Historical and Modern Image-to-Image Translation with Generative Adversarial Networks

Vinson Luo
Physics
Stanford University
vluo@stanford.edu

Michael Straka
Mathematical & Computational Sciences
Stanford University
mstraka2@stanford.edu

Lucy Li
Symbolic Systems
Stanford University
lucy3@stanford.edu

Abstract

We investigate the use of Cycle-Consistent Adversarial Networks (CycleGANs) for the translation of historical images to their modern color equivalents. CycleGANs have previously been used for a variety of geometric and color-based image-to-image transformation tasks, but here we face the particular challenge of complete colorization of a historical photograph using an unpaired training dataset. We hypothesize that due to differences between modern and historical grayscale photographs, ordinary approaches for translating grayscale to color with paired datsets will be insufficient.

1. Introduction

A simple image editor can easily convert a color image to grayscale, but the reverse is more difficult. Doing so requires information about common associations between colors and objects in the world: grass tends to be green, but how does one choose colors for flowers? Were suits that appear to be dark gray in an old photo actually dark gray in real life? Image colorization for historical photos can be particularly challenging, simply because in the majority of cases there is no ground truth coloring for such images.

In this present paper, we hope to modernize historical photos. This is primarily a colorization problem, as all of our historical photos are in black and white. However, we hypothesize that translation of early photographs to modern equivalents requires additional knowledge of the photographic styles associated with earlier time periods and present day. Differing color balance and other camera specific artifacts found in historical photos make the task of historical to modern image translation fundamentally different from the standard grayscale to color translation.

We design a model that can perform this image translation task after training on two separate sets of historical and modern images. Because our data is entirely unpaired, we intend to attack this problem using a Cycle-Consistent Adversarial Network (CycleGAN) [22], which has been shown to yield promising results in unpaired image-to-image translation. In order to constrain the learned mappings and facilitate training, we also incorporate elements of direct grayscale to color regression into our model.

2. Related Work

The basic image colorization problem (translation between modern grayscale and color images) has been attacked in a multitude of ways. Some of the earliest approaches to image colorization were non-parametric, relying on color reference images similar to a target grayscale image as source data for the colorization task. Using the Image Analogies framework [6], these approaches then transfer color onto target grayscale images using analogous regions from the source data [19, 4, 14, 15]. While these colorizations are often quite realistic, they also require similar source data in order to function well.

Other works approach the colorization problem from the perspective of regression. The large amount of easily obtainable paired data for the image colorization task makes Convolutional Neural Networks (CNNs) well-suited for this particular task. CNN-based regression methods have been able to successfully colorize images, surpassing the performance of non-CNN state of the art models without the need for curated reference data [1, 7]. Regression methods using standard L1 or L2 losses, however, can suffer from desaturation in output images, as they are trained to predict average color value over the distribution of possible colors for each pixel. For objects that have clear multimodal or nonconvex color distributions, this is undesirable, and several methods involving regression on histograms or classification of discretized colors have been applied to produce more vibrant colorizations [10, 21].

Generative Adversarial Networks [3] (GANs) provide an additional way to attack the colorization problem, learning to discriminate the characteristic features of the modern color domain. Isola *et al.* trained a "U-net" CNN generator [17] augmented with a PatchGAN discriminator [11] in an adversarial framework to realistically color grayscale photographs. The use of the PatchGAN discriminator (which only operates over local regions of the image) helped the generator learn to colorize higher frequency features in images, while a traditional L1 loss was used to capture lower frequency features.

The approaches described thus far have all used paired training data where the correct colorization of every image is known. More recent work has investigated the ability of GANs to also provide a structured framework for unpaired image-to-image translation [12, 13]—in this case, the discriminator constrains the translation output to match the output domain. To combat the issue of mode collapse, Zhu *et al.* introduced the idea of adding a cycle-consistency loss to constrain image translation output to contain much of the information of the input [22]. They applied the CycleGAN framework to several different image-to-image translation problems, including artists' styles and photos, apples and oranges, zebras and horses, winter and summer, and maps and aerial photographs. They achieved good results on color and texture changes and struggled more with geometric changes, such as translating between dogs and cats. We use CycleGANs as the starting architecture for our historical to modern image translation problem [1].

