“Nothing’s Black or White”

— Nelson Mandela

Title	Final Report: Image Colorization
Course	APS360 - Applied Fundamentals of Machine Learning
Team Number	70
Team Members	Kanav Singla - 1004997827 Aniruddha Redkar - 1004696858 Adrian Ramlal - 1004811858 Siddharth Vijay - 1004186640
Date Submitted	09/04/2021
Word Count	2487
Penalty	0%

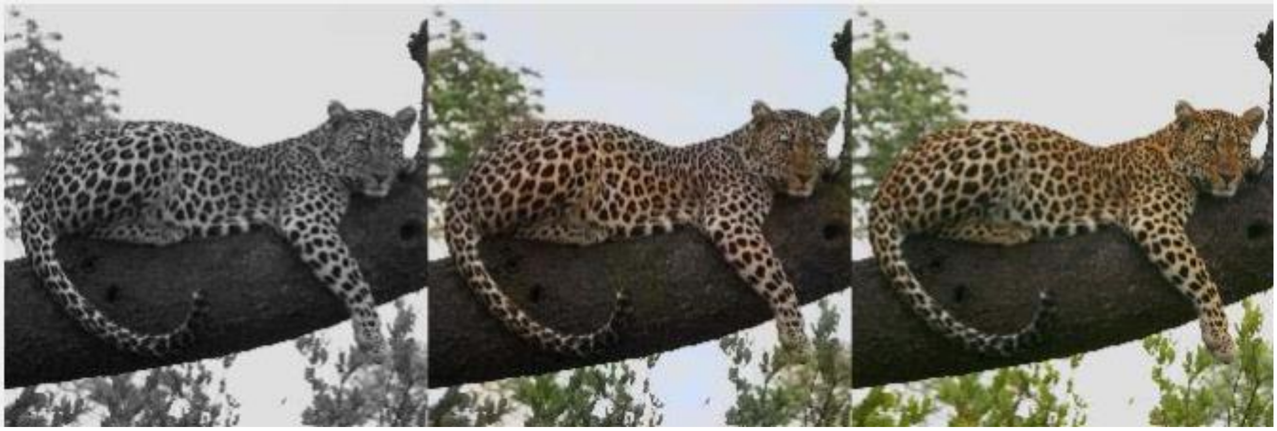
1.0 Introduction

1.1 Motivations and Goal

A large part of historic and scientific research is affected due to grayscale images. Electron microscopes and the Hubble telescope generate grayscale images, proving the variety of practical applications for automatic colorization in the fields of nanotechnology, medicine, astronomy etc [1]. Additionally, use cases exist in the colorization of historical images and films [2].The goal of this project is to bring new life to these grayscale images by using neural networks to learn colorization techniques from colored images.

1.2 Why Machine Learning (ML)?

This problem has no exact solution yet, because there is no deterministic relation between the grey image luminance and the chrominance values if it was a colored [3]. Alternative solutions to ML can be categorized into Manual and Semiautomatic. Manual Coloring requires a large amount of skill, time and effort. Semiautomatic Coloring is easier but yields unsatisfactory results [4]. Neither are practical with the advancements in science and technology so we will employ neural networks and artificial intelligence to efficiently colorize



2.0 Illustration / Figure

Fig 1: *Grayscale image (left), Colorized Image generated by model (middle), True Color image (right).*

We aim to create a model that can take in a grayscale image (left) and perform image colorization (middle). The image on the right is the true color and is used solely for comparison purposes.

Our goal is to be able to produce these colorizations by extending the idea of convolutional model architectures used in the past (see Fig 2.0), to the concept of using 2 competing neural networks to induce growth and increased performance, that is a generative adversarial network (GAN). The project design flow is as follows:

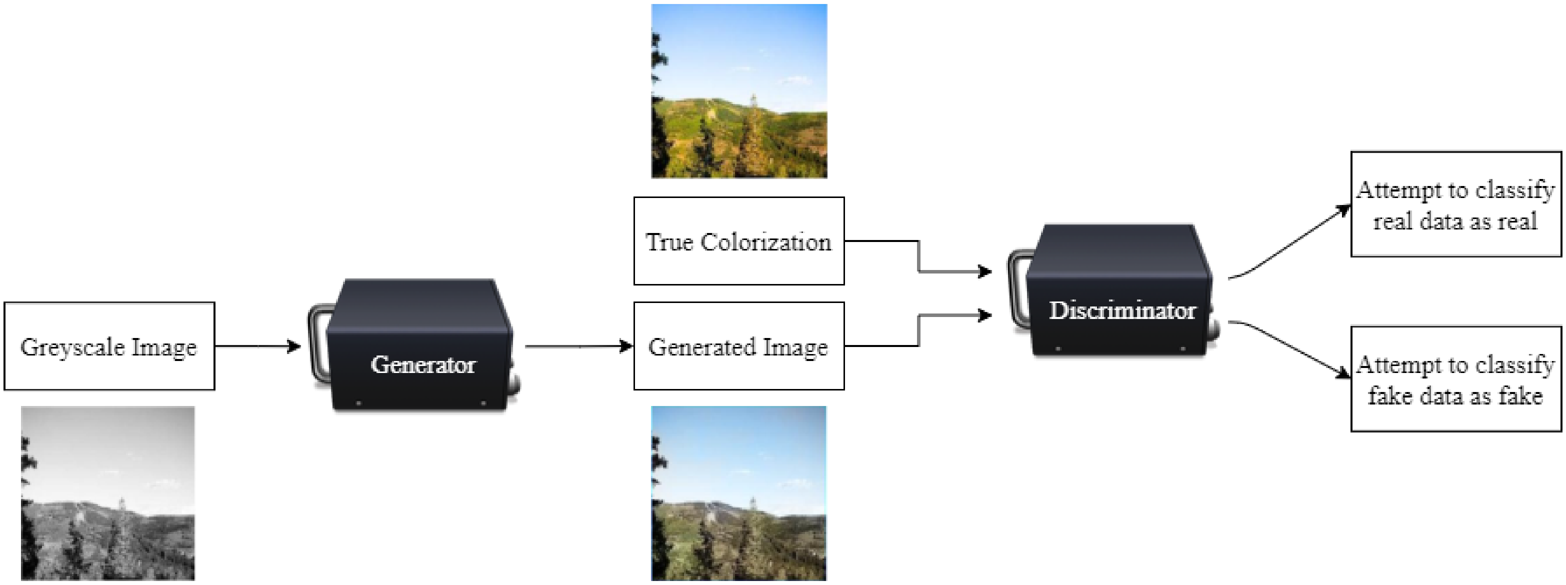

True Colorization
Greyscale Image
Generator
Generated Image
Discriminator
Attempt to classify real data as real
Attempt to classify fake data as fake

Fig 2: *Image pictured above details the flow of data (grayscale images) -- input passed into a generator model which outputs the colorized image , then passed into the discriminator*

Once our GAN is sufficiently trained, the generator should be able to produce a colorization of the picture it has never seen before in the testing and application phase.

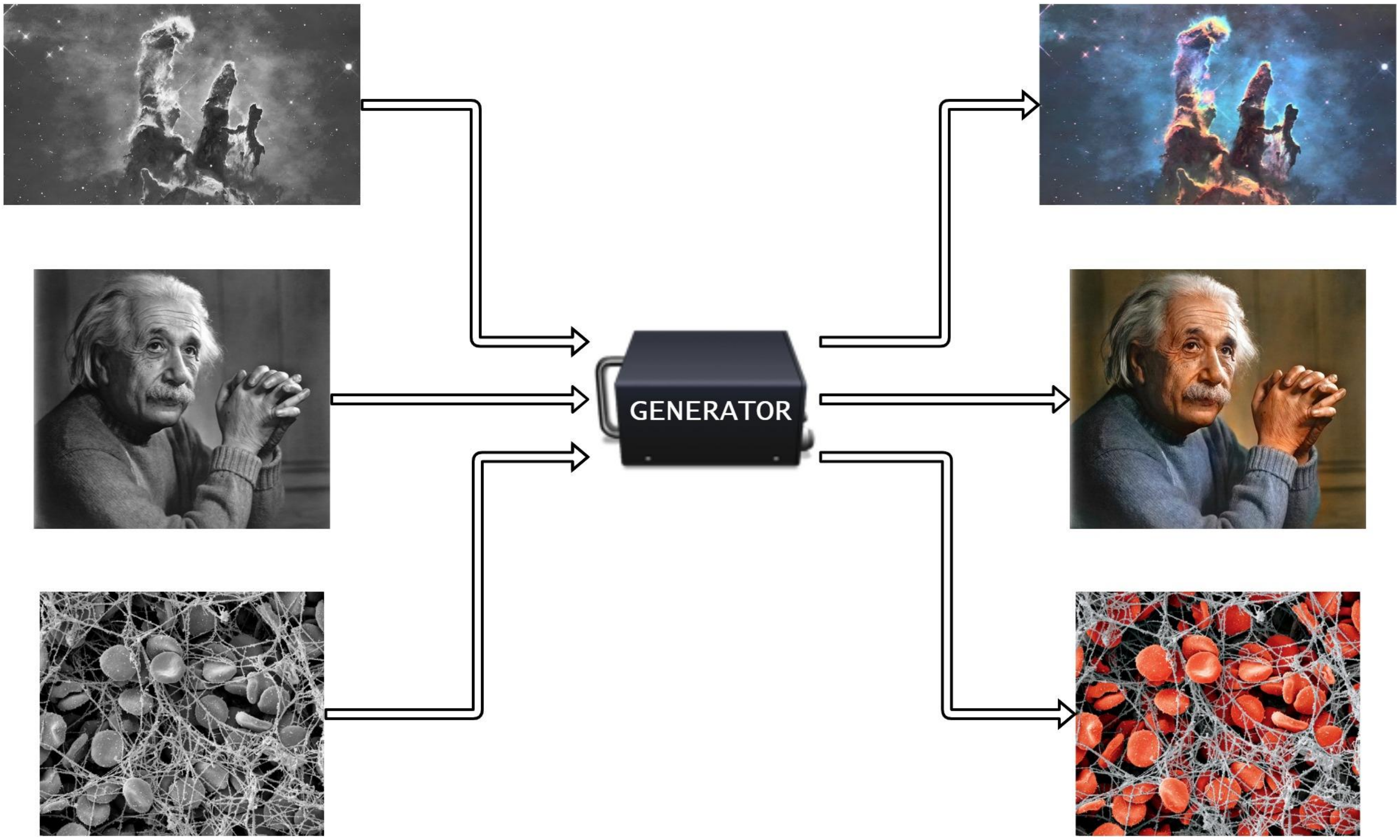


Fig 3: *Showing the performance goal for the Generator. Astronomical Image (Top), Historical Image (Middle) and Medical Image (Bottom) [5]*

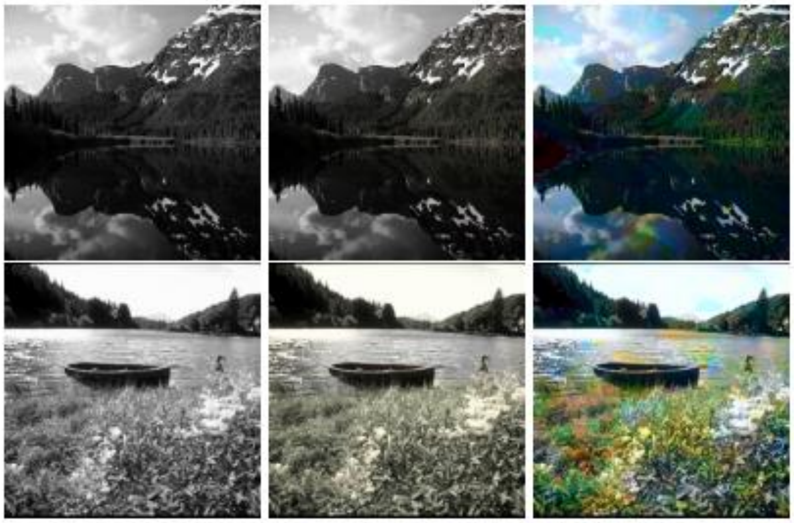
3.0 Background & Related Work

Prior colorization algorithms have developed similar systems, which leverage large-scale data and CNNs. Their major differences arise in the way they treat the data for modeling the correspondence between grayscale and color [6]. The two method categories being Scribble-based and Transfer-based [7].

