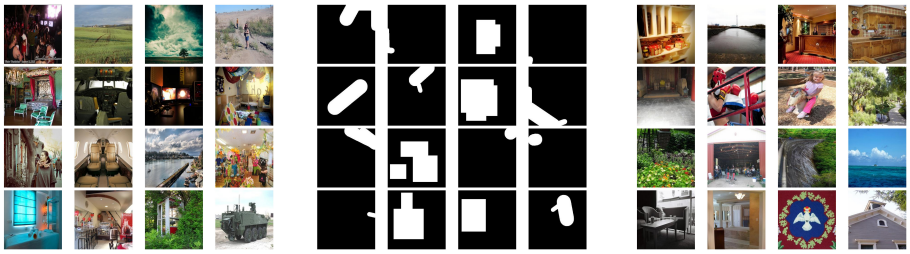

Figure 6: The figure shows the random samples of images from training and validation set a) shows random samples of ground truth images from training set b) shows random samples from masks of images of validation set and c) shows random samples of ground truth images from validation set

There are three versions of Places365 dataset, first one is **train-standard** containing about 1.8 million images, **train-challenge** containing about 8 million images and finally **validation** set contains 36,500 images.

For purpose of training, we used train-standard set's first 20 images from each type of scenes summing upto total of 7300 images i.e. 20 images from 365 scenes each and for evaluation purpose we used the complete evaluation set.

For using this dataset for our image inpainting purpose, we use the mask generation technique illustrated by LAMA's authors, then we applied the masks on all images to obtain the masked images and the original images served as ground truth for us.

In figure 1, we have shown some ground truth images from training set, random masks of images from validation set and some ground truth images from validation set.

In table 1, we have shown various versions and splits of Places365 dataset we used

4.2 Ensembling

For ensembling both the models there were two alternative i.e. boosting or bagging but since bagging would have involved parallel working of both models so it would have easily broken our memory constraint while boosting would have fitted into our memory constraint but the fact that there is n't any best way to measure the error which could satisfy our purpose , So to choose a middle path

, it was decided to use **stacking** that first the input will be processed by first model and then after clearing the memory properly the same input file will be processed by second model and the memory will be again cleared and later on the outputs of both models will be processed and combined by denoising and combiner models.

4.3 Transfer Learning

So although the models were now in our memory constraint but still there was one problem that the training time was too much even when it was decided that the models will be trained on only a 25 percent subset of Places365 standard train dataset which consist of 1.8 million images but still the training time was too much to be done on google colab since the models are quite deep and complex and also the number of models are more so to solve the problem it was decided to opt for transfer learning and it proved to be a great. Where the pretrained models of the original authors of models(LAMA and MAT) were used to avoid training which also helped in improving the performance. Although for denoising models and UNET based ensemble model were trained due to unavailabilty of trained weights for similar task , additionally these models were much lighter than LAMA and MAT so it was decided to train and they were trained on 7300 images consisting of 20 images from each scene category from Places365 standard train dataset.

4.4 Combiner model

Finally a model was made which could combine both the models results which was similar to a classic U-Net without residual connections consisting of three downsample blocks and three upsample blocks along with ReLU as activation layer.But the model seemed to make very little progress while training and model seemed to overfit which was easily avoided by changing the activation functions from ReLU to LeakyReLU. But still the outputs of the ensemble models were not upto mark and it was observed that there was some noise on output produced by LAMA model so it was decided to use a denoiser model on LAMA's output and then the ensemble model was run on its output and MAT's output but this time the final output contained some noisy patches which was quite easily solved by using another denoising model of same architecture. Then the final output was ready.