3. Methods

3.1. Default CycleGANs

Formally, our task reduces to translation between the domain of historical images X and the domain of modern images Y. The basic GAN architecture attempts to learning two mappings $G : X \to Y$ and $F : Y \to X$ that translate between the two domains, and introduce two discriminators D_X and D_Y that compete against our mappings. Specifically, D_X outputs the probability that its input is a true $x \in X$ as opposed to a generated $F(y)$, while D_Y does the same for $y \in Y$ and $G(x)$.

However, this alone is a highly unconstrained problem, as there exist many undesirable mappings that can complete this task. A G that simply maps all x to a random images in Y, for instance, may be difficult for D_Y to discriminate but does not complete the desired task of translation. To counteract this, we introduce the additional constraint of cycle-consistency: for any x and y, we force $F(G(x)) \approx x$ and $G(F(y)) \approx y$. In essence, this forces images to retain the parts of their representation that are independent of the domains X and Y during translation so that reversing the translation is possible.

Mathematically, this is implemented through the use of hybrid loss function with two components: $\mathcal{L}_{GAN}$ and $\mathcal{L}_{cyc}$. $\mathcal{L}_{GAN}$ is composed of two terms, one for F and and one for G:

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_Y) = &\sum_{y \in Y_{data}} \log D_Y(y) \\ &+ \sum_{x \in X_{data}} \log(1 - D_Y(G(x)))\end{aligned}$$

$\mathcal{L}_{GAN}(F, D_X)$ is defined similarly.

The cycle-consistency loss also consists of two terms measuring the distance between $F(G(x))$ and x, $G(F(y))$ and y:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) =& \lambda_1 \sum_{y \in Y_{data}} \|G(F(y)) - y)\| \\ &+ \lambda_2 \sum_{x \in X_{data}} \|F(G(x)) - x\|\end{aligned}$$

where $\| \cdot \|$ is a norm of our choice (we use the ℓ_1 norm, following the approach taken in [22]). Hyperparameters λ_1 and λ_2 denote the importance of cycle-consistency losses for modern and historical recreations, respectively.

We combine these losses:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) =& \mathcal{L}_{GAN}(G, D_Y) \\ &+ \mathcal{L}_{GAN}(F, D_X) \\ &+ \mathcal{L}_{cyc}(G, F)\end{aligned}$$

Our goal, then, is to find the G and F minimizing:

$$\max_{D_X, D_Y} \mathcal{L}(G, F, D_X, D_Y)$$

3.2. Training Enhancements

For more stable training, we make three practical modifications to the typical GAN losses, as suggested by [18]. Firstly, we use squared distance rather than log probability as our loss function in $\mathcal{L}_{GAN}$. Secondly, rather than directly minimizing the GAN loss for the generator, we instead maximize the probability that the discriminator incorrectly classifies generated samples. These two enhancements improve the gradients of the loss function when the true and predicted labels are far apart. Lastly, we smooth our labels into soft labels: labels of 1 are mapped uniformly to values between 0.85 and 1.15, and labels of 0 are mapped uniformly between -0.15 and 0.15. This addition of random noise helps prevent degeneracy in the generator and discriminator.

3.3. Modified CycleGAN

We found that training the basic CycleGAN model on completely unpaired data produced entirely unrealistic colorizations with almost competely incorrect colors. Even

[1] Zhu *et al.*'s PyTorch implementation, on which we base our code, can be found here.

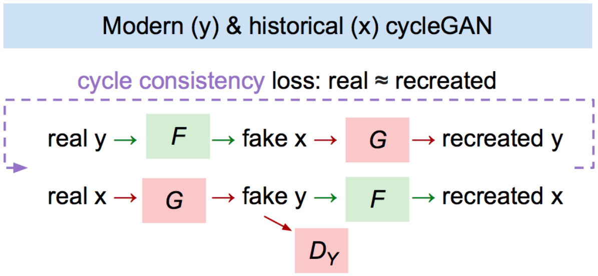
Modern (y) & historical (x) cycleGAN
cycle consistency loss: real ≈ recreated
real y → F → fake x → G → recreated y
real x → G → fake y → F → recreated x
D_Y

Figure 1. A diagram of our model used for colorizing historical images.

the modern to historical conversion, which can be approximated with a simple linear function, performed poorly, swapping white and black at times.

To stabilize our training process, we replace F with a hard-coded color to grayscale function. For a color image with channels R, G, and B, we take as its grayscale version a single channel image $0.21R + 0.72B + 0.07G$ [16]. The use of this hard-coded F function leads to the removal of D_X, leaving only one generator G and one discriminator D_Y in our model (Figure 1).