3.1 Image Colorization with Deep Convolutional Neural Networks by Jeff Hwang and You Zhou [5].

This paper presented a CNN-based system that colorizes grayscale images using a statistical learning driven approach. The baseline used was a regression model which we also employed. Results were better than regression but some outputs were under-colored or inconsistently colored [8].

Fig 4*: Showing inputs (left), regression network output(middle) and the classification network output(right). The regression network's output is severely desaturated and generally unattractive relative to the classification's network [8].*



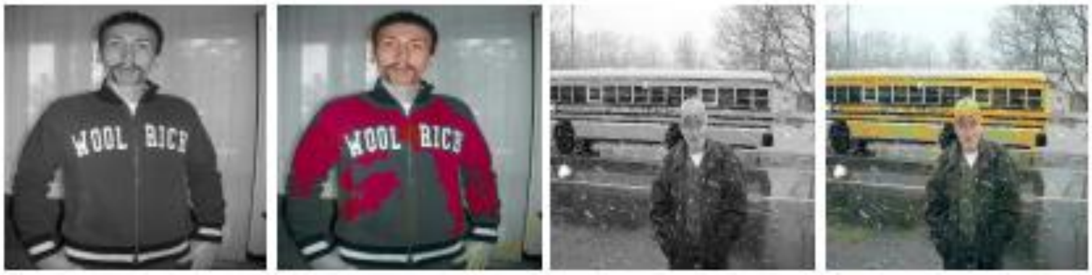


Fig 5*: Showing examples of color inconsistencies inaccurately predicted by the model [8].*

3.2 Automatic Colorization by Ryan Dahl [6].

This paper described a CNN model using 4 layers from VGG16. Intermediate outputs produced by the encoding portion of the network were residually connected to outputs produced later by the decoding portion of the network. Performance was great on foliage and skin but the images were generally desaturated [9].

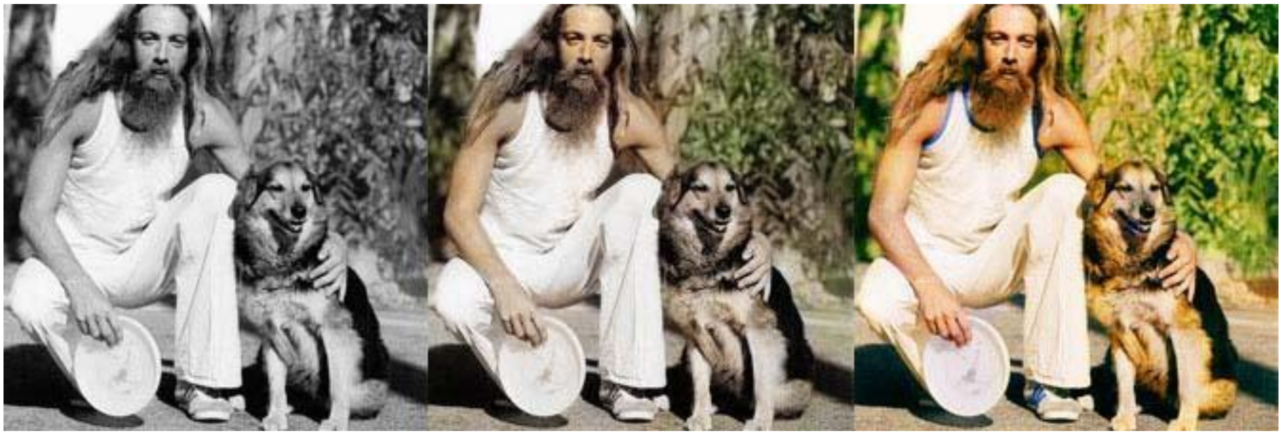


Fig 6*: Showing good performance on skin tones. The prediction(middle) is less saturated then the real image(right) [9].*

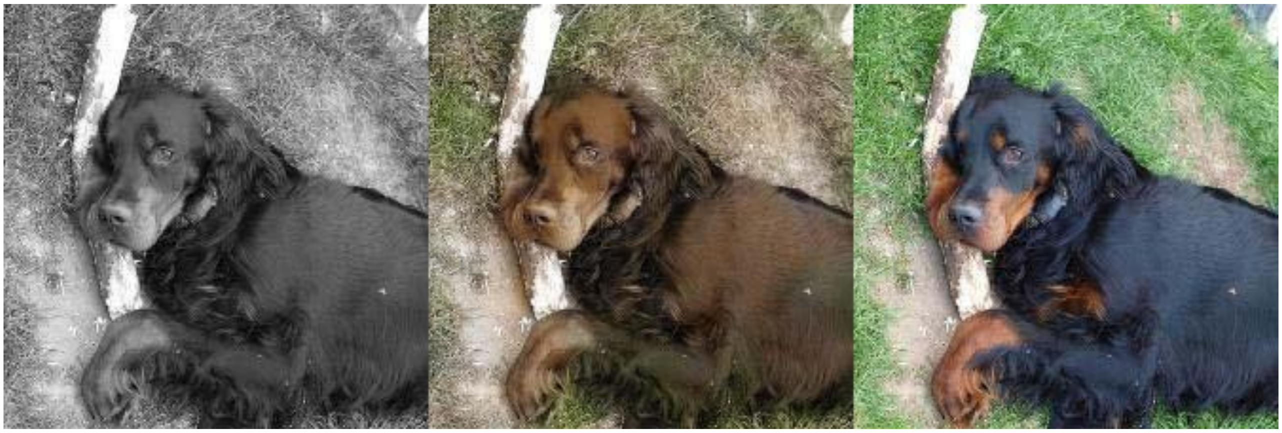


Fig 7*: Showing mediocre performance on animal hair colors. The predicted (middle) dog is brown while the real dog(right) is black. Poor saturation is also observed [9].*

4.0 Data Processing

The model utilizes images of dimensions 224 × 224 with 3 channels. These images will be in the CIELUV color space [10]. The luminance, L, channel will be used to train the model and the chromaticity values, U and V, channels will be used as the ground truth labels.

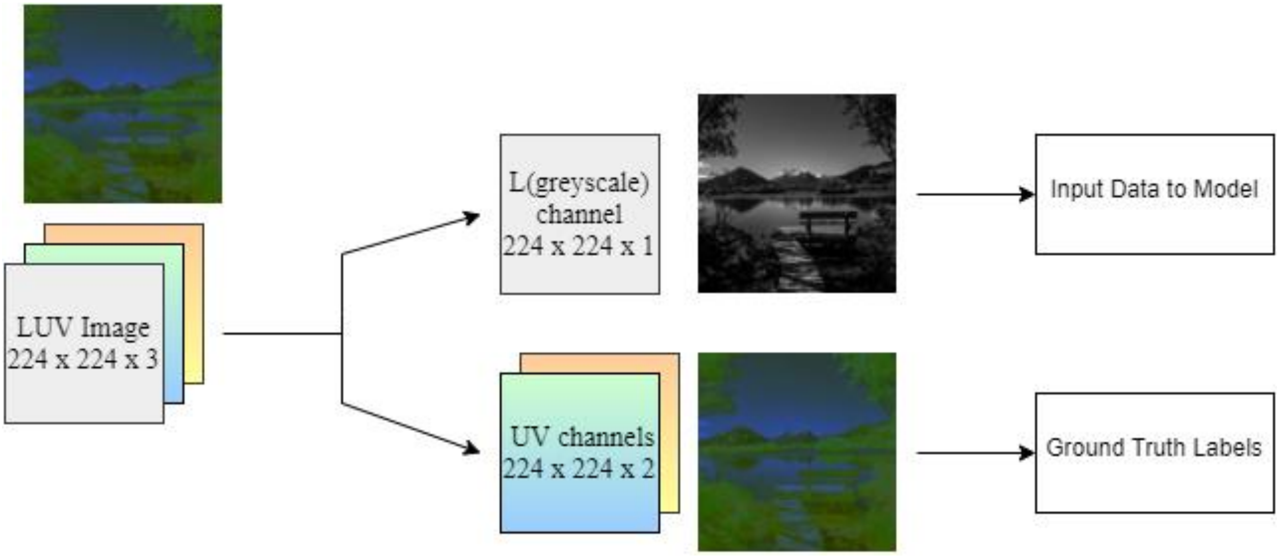


***Fig 8**: Showing data input to and utilization by the model.*

4.1 Sources of Data and Intended Usage

Subsets of ImageNet and Flickr30k were used to ensure a diverse set of training images that will expose the model to as many different object and colorization techniques as possible. CIFAR-10 was used for validating our concepts for the baseline and initial models.