4.5 Denoising Model

Denosing models played a pivotal role in improving our final output, at first instance the denoising model helped to improve the output produced by LAMA model which was fed to our ensemble / combiner model and at second instance denoising model helped to correct the random blocks of noise in output of combiner model and then the final output was made. The denoising model is a convolutional neural network(CNN) consisting of encoder and decoder block

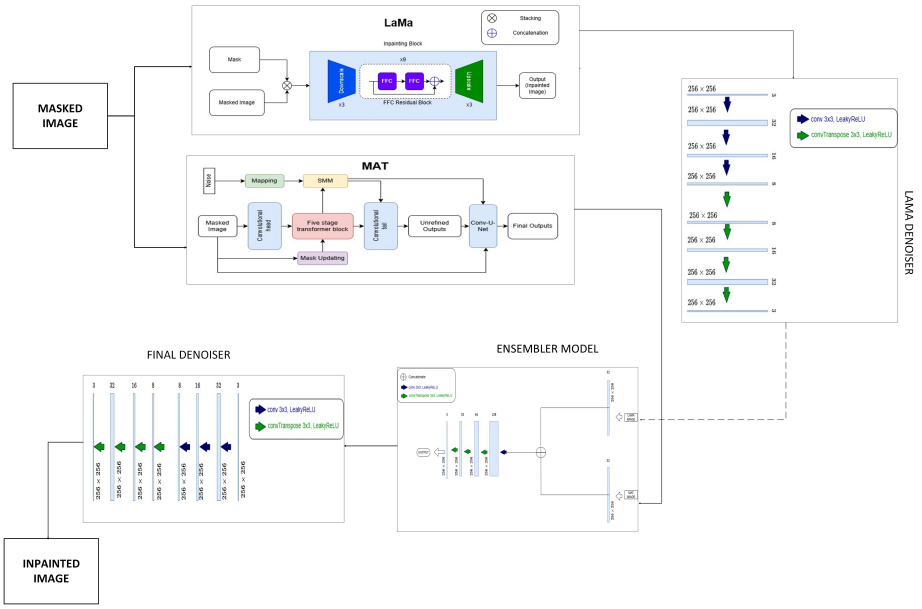

Figure 7: The figure illustrates the workflow of our technique, starting with the masked image which is fed in parallel to both LAMA and MAT model, then the output images of LAMA are fed to LAMA DENOISER, then both the MAT and LAMA DENOISER images are fed to combiner model and finally the resulted image is passed on to a FINAL DENOISER for a final finishing touch and to obtain the final output image

where the encoder block consisted of three convolutional layers along with LeakyReLU for downsampling and decoder consisted of three convtranspose with LeakyReLU for upsampling and finally a sigmoid activation function at end.

4.6 Workflow

First of all the LAMA model was used to and it was loaded with the pre-trained weights and inference was ran over validation set of Places365 dataset which was processed according to the requirements of model then the resulting images were stored for later use in combiner model, also the masks were also stored to ensure uniformity across outputs of both models. Then the same process was repeated for the pre-trained MAT model and its results were also stored and it is important to note that the masks were same. Then the output of LAMA model were passed on to our previously trained denoising model and then the output of MAT and denoising models were fed to combiner model then the resulting output was again passed on to our other previously trained denoising model to generate our best outputs.

One small issue was that the saved masks were being loaded as RGB images and had values in range of 0 to 255 which was inappropriate for our purpose of masking the images which was solved using threshold function of opencv.

Equation :-
Y = Final denoiser (Combiner Model(Lama Denoiser (LAMA (X)), (MAT (X)))

4.7 Loss Function

For purpose of computing loss for training our ensemble model four different loss functions were used to get the best measure of loss. Firstly Style loss was used for analysing texture, patterns along with capturing complex stylistic features far beyond simple color distributions of images. Next Edge loss was used which help to ensure sharpness and clear edges which further help in maintaining the structural integrity of objects in image. Next Perceptual loss based on pre-trained VGG model , it was used for comparing feature similarity ensuring structural and semantic similarities and at last MSE loss was used to measure differences at pixel level.

$$\mathcal{L}_{total} = \lambda_{style}\mathcal{L}_{style} + \lambda_{edge}\mathcal{L}_{edge} + \lambda_{perceptual}\mathcal{L}_{perceptual} + \lambda_{MSE}\mathcal{L}_{MSE}$$

For purpose of training our denoising models only MSE loss was used due to simplicity, effectiveness and proven record of good results for denoising task as it measures difference at pixel level and in case of task like denosing it is much needed.

$$\mathcal{L}_{DenoisingModel} = \mathcal{L}_{MSE}$$

4.8 Data Preparation

For training of combiner model and denosing models we used a subset of Places365 standard train set where we took 20 images from each scene category totaling to 7300 images.