While somewhat similar to ordinary paired image-to-image translation, our method of fixing F in a CycleGAN framework allows more nuance in the treatment of historical images. G is not supposed to be a perfect inverse of F, as differences exist between our modern color to grayscale conversion function F and the true modern color to historical grayscale conversion function. The fact that the closeness of this inversion is controlled by two tunable cycle L_1 losses means that only approximate inversion is enforced, allowing for more nuance in the historical to modern translation. We can manually adjust the hyperparameters λ_1 and λ_2 to achieve the desired balance between realistic conversion and stable training.

3.4. Network Architecture

For our generator G, we used the image-to-image generative network described in Johnson et al. [9]. Our generator consists of 2 downsamplings of the image, followed by 9 ResNet blocks, followed by 2 upsamplings of the image [5]. The intermediary ResNet blocks have 256 channels each, which are cut in half by each downsampling and eventually convolved into a single channel grayscale image.

For our discriminator D_Y, we use a 70×70 CNN-baseed PatchGAN discriminator, as in [22, 8]. This discriminator operates independently over each 70 × 70 pixel patch in the image, using a series of purely convolutional downsamplings to distill each patch into the probability that each patch is real or fake. The GAN loss is then computed as the average loss of the discriminator over all pixel patches.

4. Datasets

We had multiple pairings of historical datasets and modern ones. We focused on two domains: people and landscapes. We used these domains to limit the variety of objects that our model would need to learn to color. We resized all input images to 256 x 256 pixels and augmented the existing data using random cropping and horizontal flipping. Note that during testing, we generate images in same size batches for both modern recreations and historical colorizations so that the number of results is limited by the smallest test set. This means that we evaluate our test results for people colorization on 250 images and for landscapes on 95 images.

4.1. People

Our historical dataset of people consisted of scans created by the State Library of New South Wales and the University of Technology Sydney (UTS). This dataset, called the Hood Collection part I, has been shown to be usable for image labeling and retrieval tasks and dates from the 1910s to 1950s [20]. We extracted photos from the Hood Collection labeled as "groups of people", removing ones that were overly damaged. We partitioned the data into 935, 262, and 265 images in our training, test, and validation sets, respectively.

Our modern dataset of people consists of images collected from Flickr by Gallagher & Chen 2009 [2] depicting people at social events. We removed grayscale photos and used a subset of this dataset to yield 1,510, 250, and 290 images in our training, test, and validation sets, respectively. Aside from color, our modern and historical datasets also differ in camera quality, racial diversity, and clothing styles. Historical images tend to feature fewer close-ups of people's faces and only white people. The modern dataset also featured many weddings.

4.2. Landscapes

Our historical dataset of landscapes were downloaded from the Wikimedia Commons categories ”Photographs of the Zion and Bryce Canyon National Parks, 1929, compiled 1929 - 1929”[2] and ”US National Archives series: Ansel Adams Photographs of National Parks and Monuments, compiled 1941 - 1942, documenting the period ca. 1933 - 1942”[3]. We created a training set of 400 images and a test set of 95 images.

Since our historical landscape dataset featured national parks in the Southwestern United States, for our modern landscape dataset we used Wikimedia Commons category ”Zion National Park”[4]. We removed grayscale and non-nature images from this set and had a training set of 406

[2] Historical Zion and Bryce Canyon photos can be found here.
[3] Ansel Adams photos can be found here.
[4] Colored Zion National Park photographs found here.

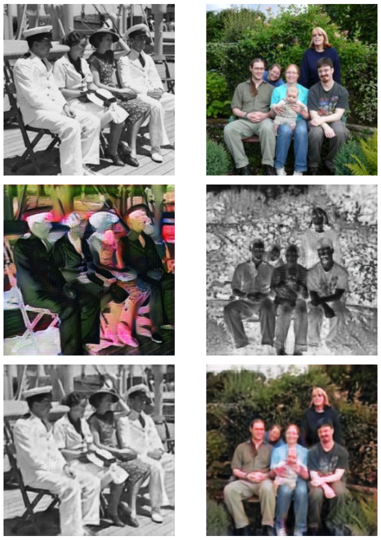

Figure 2. Results at the 10th epoch of training. Top row contains original photos with the historical image on the left and the modern image on the right. The middle row has generated ”fake” photos, and the bottom row the reconstructed images.

images and a test set of 166 images. Grayscaled modern landscapes, to a human observer, resembled historical landscape images more so than grayscale modern people with historical people images. This is likely due to the more timeless and almost-identical subjects of the historical and modern landscape datasets.