Dataset	Training	Validation	Testing
ImageNet[11]	12000	2000	1000
Flickr30k [12]	12000	2000	1000
CIFAR-10 [13]	5000	1000	1000

***Table 1**: Showing datasets and the number of images allocated to each phase. A variety of sets were used to diversify the inputs to models.*

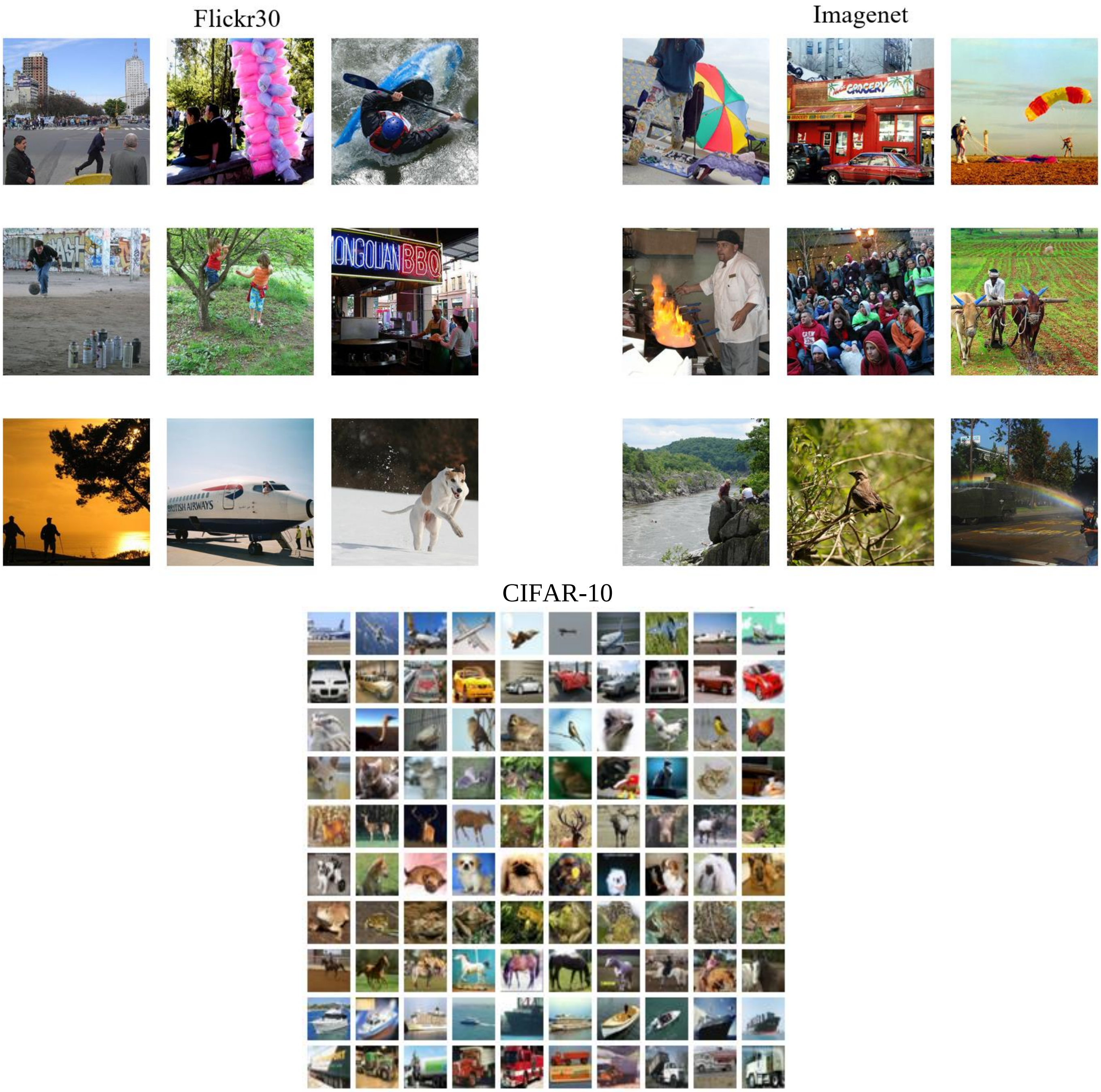


***Fig 9**: Showing samples from Flick30, Imagenet and CIFAR-10 [11 ,12, 13].*

We also created an additional testing set to determine how well the model generalizes with images where there are no colorized forms and images where there are not intuitive colors to use in the conversion process. This set consisted of 50 images in the Medical, Astronomical and Historical fields [5, 14, 15].

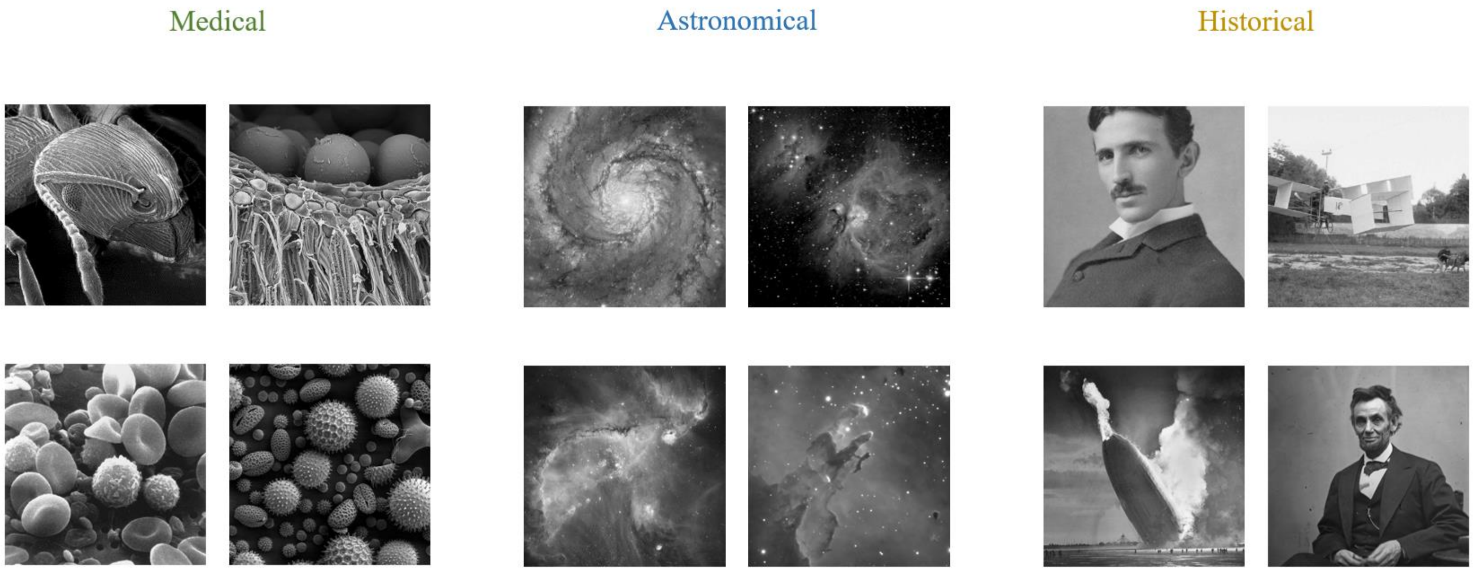


Fig 10*: Showing samples from the constructed testing set [5, 14, 15].*

4.2 Cleaning Process for Training

The above data sets consist of images of varying pixel dimensions with 3 color channels (RGB). The cleaning process will convert these to a 224 × 224 × 3 image in the LUV color space. Skimage’s color library is used for these conversions.

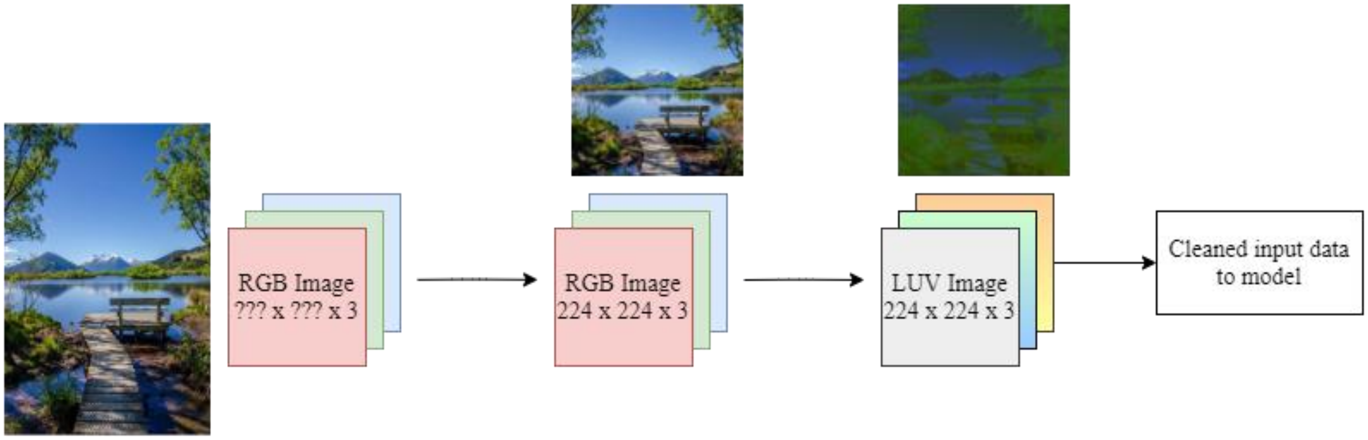


***Fig 11**: Showing the major steps in the cleaning process. The first RGB image is of arbitrary size and the output LUV image is expected by the model.*

