Image Colorizer
ITCS 4152/5152

Anirudh Cheruvu
UNC Charlotte
acheruv1@uncc.edu

Hamilton Sheppard
UNC Charlotte
hsheppa1@uncc.edu

Kevin Mai
UNC Charlotte
kmai2@uncc.edu

Abstract

UNC Charlotte consists of an extensive archive of pictures from the University Archives which consists of black and white images that showcase the history and architecture of the University. Bringing these pictures into a more relatable light enables the present generation to reconnect and engage with the picture rather than looking at a black and white picture that the newer generation will not find interesting. Our plan is to transform these old images that are in black and white into an RGB color format. Although existing diffusion colorizer models exhibit good performance, we aim to fine tune them using the present UNCC pictures (converted to Black and White) and get results more aligned to that of UNCC's architecture and the color theme.

1. Introduction

Image colorization, is the process of adding many colors to existing black and white images that we adore or would want to bring to life. Image colorization models aim to understand the spread of color and interact with the given space of an image, where it may be the environment or buildings. There has been many techniques that have been developed to handle tremendous amounts of data sets.

Prior works have studied colorization of images. In GAN-Based image colorization[1], they proposed the use of conditional GANS for colorization of images. This is a self-supervised learning approach for image colorization. Farella et al.[2] shows how past aerial images can be colorized using deep learning with different models trained on the dataset. In Learning Diverse Image Colorization[3], their goal is to diversify intrinsic problems of colorization to produce multiple colorization's to display images. Using their model they display diverse colorization's than a CVAE model and cGANS model.

In this work, called the Image Colorizer we biased the model around The University of North Carolina at Charlotte, where we made the base prompt include the things around campus. We compare pre-existing models and fine tune the one we build to get the best result we can, through rigorous testing and coding. Of course, there are many colorizers on the internet that colorizes images. We evaluate three state of the art models. Our fined tuned model exhibited good performance.

Here are the following contributions of this project:

1. We search and source 50 archival images of UNC Charlotte campus, dating back from 1960's and 200 modern images from 2022 through multiples sources as well as capture some images from the campus ourselves and create a dataset.
2. We test the performance of base stable diffusion model(1.5) [4] to colorize the original black and white images and then apply super-resolution [4] over the images to re-scale them to higher resolution. Then we retest the performance of the stable diffusion model over the rescaled images.
3. Then, we use the controlNet [5] for our pipeline to control diffusion models by adding extra input condition i.e, a canny image of the input image to help the model understand the images better and run 8 tests to find the optimal parameters such as input prompt, strength, negative prompt, guidance scale, number of iterations.
4. We now use LoRA finetuning [6] to fine tune the stable diffusion model using 177 modern images with an added text prompt and saved 13 checkpoints of the model over 6500 inference steps.
5. We now run a total of 18 tests across 3 specific checkpoints, i.e 500, 3500 and 6500 to find the right input parameters to get the best suited colorization on the 50 black and white images and also run 7 tests on each of the 3 the checkpoints over the 20 images in the validation set for metrics.
6. And lastly, we employ Fréchet Inception Distance, Learned Perceptual Image Patch Similarity, Peak Signal-to-Noise Ratio, and the colorfulness metrics to evaluate our fine-tuned model against competing models: our base stable diffusion model; two tests of our base stable diffusion model with controlNet applied to our pipeline; and a state-of-the-art NoGAN diffusion model over 20

validation images not overlapping with the 177 modern images used for fine-tuning.

2. Data Collection

At the start of the project, we needed to collect images to fine-tune. We wanted to collect Black and White images and colorized images to use in our model. We looked to find the best results and into many sites, we compiled our findings from the best sources we could find.

2.1. Old UNCC images

We first needed to find Black and White images so we sifted through UNCC archives which are from goldmines.charlotte [7] and findingaids.charlotte [8] these websites are owned by UNCC. In UNCC's goldmines [7] We managed to collect about 50 black and white images from UNCC when the college was still developing.

2.2. High Resolution Images

Collecting 200 images of modern High Resolution pictures of The University of North Carolina at Charlotte we found images from ucomm.charlotte , UNCC's own flicker which housed many collections consisting of construction, vintage, academic life and campus culture. We also were looking into UNCC's official shuttershock images. Also taking pictures on our own we seemed fit for data collection.