5. Results

5.1. Hyperparameters

For almost all of our hyperparameter values, we choose the values used in Zhu *et al.* [22], as they have been demonstrated to work well in image-to-image translation tasks. We use the Adam optimizer with an initial learning rate of 0.0002 and a momentum of 0.5 to optimize our parameters. In order to stabilize gradient updates and take full advantage of our GPU, we used the highest batch size possible given our memory constraints, which turned out to be 4 for our particular choice of generator.

The two hyperparameters λ_1 and λ_2 were the hyperparameters that we experimented most on. We found that values of $\lambda_1 = 13$, $\lambda_2 = 10$ yielded the most reasonable colorizations, striking a balance between the implausible color guessing associated with low values of these two parameters and the sepia-like colorization associated with high values of these two parameters.

5.2. Default cycleGANs

Using Zhu et al.'s [22] PyTorch implementation of CycleGAN, we conducted a preliminary investigation of CycleGAN's ability to generate grayscale and color images[5]. We produce the final losses in Table 3 and images in Figure 2. These results are less than stellar. It seems that the model inverted the colors for $X \to Y$ and $Y \to X$ so that lighter colors or shades became black, and darker colors or shades became white. While the model seemed to be detecting edges, it was nevertheless unable to color objects reasonably. At the same time, however, the recreated images $F(G(x))$ and $G(F(y))$ shown in Figure 2, often matched the originals quite closely.

	Baseline train	Baseline val
cyc_X	1.122	1.028
cyc_Y	1.111	0.8297
D_Y	0.3788	0.3919
G	0.3741	0.2926
D_X	0.3424	0.1420
F	0.2543	0.09578

Table 1. Training set and validation set losses on batches of size four using our baseline model, which was trained through 10 epochs. cyc_X and cyc_Y are the two terms that are added to produce the cycle consistency loss $\mathcal{L}_{cyc}(G, F)$. D_Y and D_X are intermediate discriminator losses, and G and F are the generator losses $\mathcal{L}_{GAN}$.

	people		landscapes	
	train	test	train	test
cyc_X	1.017	0.317	1.329	0.560
cyc_Y	0.918	1.448	1.458	1.169
D_Y	0.137	0.641	0.103	0.293
G	0.676	1.424	0.686	0.084

Table 2. Training set and test set losses on batches of size four using our final model, trained through 52 epochs. cyc_X and cyc_Y are the two terms that are added to produce the cycle consistency loss $\mathcal{L}_{cyc}(G, F)$. D_Y is the discriminator loss, and G is the generator loss $\mathcal{L}_{GAN}$.

	L1	L2
People	0.0298	0.0114
Landscapes	0.0968	0.0384

Table 3. Average per-pixel L1 and L2 losses computed between test sets of 250 original people photos and their recreations and 95 original landscape photos and their recreations.

5.3. Modified cycleGANs

We modified our model based on training progress and validation results of the people datasets. Hard-coding $F : Y \to X$ increased the stability of our model and prevented the grayscale generator from ”cheating” or including residual color in its output. It also prevented the ”inversion” of

[5] Details on the PyTorch Python package can be found here.