4.3 Data Pipeline

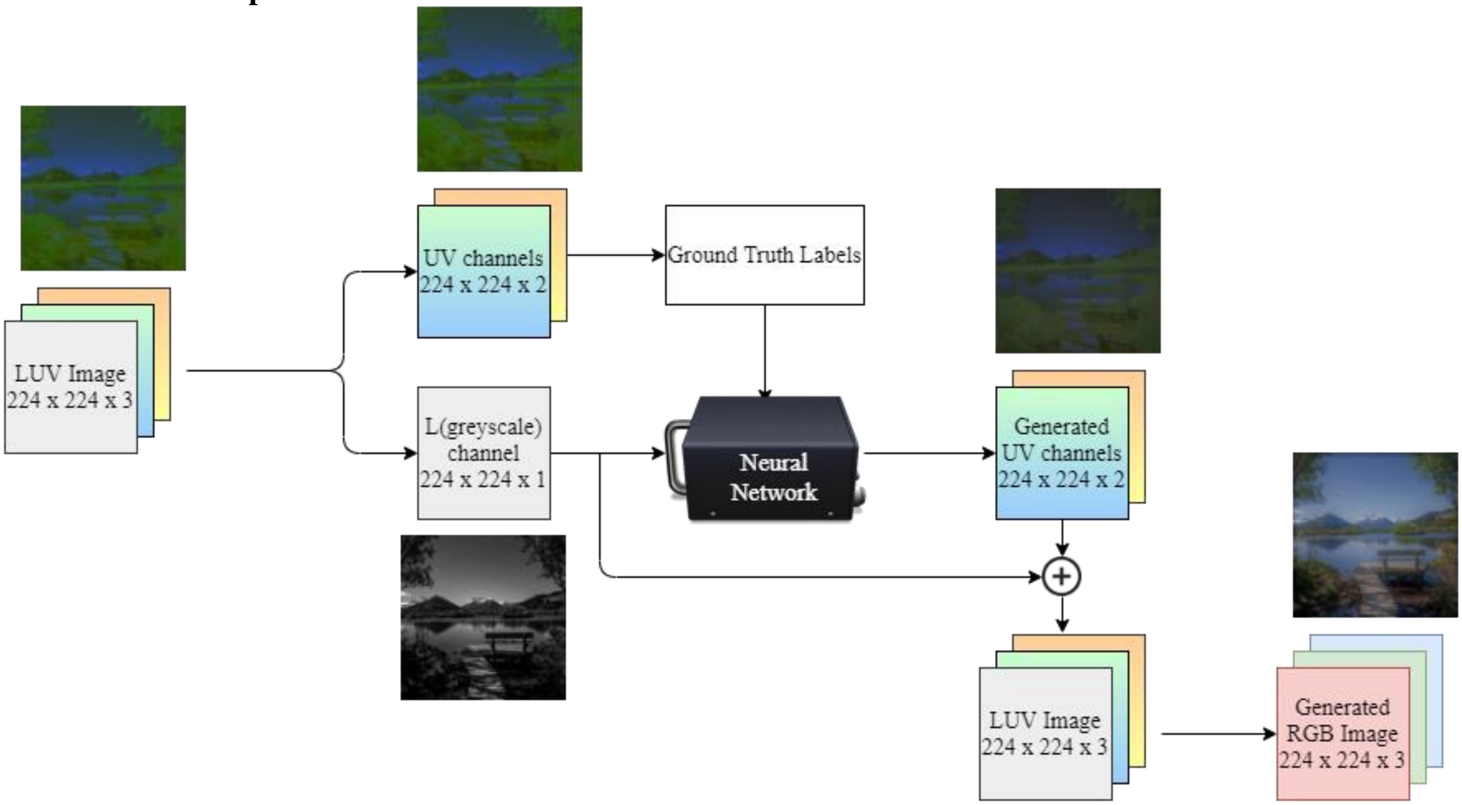


***Fig 12**: Showing the data pipeline for the entire model.*

4.4 Cleaning Process for Testing

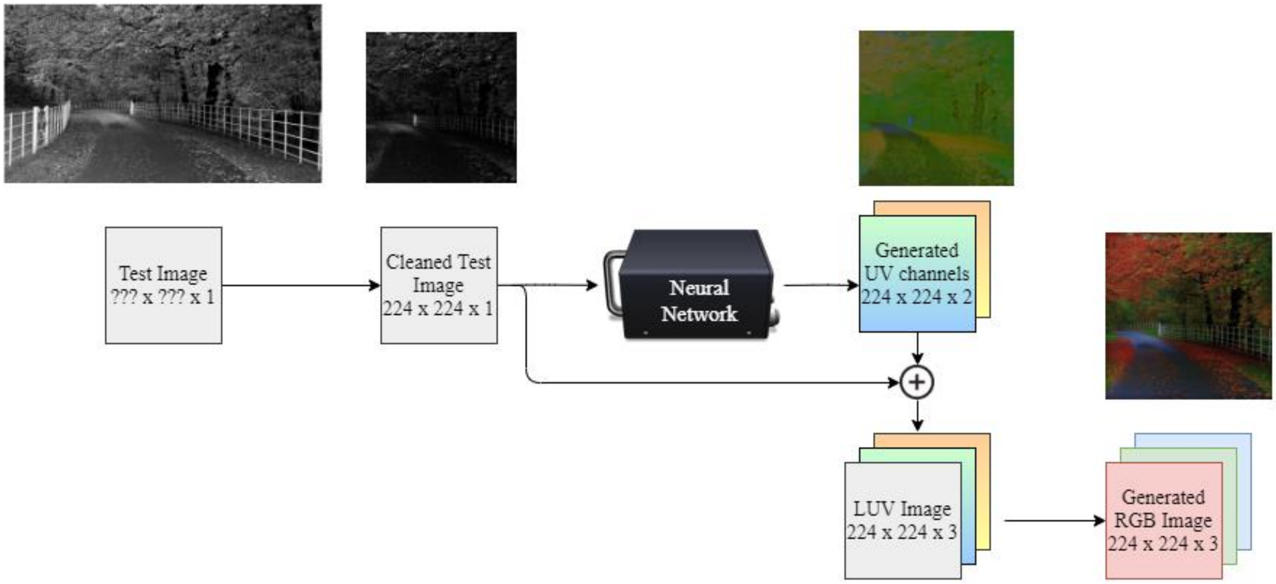


***Fig 13**: Showing the major steps in the cleaning process for test data and creation of output images*

5.0 Architecture

5.1 Initial: The Classification Based Model

Initially our classification model stacked multiple Convolution-Pooling-Normalization-Activation standard blocks to classify each pixel into 24 (different colors) categories. We use cluster centres of these 24 colors, produced by running k-means clustering over colours for the labels. To further simplify things, the distance is measured in RGB space.

The basic CNN architecture was further improved by making the following challenges:

- Using Strided & Transposed Convolutions instead of using upsample function, so that information is better preserved while upsampling
- Introducing skip connections between earlier and later layers, so that the original intensities of the pixels can be passed onto the later layers where most of the learned features are pretty abstract

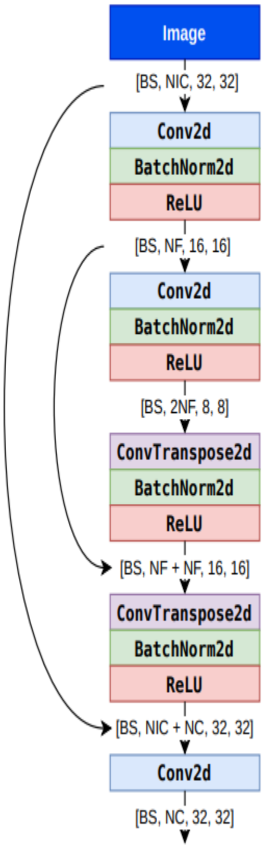


Fig 14: *U-Net Model Architecture*

5.2 Final: Generative Adversarial Network (GAN)

The Generative Adversarial Network (GAN) architecture that we employed for our core model consists of 2 convolutional network models being trained in parallel, competing with each other, and consequently increasing their performance. The two models used are labeled as the Generator and the Discriminator. The former is responsible for taking in grayscale images and colorizing them. The discriminator is responsible for determining if any given image is real or fake. As such, the generator tries to colorize images such that they look so real, the discriminator would be unable to identify the image as fake.

The Generator architecture is inspired by the U-Net architecture shown in Fig. 15 and is a series of convolutional and deconvolutional blocks paired with skip connections. This is the same method employed for our Generator as well -- we note 5 convolutional blocks (consisting of a Conv2D, a BatchNorm2D, and a ReLU) as well as 5 deconvolutional blocks (consisting of a ConvTranspose2D, a BatchNorm2D, and a ReLU) connected via skip connections. As for the Discriminator, we simply use a series of convolutional blocks (same structure as blocks found in the Generator) and connect those to a fully connected layer and a sigmoid to return a probability of a given image being real.

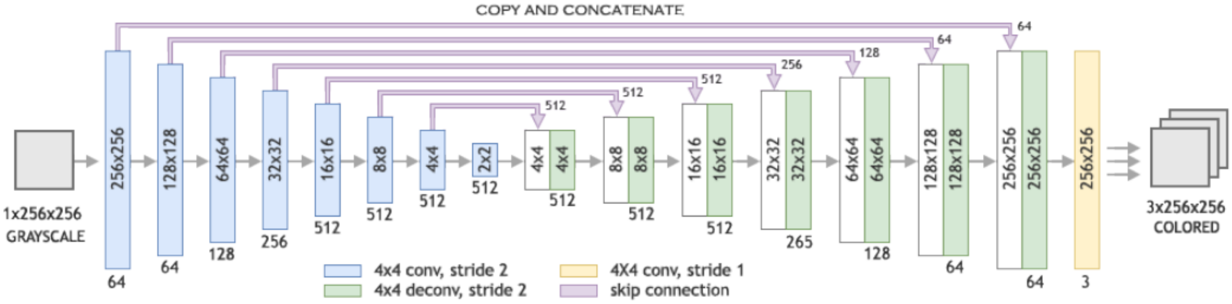
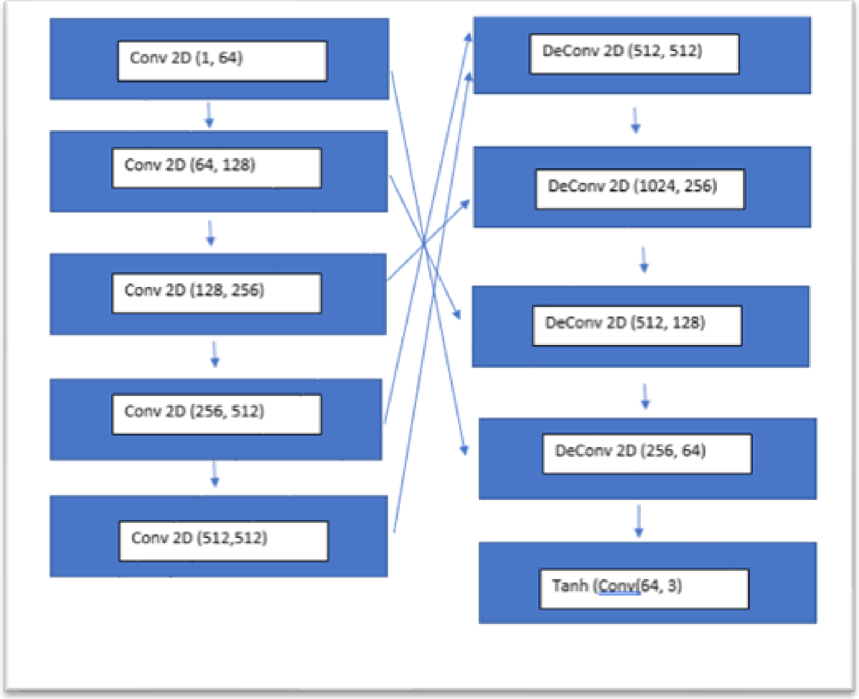
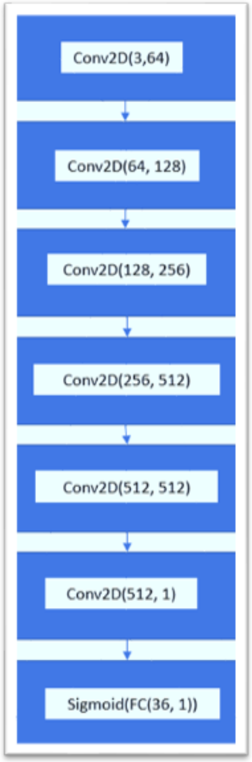