3. Colorization of Old UNCC Images

In this section we discuss the journey we embark to explore various approaches to solve the task of colorizing the Old UNC Charlotte images and show the progressive improvement we've had over the results for colorization.

3.1. Generative Models

There exists pre-trained generative models that produce accurate results. We utilize a base diffusion model and make incremental enhancements towards its ability to colorize images. Throughout our journey, we work a) a base diffusion model, b) a diffusion model with controlNet applied to its pipeline, c) a state-of-the-art NoGAN model, which is a class of synthetization algorithms that do not rely on neural networks for training, as traditional GAN models do, and d) our base diffusion model fine-tuned through the use of Low-Rank Adapation[9].

3.2. Colorization using Stable diffusion

The first approach for colorization is using the base diffusion model(stable-diffusion-v1-5) [4] through the image-to-image pipeline and colorizing the old b/w images of UNC Charlotte. Due to the very low resolution of the original b/w images where majority of the images are with dimensions 200x162 pixels, the base stable diffusion model does not pickup enough data to re-colorize **(Figure 2)**. We tried several prompts with many combinations of varying strength, guidance scale as well as number of inference steps. We decide to use the base stable diffusion model over the larger SDXL model as SDXL works especially well with image sizes between 768x768 and 1024x1024 and is more suited to generate a high resolution image and does take more computation than the base stable diffusion model, where our task at hand does not need a high resolution image.

3.3. Colorization after Super-resolution

To mitigate this low resolution of the original images, we move on to use a super-resolution model - ldm-super-resolution-4x-openimages which super resolves a base image of resolution to a resolution of 768x640 where the larger dimension of each image is 768 pixels. After the images are super-resolved, we again use the image-to-image pipeline with the stable-diffusion-v1-5 model and find that the resultant images are way more closer to that of the original image compared to the approach 1. In **Figure- 3**, it can be observed that the stable diffusion does have enough information create an image close to that of the original but not colorized optimally.

3.4. Colorization using controlNet

In this approach, we use ControlNet [5] - a neural network structure to control diffusion models by adding extra conditions, where additional input conditions are used to control pre-trained large diffusion model. These additional conditions can be pre-processed images such as **canny** (A monochrome image with white edges on a black background), **depth** (A gray-scale image with black representing deep areas and white representing shallow areas), **human scribbles** (A hand-drawn monochrome image with white outlines on a black background) and more to a total of 8. We decided to specifically use the checkpoint which corresponds to the canny images for our task at hand.

In **figure-4**, we can compare an image with its corresponding canny image. We process the canny images for the super-resolved images of the original b/w images and use them for each of the tests to follow. The model now takes in, the canny image and the super resolved image to generate a colorized image with the additional parameters discussed. We run a total of 8 tests over the old b/w images to fine-tune the parameters for this pipeline, i.e prompt, strength, guidance scale, negative prompt. For each test, we change one parameter to find its optimal value and then move to a different parameter. The trade off for the results is between the number of images that the model can colorize and the closeness of the colored images to that of the original b/w images.

(a) Black and white images collected

(a) Black and White image of UNCC

(b) Stable Diffusion Model result of above image

Figure 2. Comparison of the original black and white image to colorized using baseline stable diffusion model without super-resolution

(a) Super resolved UNCC image

(b) Stable Diffusion Model result of above image

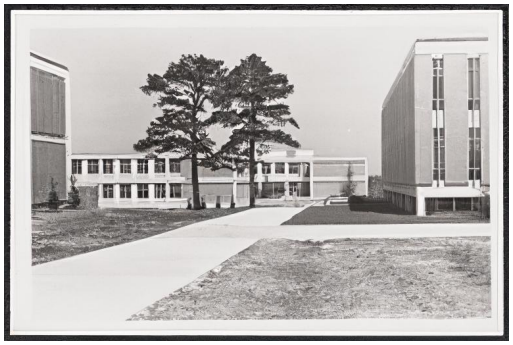
Figure 3. Comparison of the Super-resolved black and white image to colorized using the base stable diffusion model