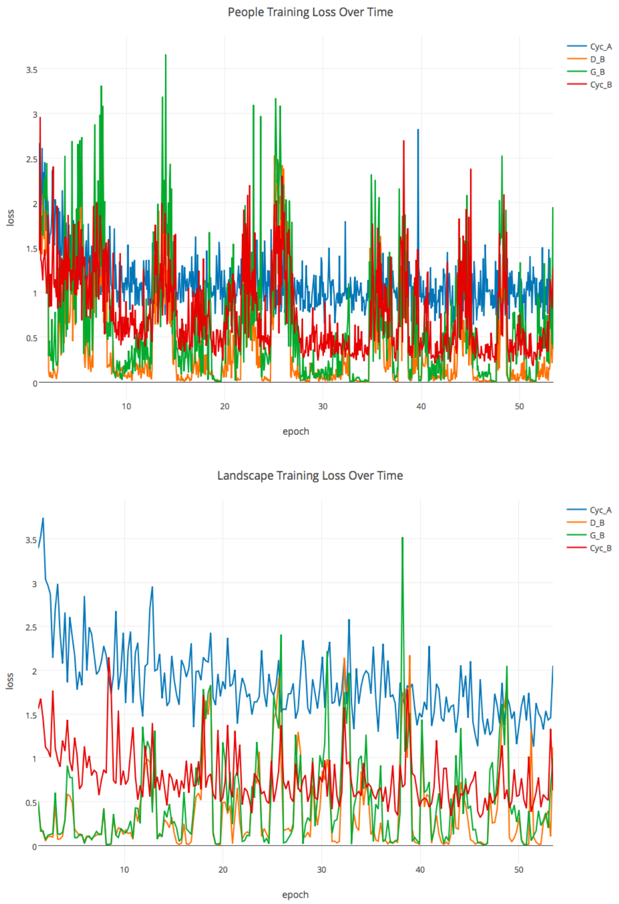

Figure 3. Training loss over 52 epochs with our people (top) and landscape (bottom) datasets. D_B (= D_Y) and G_B (= G) are discriminator and generator losses for colorized historical images. Cyc_B is the cycle consistency loss based on the L1 difference between original historical images and their recreations. Cyc_A is the cycle consistency loss between original modern images and their recreations.

faces and other object colors that we saw with default CycleGANs, and some images within 10 epochs of training featured skin-colored faces. We ran our models for people and landscapes for 52 epochs along with batch normalization and dropout. As seen in Figure 2, our cycle consistency losses seemed to plateau near the end, while our generator and discriminator losses oscillated far more[6].

Historical colorizations are difficult to evaluate due to the lack of ”ground truth.” We determined the quality of our generated historical colorizations using our own subjective judgment. We also evaluated our model by calculating the L1 and L2 losses of our recreated modern test set images (Table 3), but it is important to note that a model that creates quality modern recreations does not necessarily create quality historical colorizations. For example, our modified model produced the historical colorizations seen in Figures 4, and our modern recreations did not feature colors that were as bright as the originals (Figure 6). However, if we attempt to brighten these recreations by appending discriminator D_Y after their generation to create an additional loss factor for D_Y and G, our historical colorizations hallucinated unrealistic color patches to an even greater degree. In general, colorization patterns such as those seen in the bottom row of Figure 4 occurred more prominently in historical colorizations than in modern recreations.

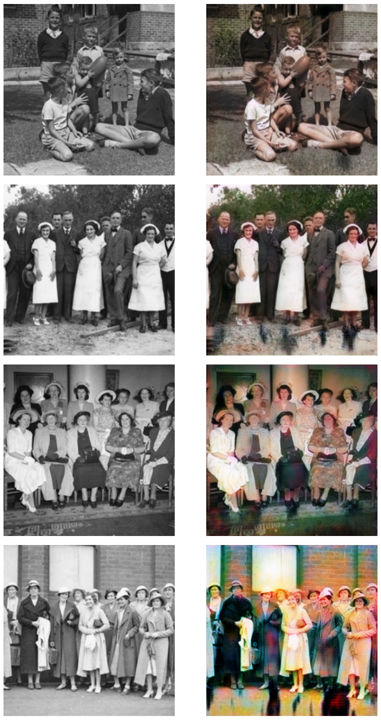

Figure 4. Reasonable historical colorizations usually depicted skin-colored faces, green foliage if any, and subdued brown tones. Less ideal colorizations, such as the one in the bottom right, involved hallucinating bright yellow, red, green, or turquoise patches. The originals are on the left and the generated images on the right.

We also compared our results qualitatively to the results obtained by [21] by computing their model on our original black and white photos and comparing them to the color photos generated by the same images by our model (Figure 7). In general we found that their model tended to opt for brighter colors when dealing with multimodal problems

[6] For our people model, we designed it so that with some probability p, we would conduct ”paired training.” Real x, as seen in Figure 1, would be a grayscaled modern photo, fake y would be a modern colorization, and recreated x would be recreating the grayscaled modern photo. We calculated a ”direct loss” between fake y and real y, with a weight of 10. Our cycle consistency loss weight between real y and recreated y had a weight of 10. We later discovered that this ”paired training” model with $p = 30\%$ is theoretically equivalent to setting the cycle consistency weight on modern originals and their recreations to 13.