For evalutaion we use the validation set of Places365 dataset consisting of 36,500 images, then we remove all the images which don't have all the three color channels and also we resize all the images to 256X256 size to ensure uniformity in images. Also, we take care of the fact that even masks are read as color images and have values between 0 to 255. So by using the threshold function we convert all the values to 0 and 1's. Then we convert the images to tensors using a transform function.

4.9 Step by step procedure

1. Get the data of Places365 validation set and preprocess it as explained in above sections
2. Generate masks using the technique introduced in LAMA paper and store them for later use and make masked images .

Figure 8: The figure illustrates the various phases of our approach's working firstly there are ground truth and mask which is applied on ground truth to create masked image which is processed by LAMA and MAT then the LAMA's output is denoised and then both MAT's output and denoised LAMA are combined to make semi prediction then applying the denoiser model again provides us with the final output image

3. Perform inference of pre-trained LAMA model on our masked images and store the results .
4. Repeat the step 3 for pre-trained MAT model.
5. Now apply the LAMA denosing model on the resultant images of LAMA model.
6. Feed the converted images and MAT images to Combiner model .
7. Now apply final denoising model on the resulted images to obtain the final and improved output images.

5 Experimentation

In this section we will tell you about the various experiments we had done for finding the best approach for image inpainting. Talking about the base models which in our case are LAMA and MAT but we had a wide choice for them also like HINT based on masked pixel downshuffle and spatially activated attention layer, Graphfill based on utilizing graphs neural networks and FFC, GSDM based on diffusion and SPM(structured prediction models) and many more. So for choosing the models we carefully analysed all the possible combinations along on various parameters like their USP, computational requirements, problem they were solving, inter compatibility etc. These all lead us to choosing LAMA and MAT as the best possible combination.

Talking about the ensembler model we had a wide range of options to choose from like attention based fusion, Multi pixel downshuffle based ensembling, multi scale ensembling that to of two types i.e. feature map based and resolution based and list go on. So we tried and tested a set of techniques which we believe could help and at the end we choose a U-NET based ensembler due to its best performance. We have illustrated the performance of all different experiments in table below on three metrics: LPIPS, SSIM and FID. This is shown in table

ARCHITECTURE	LPIPS	FID	SSIM
Feature based multi scale ensemble	0.9041	65.63	0.03
Resolution based multi scale ensemble	0.6091	60.23	0.18
Masked pixel downshuffle based ensemble	0.1120	35.6	0.892
UNET based ensemble	0.0194	19.23	0.928

Table 3: This table illustrates the performance of various ensembling architectures on metrics, All these methods were applied directly on LAMA and MAT outputs to get a fair judgement

	LPIPS	FID	SSIM
UNET ensemble	0.0194	19.23	0.928
Denoised LAMA with UNET	0.0204	18.43	0.9398
Final output	0.01294	0.9873	0.9764

Table 4: This table illustrates the performance of various stages of our architecture development firstly we just combined LAMA and MAT with UNET based ensembler, then we applied denoiser on LAMA and in the final output we combined denoised LAMA and MAT with UNET and then also applied denoising model.

1.

Talking about overall architecture after choosing UNET as our ensemble model and LAMA and MAT as base models, we yet had to choose for what is to be done to improve the outputs even more like we had decided to go for image harmonization to enhance the matching of recreated region and exsisting region but that does n't helped since our architecture was already handling it well then we decided that we need to use denoising models, firstly we applied it on LAMA's output as it was noisy which helped in improving outcomes and then on observing we noticed due to this our final output was being noisy also so we again used a denoising model which helped us even more. This is shown in table 2 and figure 4.