Fig 15*: The U-Net architecture which would inspire our Generator CNN [17]*


Conv 2D (1, 64)
Conv 2D (64, 128)
Conv 2D (128, 256)
Conv 2D (256, 512)
Conv 2D (512,512)
DeConv 2D (512, 512)
DeConv 2D (1024, 256)
DeConv 2D (512, 128)
DeConv 2D (256, 64)
Tanh (Conv(64, 3)


Conv2D(3,64)
Conv2D(64, 128)
Conv2D(128, 256)
Conv2D(256, 512)
Conv2D(512, 512)
Conv2D(512, 1)
Sigmoid(FC(36, 1))

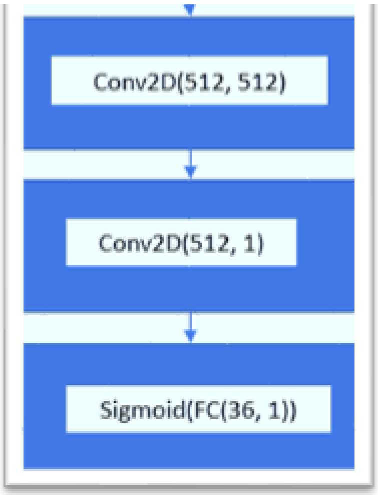

Conv2D(512, 512)
Conv2D(512, 1)
Sigmoid(FC(36, 1))

***Fig 16**: The architecture used for the generator(left) and discriminator(right) models in our GAN.*

6.0 Baseline Model

The baseline model that we are using is a regression model described in [8]. They consider the

Image Colorization problem to be a classification problem but use a regression model as their baseline, giving us a precedent to use it as ours. The model is widely used as a baseline for image colorization. The image below details the architecture of the regression model, consisting of two main components -- an encoding process in the form of conv2D, BatchNormal2D and ReLu and a decoding process in the form of upsampling.

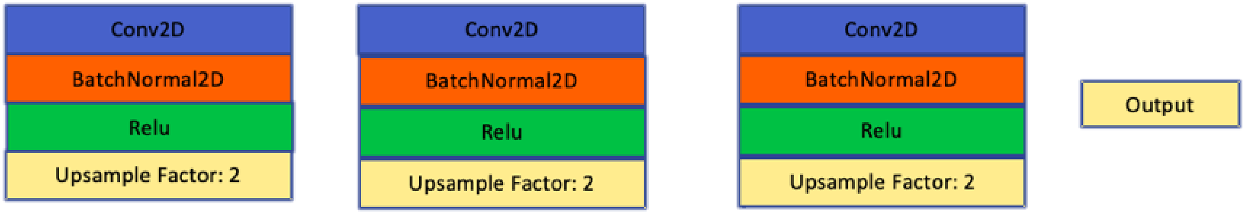


Fig 17*: The above image shows the architecture for the baseline model.*

The loss function used is the MSE loss

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Fig 18*: The above equation is the MSE loss function*

7.0 Quantitative Results -

7.1 Baseline Regression Model

As accuracy is more of a classification concept,we used MSE as a quantitative assessment tool by minimizing the squared distance between the predicted and true color value.

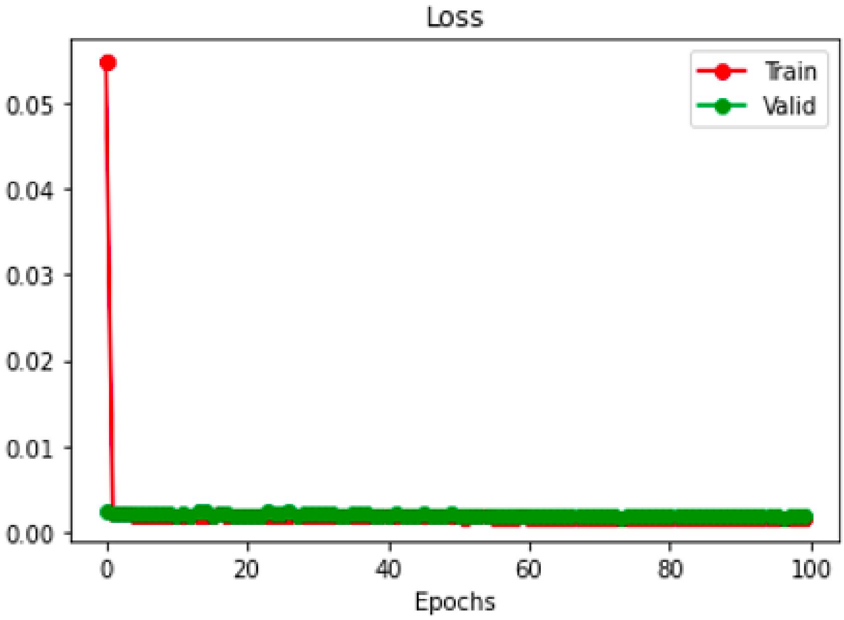


Fig 19*: The above image shows the Loss vs Epochs for the regression model*

Epochs	**Train Loss**	**Val Loss**
100	0.0017	0.0035

Table 1**:** *Final Quantitative results of our Baseline Regression Model*

7.2 Classification Based Model

For quantitative results, we see a Val accuracy of 43% on our first model and an improved Val accuracy of 59.8% and more generalizability on our final classification model (U-Net).

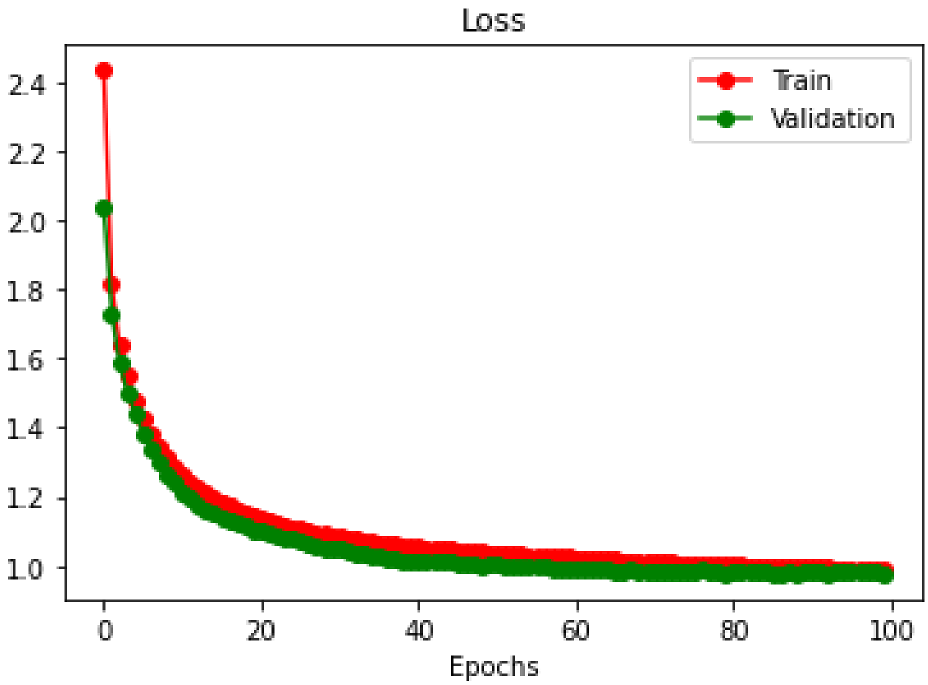


Fig 20: *The above image shows the Loss vs Epochs for our final classification based model (loss is computed over RGB space)*

Epochs	**Val Loss**	**Val Acc**
100	0.9769	59.8%

Table 2: *Final Quantitative results of our model for the given number of epochs*

7.3 Generative Adversarial Model (GAN)

In order to quantify the progress of the GAN we can note a couple values of interest; the loss of the Generator/Discriminator and the accuracy of the discriminator. As such, the latter is not as relevant to us since the loss of the discriminator does not indicate to us how well the generator is actually producing images. To elaborate on this, consider if the discriminator accuracy is very high -- this could indicate that the generator is simply performing poorly, and the opposite could indicate that the discriminator itself is not training properly, not necessarily that the images generated are good. Thus we omit the accuracy and look at the loss of the models.

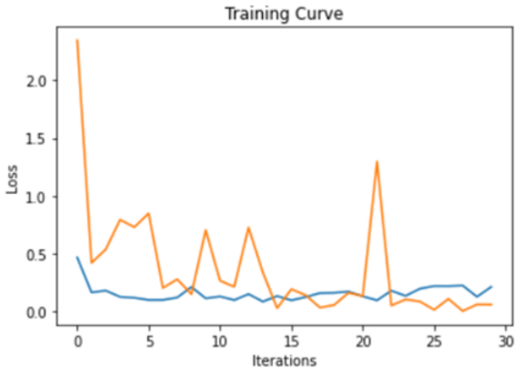


Fig 21*: The above image shows the Loss vs Epochs for the GAN model -- the generator loss is in orange and the discriminator loss is in blue*

Epochs	**Final Generator Loss**	**Final Discriminator Loss**
30	0.15696	0.33276

Table 3**:** *Final Quantitative results of our GAN -- final losses for generator and discriminator*

8.0 Qualitative Results

Qualitative result is the most important criteria for model performance for this project as humans subjectively perceive images so by looking at them.

8.1 Regression Model

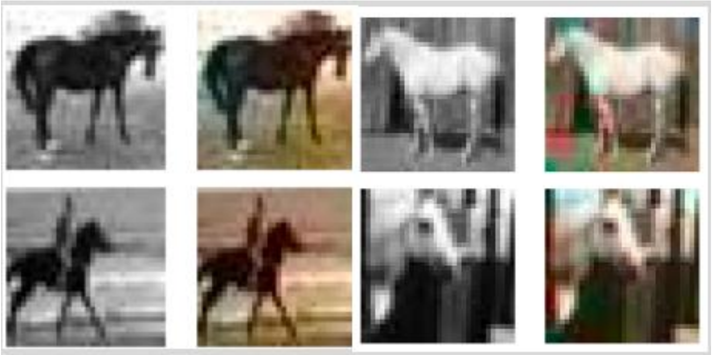


***Fig 22**: Qualitative result for regression*

The images used in the regression model are of size 32 x 32 as the complexity of training images in regression is a function of number of pixels. A surprising result is that we can see that the results we get on desaturated colors than a highly vibrant image.

8.2 Classification Based Model

The quality of results are improved by adding the residual connections and transposed convolutions. The results get slightly sharper, though still not too close to the ground truth in terms of pixel intensities.