We target the number of colorized images to be greater, the artifacts(buildings, people) in the colorized images are too far and altered from that of the original b/w. But when we target the closeness, the number of images that are colorized are reduced. We are able to find optimal parameters at test-6 where the **prompt** is "Color film, unc charlotte, 8k resolution, best quality, clay buildings, green grass, blue sky, mirror clear, trees green, water blue, black to color", the **negative prompt** is "ugly, deformed, disfigured, poor details, bad anatomy", The number of inference steps set to 120, **strength** set to 0.9, and the **guidance scale** set to 7. For test-6, the model was able to colorize 48 images with varying colorization results with some images close to original and others not so much out of the 50.

3.5. Generation parameters

Below are the generation parameters we

1. **Prompt** - This is the text description or instruction that guides the generation process of the diffusion model. It serves as a starting point for the model to understand the desired output.
2. **Negative Prompt** - This is parameter that allows you to specify things you don't want to be included in the generated output. It can help steer the model away from undesirable content or styles.
3. **Number of Inference Steps** - This parameter determines the number of iterations the diffusion model will



(a) Super-resolved Black and White image of UNCC

(b) Canny image of the above image.

Figure 4. Comparison of the super resolved black and white image with its canny image.

(a) Super resolved UNCC image

(b) ControlNet Model result of above image

Figure 5. Comparison of super-resolved black and white image with the controlNet with stable diffusion.

run during the generation process. More inference steps generally lead to higher-quality outputs but also increase computational cost.

(a) Super resolved image UNCC

(b) ControlNet Model result of above image

Figure 6. Comparison of super-resolved black and white image with the controlNet with stable diffusion.

4. **Strength** - A higher setting on the strength Parameter introduces greater levels of noise, resulting in increased random variations and reduced consistency with the original image. Where as the lower value introduces less noise and is more close to that of the original image with reduced randomness.
5. **Guidance Scale.** - This parameter is responsible for controlling the strength of the guidance from the input prompt. Higher values would result in a stronger emphasis on the input text prompt, possibly leading to a more accurate representation of the prompt in the output image.

3.6. Qualitative evaluation

There is an apparent increase in the quality of colorization that is undergone between our base SDM and after we introduce ControlNet, to apply additional inputs. Various experiments show that applying Super-Resolution to our images, ControlNet to our base SDM, and performing prompt refinement, with the introduction of a negative prompt to counter unwanted results, our Image Colorizer does perform better at enhancing the colorization of images(**Figure 5**). However, there are instances in our experiment where our outputted image fails to properly colorize to buildings(**Figure 6**).

Findings. The instances where our SDM fails to properly add the correct color or omit the building textures leads us to believe that our SDM cannot accurately colorize UNC Charlotte buildings because the training dataset that has pretrained our base SDM does not contain images of a similar color palette to these buildings. To further enhance our SDM's effectiveness, we finetune our model by introducing an additional dataset that includes colored images taken of UNC Charlotte campus today.

4. Colorization using LoRA Finetuning

The datasets used to train our base Stable Diffusion Model (SDM) hinder their performance in capturing diverse colors and textures. **L**ow-**R**ank **A**daptation, known as LoRA, is a computationally inexpensive approach that allows for the finetuning of our SDM. LoRA proves effective in generating colorized renderings of archived UNC Charlotte images. Complex finetuning approaches require that the SDM's pre-trained weight, W, is merged with the weight created from finetuning, (W). However, LoRA decomposes (W) to two trained, lower ranked, smaller matrices, (A) and (B). These matrices act as an adapter that is merged to our base model, producing a fine-tuned SDM. The **P**arameter **E**fficient **F**ine-**T**uning (PEFT) library provided by Hugging Face allows for the trained adapter to be applied to our base model, strengthening the performance of our SDM in comparison to our base model and competing state-of-the-art models[10].