6 Comparison

In this section we illustrate the improved performance of our model in comparison to other baseline models and show that how the combination of LAMA and MAT provides a much more robust and improved results. This is shown in table 3. We have compared our model's performance with publically available inpainting models on Places365 validation dataset on LPIPS, FID, P-IDS and U-IDS.

Our model beats all the baselines models in all metrics with closest competitor being LAMA, In FID which is better if its less, our model beats LAMA by small difference and rest all by huge and in case of LPIPS which is better to be less, our model reduces it to more than 50% than that of LAMA. Talking

	FID↓	LPIPS↓	P-IDS↑	U-IDS↑
Ours	0.9873	0.0129	22.71	38.26
LAMA	0.99	0.0354	13.09	32.29
CMGAN	1.628	0.189	20.96	37.42
CR-FILL	9.657	0.233	5.53	22.90
CoModGAN	2.92	0.111	19.64	35.78
ProFill	7.7	0.230	3.87	21.19

Table 5: this table shows the performance of our model in relation to other baseline models which clearly tells the better results of our model

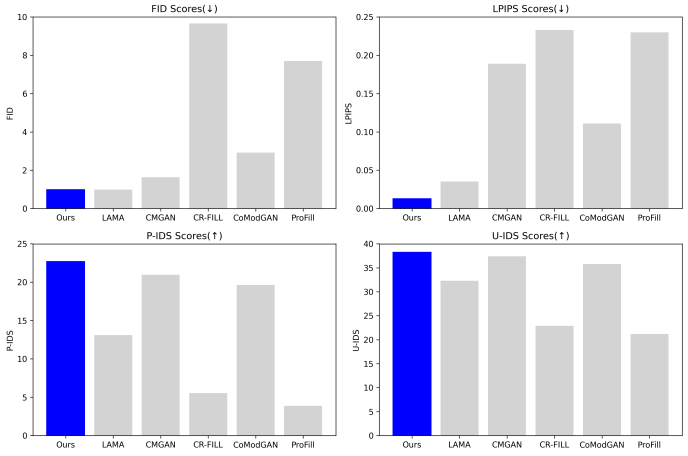

Figure 9: Figure showing the competitive performance of our model with all other baseline models

about P-IDS and U-IDS which are good if more our model again out performs all the rest models although by an approximate improvement of 10% in both cases which showcases the robustness of our model.

7 Summary

In recent times various techniques in field of image inpainting are introduced for solving problems like contextual ambiguity, visual inconsistency, low resolution, low receptive field and many other. With most of the techniques using transformers and diffusion as base. While quite innovative techniques are also introduced including using graph based approach , using signal processing techniques for image inpainting. Also some techniques have introduced multi-modals or multi-task generative models which can do a wide set of tasks like, object removal, text guided image [**?**], context aware, and shape guided inpainting.

While GAN's and Variational Auto Encoders (VAE) which were used inten-

sively for tasks like image inpainting have seen a downfall in their usage after the introduction of transformers and diffusion. Where transformers excel in semantic understanding of images and diffusion excel in generating high quality and realistic images.

For text guided inpainting most of the techniques use learnable prompts which guide the model to generate content as per instructions. Also some techniques have used coarse to fine approach where a coarse network fill the missing regions which is later refined by a fine network to produce high quality results. Some techniques have used edge detection which provide a rough structure of image to help the model in generating contextually stable results.

Most of the techniques use perceptual loss, style loss, total variational loss and adversarial loss. Commonly used evaluation metrics include SSIM(Structural similarity index measure), FID(Fréchet inception distance) , PSNR(peak signal-to-noise ratio) and LPIPS(Learned Perceptual Image Patch Similarity). Commonly used includes Places365 and CelebA but there are very less models trained on large scale datasets like LAION-5B.