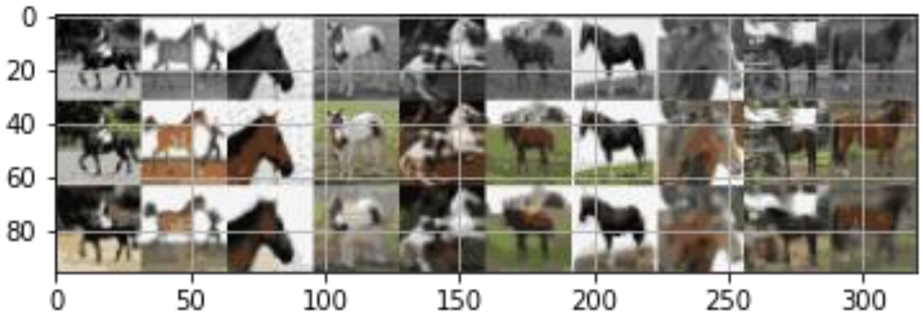


Fig 23: *Image colorization (bottom) done by our Simpler CNN model, ground truth (middle) and grayscale model inputs (top)*

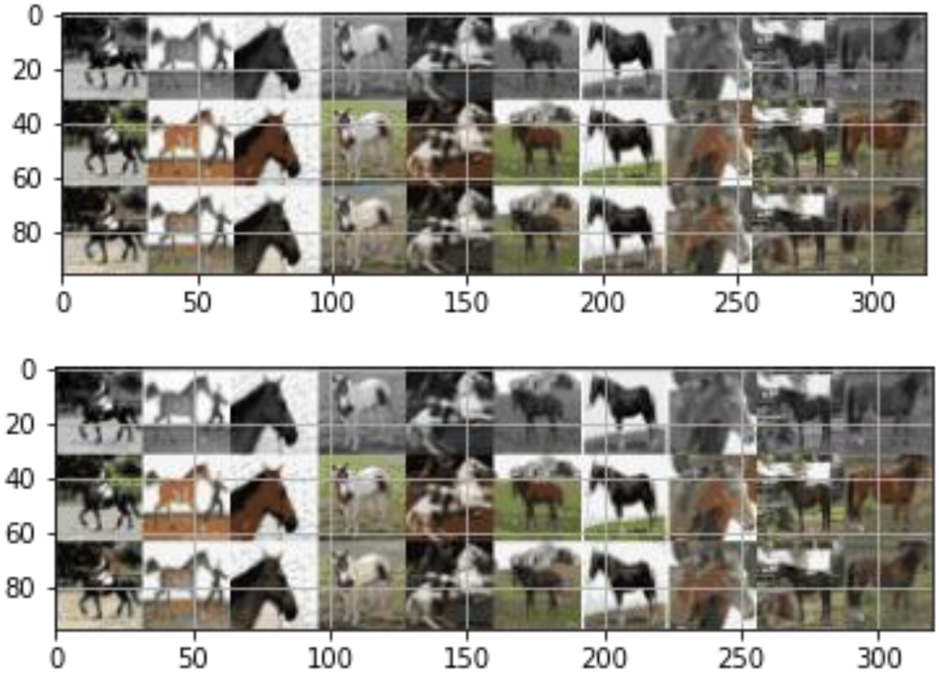


Fig 24: *Image colourization performed by our U-Net model after 50 epochs (above) and 100 epochs (bellow) of training; for each graph we have grayscale model inputs (top), ground truth (middle) and Colourized results (bottom)*

8.2 Generative Adversarial Network (GAN)

We see the results of the GAN as it colorizes a series of images; we find that it is able to easily detect human skin colors and distinct features but has trouble with more abstract things. For a more detailed discussion see Sections 9 and 10.

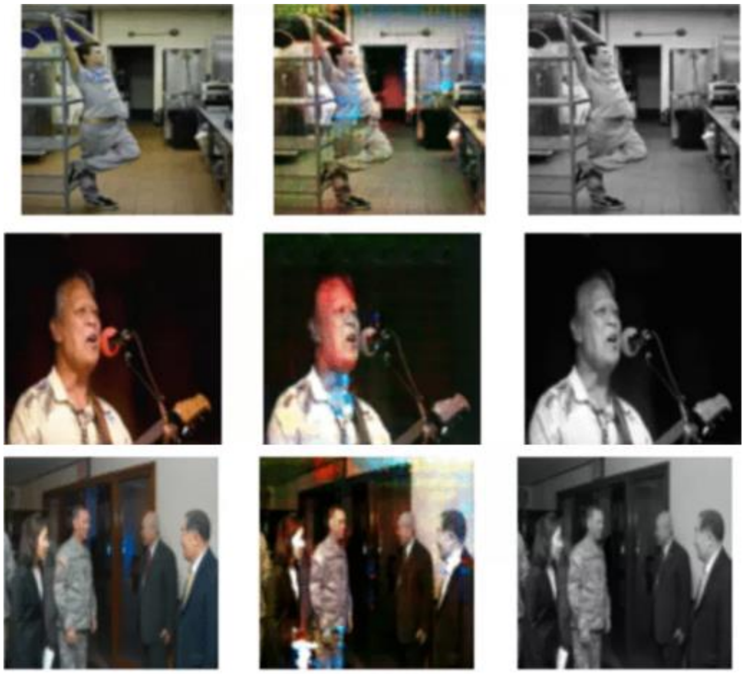


Fig 25: *Image colorization done by GAN; left (ground truth), middle (GAN colorized image), right (grayscale)*

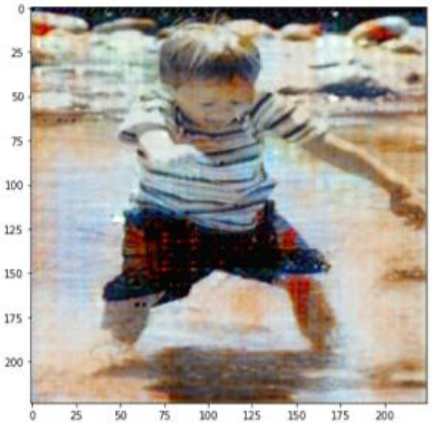


Fig 26: *Single Image colorized by GAN; further discussion on colorization results found in Section 10.*

9.0 Evaluate Model on New Data

Our new data consists of a subset of Flickr30k and Imagenet which were not used in training or validation. We also analyse our model's performance on a constructed dataset in the use cases the model can be applied to determine how well it generalises in each of these fields. Only qualitative results will be analysed since without ground truth images a loss-function is difficult to define.

9.1 New Images from Flickr30k and Imagenet

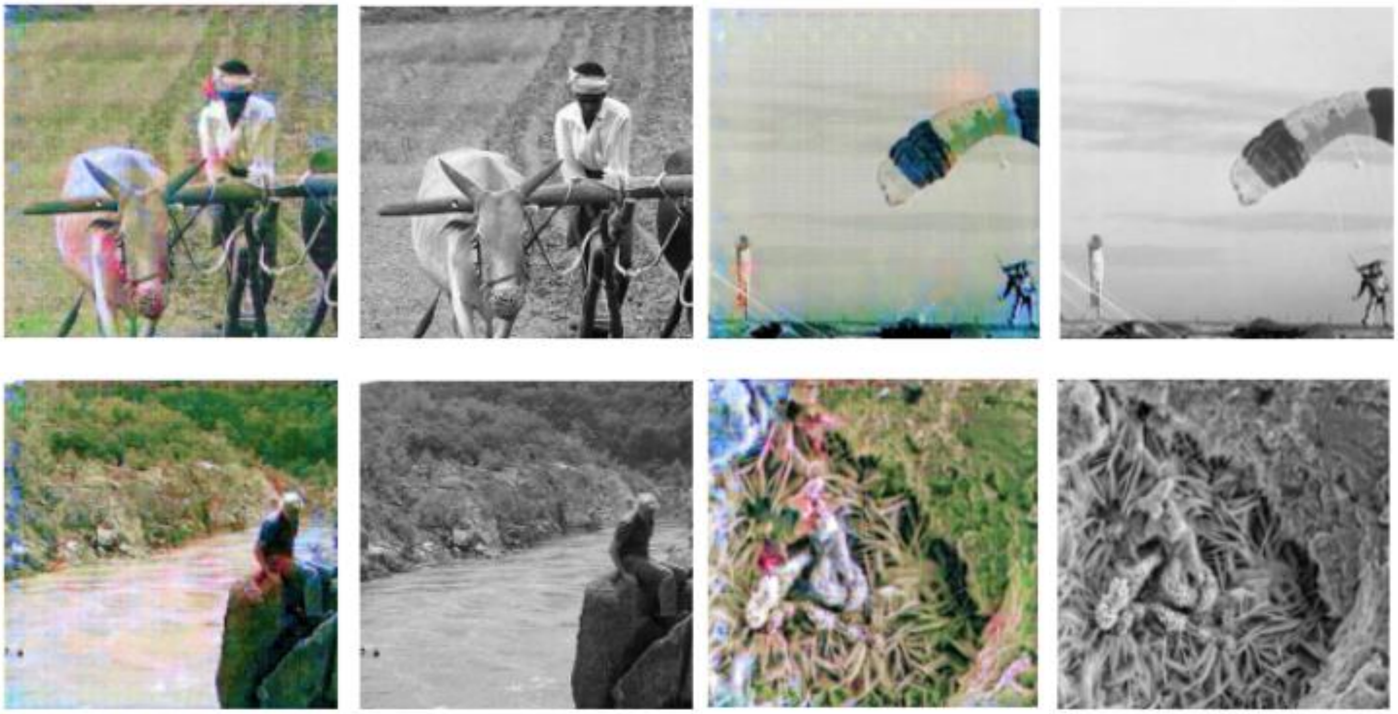


Fig 27: *Test results of colorising random images in our datasets*

The model tends to default to the safest color that is brown

Test images from Imagenet and Flickr30k were very good when compared to the model's performance on training images. This was expected since the shapes and a lot of other features will have been learned. The colorization techniques will perform very well on these since they are the same "flavour" as the training images. These grayscale images also used the same conversion method to grayscale (skimage) as the method used on training images.