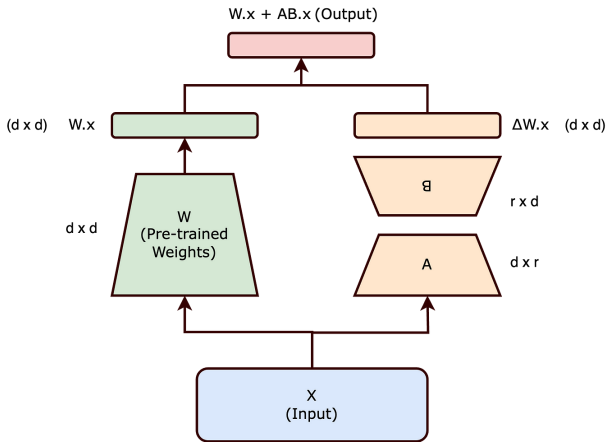


Figure 7. Low Rank Adaptation (LoRA) architecture

Between the original B/W image, Super-Resolved colorized image, and newly outputted image from our fine-tuned SDM, there is a clear enhancement in the color vividness and texture that is apparent on the buildings.

(a) ControlNet result for a black and white image.

(b) After implementation of LoRA finetuning, super-resolved image of UNCC

Figure 8. Comparison of enhancement by LoRA finetuning

5. Colorization of High Resolution Images

After building our validation set from our 80 20 data split We conduct a series of experiments on our validation set, which we use as our ground truth, and evaluate how effective our fine-tuned SDM performs against our base SDM, earlier iterations of our fine-tuned model, and state-of-the-art NoGAN DM.

5.1. Metrics

We employ **F**réchet **I**nception **D**istance (FID), **L**earned **P**erceptual **I**mage **P**atch **S**imilarity (LPIPS), **P**eak **S**ignal-to-**N**oise **R**atio (PSNR), and the colorfulness score.

FID quantifies the realism and diversity of images generated by our SDM, assessing the impact of our SDM's network on the colorization of images. FID is based on The Fréchet distance, which measures the similarity between two curves[11]. FID scores are measured by computing the differences between the representations of features, such as edges, lines, and shapes present in the images that are transformed into a latent space; the lower the FID score indicates the greater the similarity between two images.

LPIPS measures the perceptual similarity between two images, matching closely to human perception. Scores are generated comparing the activation functions of a model be-

Figure 9. From left to right: First image is the original black and white image of UNCC; the second image is the result of our base stable diffusion model; Third is the result of the base stable diffusion after super resolution is applied; Fourth is the result after controlNet is applied to our base stable diffusion model pipeline, to apply extra input conditions; Fifth is the result after fine-tuning using LoRA

tween two different images. The lower the LPIPS score indicates the greater the similarity between two images[12].

PSNR describes the ratio between the maximum power of a signal and the power of corrupting noise that affects an image. The higher the PSNR score indicates lower distortion and higher similarity between two images[13].

Colorfulness measures the amount, intensity, and saturation of colors in an image; scaling on a range from not colorful to extremely colorful, which can be used to describe how vivid or dull an image visualizes to. The standard deviation and mean of the images color channels are computed to attain the colorfulness score[14].

Diffusion Models	**FID↓**	**LPIPS↓**	**PSNR↑**	**Colorfulness↓**
Base SDM	235.67	0.625	1.3943	12.67
ControlNet SDM Test #1	199.07	0.465	1.3949	52.95
ControlNet SDM Test #2	290.86	0.413	1.3957	44.77
NoGAN DM	174.95	0.369	1.508	37.29
LoRA Finetuned SDM	**166.26**	**0.369**	1.397	19.43

Table 1. Quantitative Evaluation

The above metrics are **averaged** scores of our validation sets across all of the images contained within them. The metrics indicate that our fine-tuned SDM does outperform our base SDM across almost all evaluations, except for PSNR. The colorfulness scores are compared against the colorfulness score given to our validation set of unedited, colored images, which averaged out to 44%. The result proves that our fine tuned SDM does not exactly match the vividness of their groundtruths for outputted images; however, our score does indicate that our SDM does not apply too much vividness or too little vividness to images. We implemented FID and LPIPS in lieu of conducting human surveys for our qualitative evaluations, in which our fine-tuned SDM outperforms other models' scores, and in fact, shares an equal metric score with the state-of-the-art NoGAN DM, at 0.369.