9.2 New Images from Constructed Application Dataset

Historical Images

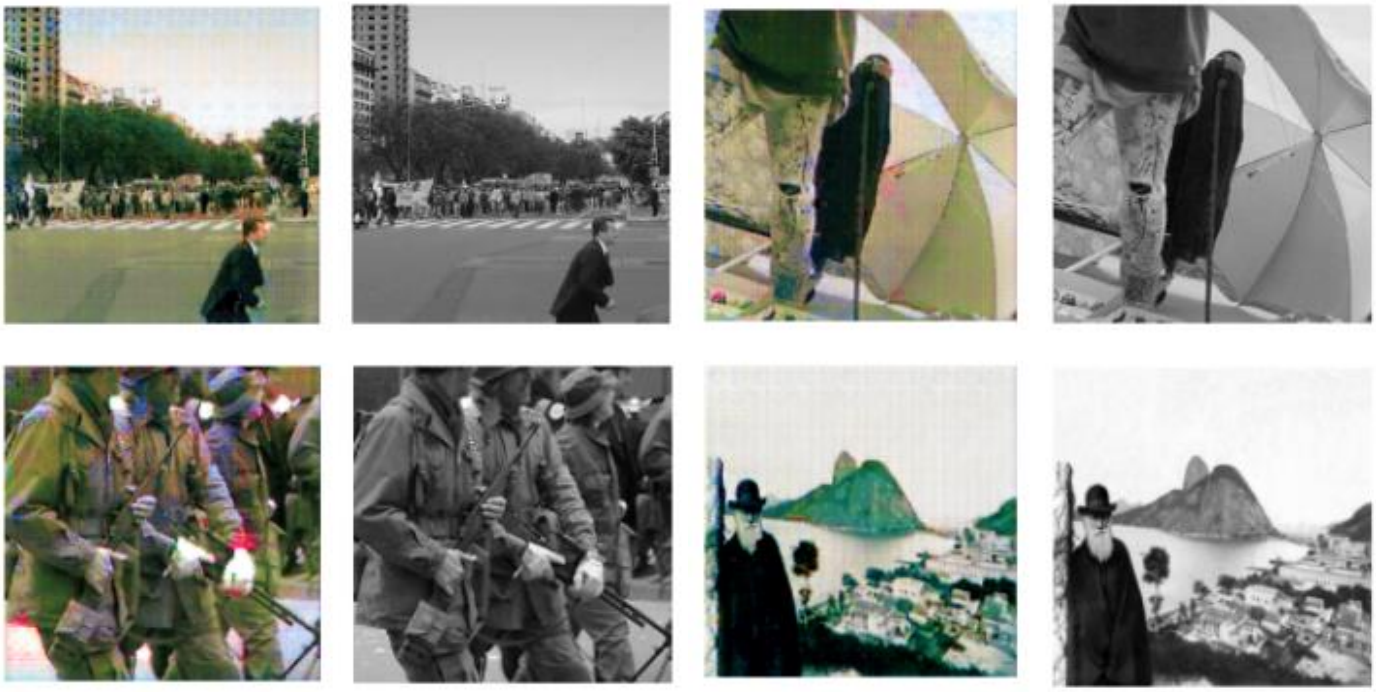


Fig 28: *Figure showing colorization of historical grayscale image.*

The performance on these images were very good. There is no definite way to verify the color accuracy, but as it is consistent in terms of color blending, lighting intensity, and basic human color intuition, we can be confident in our model. It is particularly good at sceneries and the faces seen on people are very convincing reproductions.

The image defects present appear to be ignored by the model. This is not the most aesthetically pleasing results but the model is robust enough to not allow it to disrupt its predictions

Astronomical Images

For astronomical images, scientists use filters to record the wavelengths and determine what the color should be. So in this section, we compare the qualitative colorization of the astronomical image colored manually with our GAN's generated colorized image.

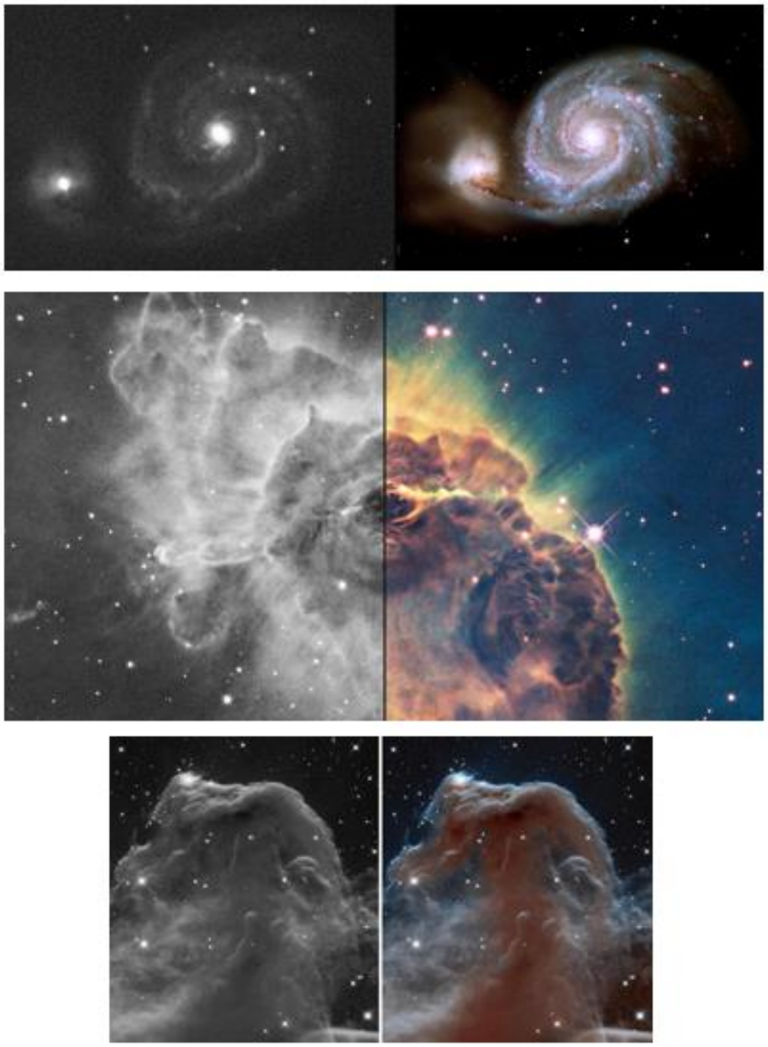


Fig 29: *Figure showing colorization of astronomical grayscale image done manually [14].*

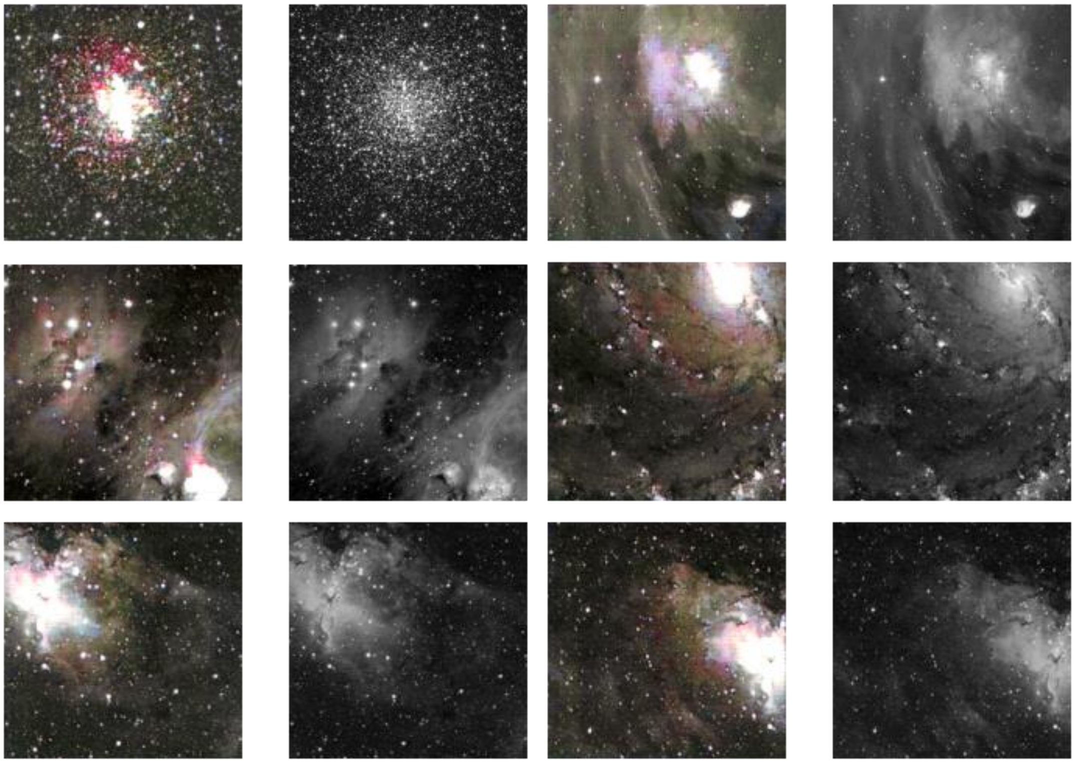


Fig 30: *Figure showing colorization of astronomical grayscale image by our model.*

The model surprisingly did a good job at determining the lightness and the saturation of the image. The depth of the image is also captured well and the shading of the image seems to be consistent with the manually colored one. This could be due to the presence of astronomical images in the Flickr30k dataset.