Both of these metrics share a similar evaluation scale: the lower the averaged score indicates the greater the similarity between our dataset of images. We are aware that our qualitative evaluation proves that our fine-tuned SDM falls short in ensuring that the outputted images depict realism and fail to match the exact hue, brightness, and saturation of the original, colored image, so we as well introduce a Pixel-Level Dissimilarity metric evaluation, PSNR. The results prove that our fine-tuned model does not significantly increase the similarity between its outputted images and the

original, colored images from the base SDM, nor outperform the NoGAN DM, as higher PSNR values indicate a greater similarity between the sets.

6. Contributions.

The team researched cutting-edge computer vision project that would positively impact and contribute to The University of North Carolina at Charlotte (UNC Charlotte).

Anirudh Cheruvu

1. Anirudh Cheruvu executed multiple iterations of tests on the pipeline from approach-4 that containes pretrained Stable Diffusion Model (SDM),controlNet with the appropriate test dataset of images.
2. Anirudh Cheruvu tested the performance of the base SDM and produced the super resolution, canny images for the pipeline in the approach 4.
3. Anirudh Cheruvu implemented LoRA finetuning on the stable-diffusion-v1-5 and saved the checkpoints 13 of the model.

Hamilton Sheppard

1. Hamilton Sheppard analyzed the results of their multiple checkpoints after LoRA finetuning to denote differences between the generated weights at specific intervals of the check-pointing steps.
2. Hamilton Sheppard calculated metrics to compare their finetuned SDM against their earlier iterations and a competing state-of-the-art model.

Kevin Mai

1. Kevin Mai researched and sourced archival images of UNC Charlotte campus, dating back from 1960's, and modern images from 2022 through multiple sources.
2. Kevin Mai analyzed the results of testing fine-tuned check points of each test of 500, 3000, 6500. Testing about multiple tests for each checkpoint and finding out the right parameters for fine tuning images.

References

[1] S. Treneska, E. Zdravevski, I. M. Pires, P. Lameski, and S. Gievska, "Gan-based image colorization for self-supervised visual feature learning," *Sensors*, vol. 22, no. 4, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/4/1599

[2] E. M. Farella, S. Malek, and F. Remondino, "Colorizing the past: Deep learning for the automatic colorization of historical aerial images," *Journal of Imaging*, vol. 8, no. 10, 2022. [Online]. Available: https://www.mdpi.com/2313-433X/8/10/269

[3] A. Deshpande, J. Lu, M.-C. Yeh, M. J. Chong, and D. Forsyth, "Learning diverse image colorization," 2017.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2021. [Online]. Available: https://arxiv.org/abs/2112.10752

[5] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[7] "Buildings — Goldmine — UNC Charlotte — goldmine.charlotte.edu," https://goldmine.charlotte.edu/islandora/object/ua%3Aunccp1-1, [Accessed 25-04-2024].

[8] "ArchivesSpace Public Interface — UNC Charlotte Finding Aids — findingaids.charlotte.edu," https://findingaids.charlotte.edu/, [Accessed 25-04-2024].

[9] F. Mameli, M. Bertini, L. Galteri, and A. Del Bimbo, "Image and video restoration and compression artefact removal using a nogan approach," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 4539–4541. [Online]. Available: https://doi.org/10.1145/3394171.3414451

[10] "LoRA," https://huggingface.co/docs/peft/main/en/conceptual_guides/lora, accessed: 2024-4-25.

[11] T. Wylie and B. Zhu, "Following a curve with the discrete fréchet distance," *Theor. Comput. Sci.*, vol. 556, pp. 34–44, Oct. 2014.

[12] "Learned perceptual image patch similarity (LPIPS) — PyTorch-metrics 1.3.2 documentation," https://lightning.ai/docs/torchmetrics/stable/image/learned_perceptual_image_patch_similarity . html, accessed: 2024-4-25.

[13] I. Follow, "Psnr," https://www.geeksforgeeks.org/python-peak-signal-to-noise-ratio-psnr/, Jan. 2020, accessed: 2024-4-25.

[14] A. Rosebrock, "Computing image "colorfulness" with OpenCV and python," https://pyimagesearch.com/2017/06/05/computing-image-colorfulness-with-opencv-and-python/, Jun. 2017, accessed: 2024-4-25.