Medical Images

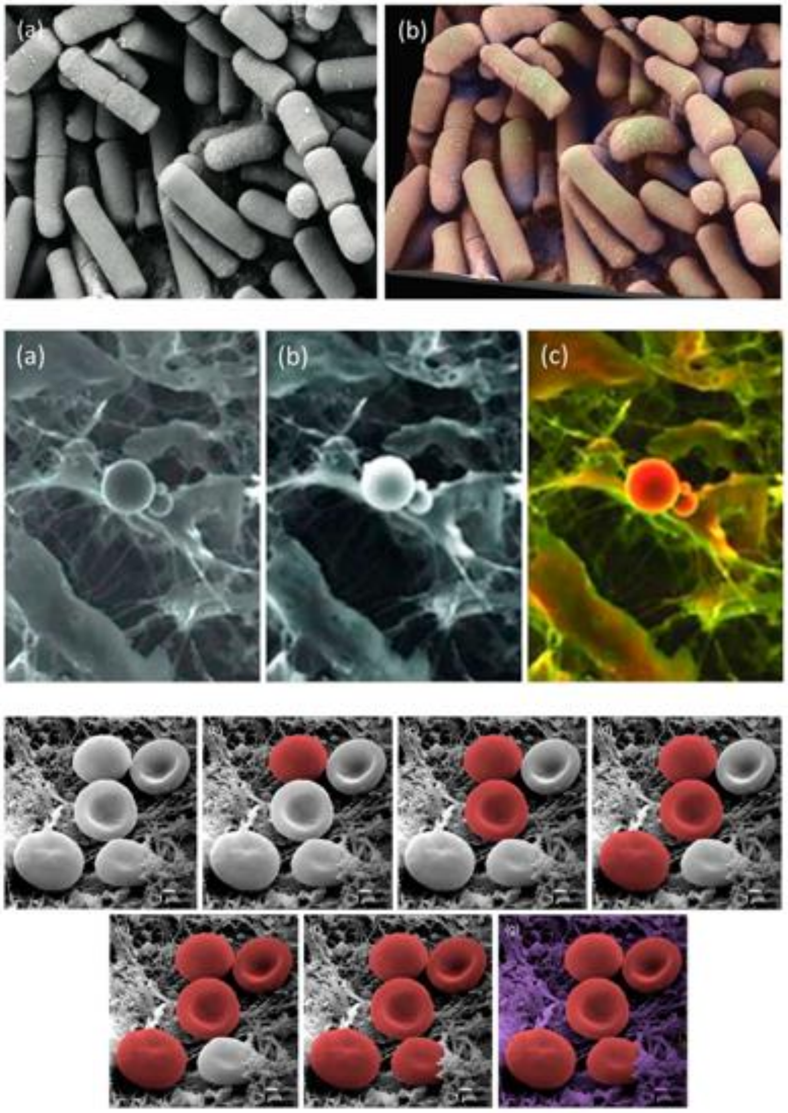


Fig 31: *Figure showing colorization of microscopic images done manually [5].*

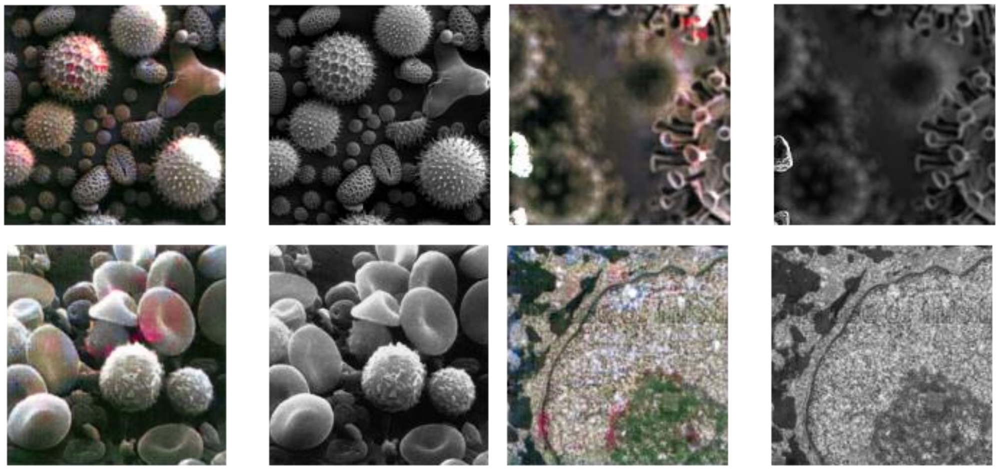


Fig 32: *Figure showing colorization of microscopic image by our model.*

The results we got are below satisfactory. Similar to Astronomical Images, we do not have access to true colorized versions of Medical Images.

The manually colourised images contain a larger degree of saturation and more colors [5]. Our model attempts to colorize these images but it is clearly not recognising the large extent of magnification that has taken place in capturing these images. This is understandable as there is nothing remotely close to these images in our training phase.

This identifies a very good limitation of our model and indicates that these types of images should either be used in training or for more concrete results this problem should be solved with a specific model for this magnification level.

10.0 Discussion

10.1 Regression Discussion

The regression baseline model didn’t give us the best colorization result. It serves as a good reference point , but isn’t excellent. The unusual result we got was the partial improper colorization of the image. This could be due to the fact that we used MSE which isn’t the best loss function for multimodal problems. It gives better results on desaturated colors than a highly vibrant image.

10.2 CNN Discussion

For the classification model we were able to improve on the results from a more vanilla classification based architecture by shifting to U-Net, which adds residual connection and strided convolutions for added benefits. This bumped our validation accuracy, by fixing the inherent difficulties present in linear CNN bases learning architectures (like weight decay and sparse weight matrices) and resulted in sharper and clearer colors for our outputs. Even though it was trained on for about 100 epochs, given the loss was calculated on RGB space; the model wasn’t able to colourize the image exactly. Moving to a 2-channel loss space could have possibly improved our results, since the model will only have to learn the feature relations of two channels instead of three. Moreover training over more images and a larger model could have helped learn the precise colourings better; as the results seem to be the fader versions of ground truths.

10.3 GAN Discussion

As can be seen in the loss graphs, we note fluctuations in the graphs instead of the typical decrease in loss that is found from most CNN models. The reason for this is the competition between the two models (generator and discriminator) -As one gets better, the other gets worse and the loss graphs follow this same trend. Shown in Fig. 6 and Fig. 7 we see that the general colors are able to be colorized but we find that the intensity of colors is much less. In Fig 7, the color of the water is correctly identified but the edges of the water are colored brown, most likely due to the fact that there are no identifiable features near the edges of the water so the model has a difficult time coloring those sections. Additionally, GANs generally increase in performance as you increase the number of training examples and the duration under which the model is learning which could be a factor.

11.0 Ethical Considerations

11.1 Discriminationatory conduct by AI

To combat Representation Bias and prevent Disparate Treatment , while we can’t give any a hundred percent guarantee, we used multiple datasets to reduce the susceptibility of our AI-based conclusions to embedded biases and discriminatory outcomes [18].

11.2 Misleading conclusion by AI

There are certain species of flowers and plants which can be identified as dangerous by the plant's shape or color. There are instances of the predicted flowers colors not matching the true color. While this is a minor issue for visual aesthetics it can be potentially hazardous in the form of allergic reactions to pets or humans [19].

11.3 Privacy Issues

As we didn’t want to use personal images of an individual, whose consent we don’t have. Only images which are legal under the Personal Data Protection Act were used [20].

12.0 Project Difficulty/ Quality

Given various grayscale images, we are predicting the colors and colorizing the images.
This is a difficult problem for many reasons, one of which being that it is ill-posed and doesn't have a single possible answer i.e. for a single grayscale image, there is a possibility of multiple, equally valid colourings. This itself brings color on to the difficulty of the problem in hand.

Our initial architectures and solution formulations add to the difficulty. The regressions loss function (MSE) had a loss function dependent on image vibrancy, we had to try multiple loss functions to get a balance between model complication and results. Once we got a sufficiently working regression, we moved on to classification.

These models both learn pixel wise, instead of learning and producing the image in one shot (like GANs). Moreover the representation power of these networks, given the limited size and available computer, isn't too strong. These inherent properties of the models made the training on bigger images extremely challenging and compute heavy. Hence, we had to reside to choosing CIFAR-10 as our dataset of choice, because of the size of the images of 32*32 pixels. Furthermore, we restricted the training and testing to only the "horse" category to further help on the computing time. This allowed us to train our model for a greater number of epochs and achieve considerably good quality results for colourization given the expressibility of our networks and the computation capability available at our disposal. Talking about our final model, GANs are computationally heavy to train and require a lot of images and training time Hence, we weren't able to train the generator to the point where it fools the discriminator consistently and achieves the singularity; but it still was able to achieve decent results as can be seen in above sections.

13.0 References

[1] Arshiya Sayyed, Apeksha Rahangdale,Rutuja Hasurkar, Kshitija Hande, ”Automatic Colorization Of Gray-scale Images using Deep Learning”. [Online]. Available: http://ijsetr.org/wp-content/uploads/2017/04/IJSETR-VOL-6-ISSUE-4-532-536.pdf. [Accessed 12 February 2020].

[2] C.Santhanakrishnan , Neeraj Durgapal , Deepak Yadav,” A Survey On Auto-Image Colorization Using Deep Learning Techniques With User Proposition”. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1362/1/012094/pdf. [Accessed 12 February 2020].

[3] S. Titus and J. N.M., "Fast Colorization of Grayscale Images by Convolutional Neural Network", Ieeexplore.ieee.org, 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8821180. [Accessed: 09- Apr- 2021].

[4] M. Imran, Colorizing Grayscale Images, 2011. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:519159/fulltext01.pdf. [Accessed: 09- Apr- 2021].

[5] C. Mignot, "Color (and 3D) for Scanning Electron Microscopy", *Microscopy Today*, vol. 26, no. 3, pp. 12-17, 2018. Available: https://www.cambridge.org/core/journals/microscopy-today/article/color-and-3d-for-scanning-electron-microscopy/96EA3848313BFFF6437B17EFC6C01411. [Accessed 9 April 2021].

[6] R. Zhang, P. Isola and A. Efros, "Colorful Image Colorization", Computer Vision – ECCV 2016, pp. 649-666, 2016. Available: https://link.springer.com/chapter/10.1007/978-3-319-46487-9_40. [Accessed 12 February 2021].

[7] G. Larsson, M. Maire and G. Shakhnarovich, "Learning Representations for Automatic Colorization", Computer Vision – ECCV 2016, pp. 577-593, 2016. Available: https://link.springer.com/chapter/10.1007/978-3-319-46493-0_35. [Accessed 12 February 2021].

[8] J. Hwang and Y. Zhou, "Image Colorization with Deep Convolutional Neural Networks", Cs231n.stanford.edu, 2016. [Online]. Available: http://cs231n.stanford.edu/reports/2016/pdfs/219_Report.pdf. [Accessed: 12- Feb- 2021].

[9] R. Dahl, "Automatic Colorization", Tinyclouds.org, 2016. [Online]. Available: https://tinyclouds.org/colorize/. [Accessed: 12- Feb- 2021].

[10] "Information on the CIE LUV colour space", Cs.haifa.ac.il. [Online]. Available: http://cs.haifa.ac.il/hagit/courses/ist/Lectures/Demos/ColorApplet/me/infoluv.html. [Accessed: 12- Feb- 2021].

[11] "ImageNet", Image-net.org. [Online]. Available: http://www.image-net.org/. [Accessed: 12- Feb- 2021].

[12] "Flickr Image dataset", Kaggle.com. [Online]. Available: https://www.kaggle.com/hsankesara/flickr-image-dataset. [Accessed: 12- Feb- 2021].

[13] A. Krizhevsky, V. Nair and G. Hinton, "CIFAR-10 and CIFAR-100 datasets", Cs.toronto.edu, 2021. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html. [Accessed: 12- Feb- 2021].

[14] "Hubble Space Telescope Images", *NASA*, 2021. [Online]. Available: https://www.nasa.gov/mission_pages/hubble/multimedia/index.html. [Accessed: 09- Apr- 2021].

[15] "National Photo Company Collection: Prints & Photographs Online Catalog (Library of Congress)", *Loc.gov*, 2021. [Online]. Available: http://www.loc.gov/pictures/search/?st=grid&co=npco. [Accessed: 09- Apr- 2021].

[16] "Image Colorization", Kaggle.com. [Online]. Available: https://www.kaggle.com/shravankumar9892/image-colorization. [Accessed: 12- Feb- 2021].

[17] Olaf Ronneberger, Philipp Fischer, Thomas Brox,“U-Net: Convolutional Networks for Biomedical Image Segmentation” Available: https://arxiv.org/abs/1505.04597 [Accessed 12 February 2020].

[18] ” Artificial Intelligence: examples of ethical dilemmas”. [Online]. Available: https://en.unesco.org/artificial-intelligence/ethics/cases [Accessed 12 February 2020].

[19] David Beaulieu,“Identifying 15 Common Poisonous Plants” Available: https://www.thespruce.com/pictures-of-poisonous-plants-2132624[Accessed 12 February 2020].

[20] Rishab Bailey,“Comments on the (Draft) Personal Data Protection Bill, 2018” Available: https://www.meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf [Accessed 12 February 2020].