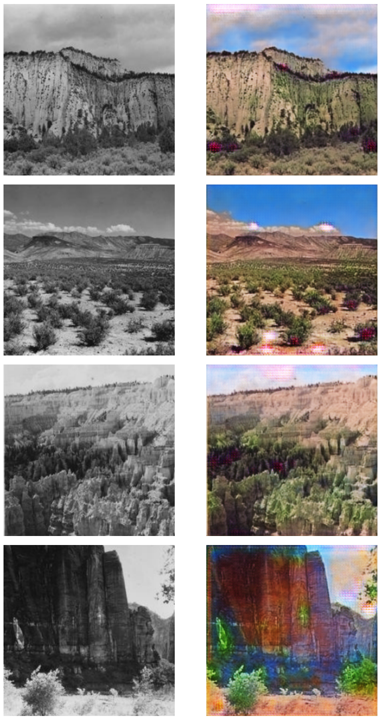

Figure 5. Our results for colorization of historical landscape images. Originals are on the left and colorizations are on the right. The bottom row is an example of a failure case. Like for images with people, failure cases usually involved bright hallucinations.

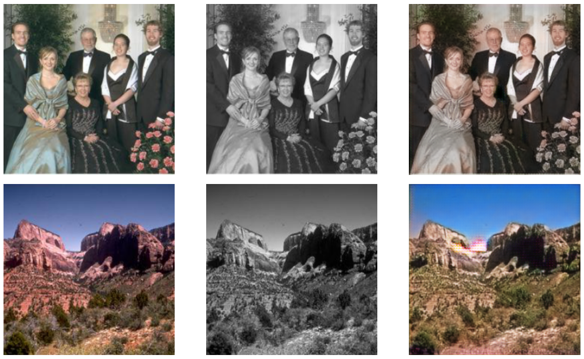

Figure 6. From left to right: original modern image, grayscale image, and our recreated image.

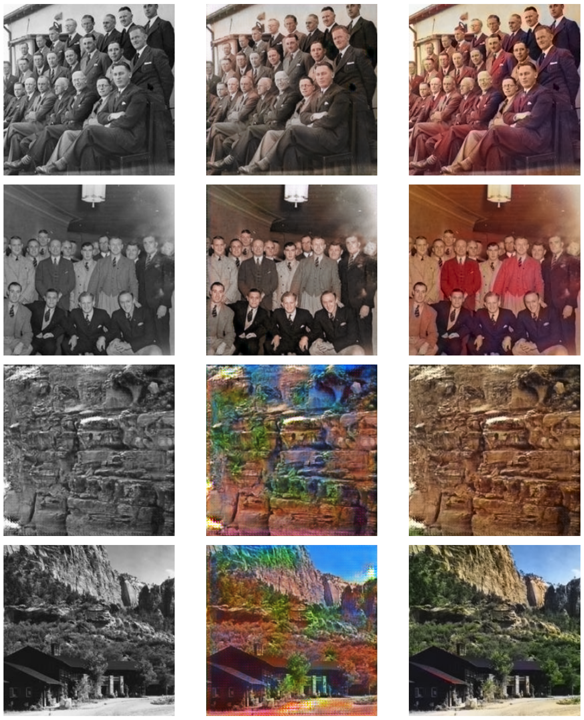

Figure 7. From left column to right column: original historical images, our colorization, and [21]'s colorization.

with many plausible outcomes, such as coloring clothing bright red. In addition, it occasionally produced images that were almost entirely colored red in an unrealistic hallucinartory fashion. This is in contrast with our model, which typically produced images with more muted sepia-toned shades of color. Their landscape colorizations were more consistently reasonable than ours. It is important to note that their training set, ImageNet, is far more diverse than ours.

Our model, when trained and tested on the people datasets, was able to capture skin colors and sometimes green grass and blue skies, but it also often reduced images to sepia or hallucinated vibrant patches (Figure 4). Surprisingly, our modern recreations were able to color not only white people but also other races. Clothing, due to its diversity of possible colorings, was especially a challenge. In earlier epochs of training, the model would occasionally color skies as blue, but it would also mistakenly color indoor walls as blue. In later epochs, our model ended up coloring fewer backgrounds, including skies, as blue.

Taking in consideration the challenges we faced with the people dataset, we decided to try a landscape dataset due to its simplicity of color choices. Our landscape test results were more convincing than those for our people datasets, likely due to this simplicity. Our model training on the landscape dataset was able to establish within the first epoch that skies are blue and mountains are gray, brown, or green. Our training results were consistently plausible. It may be that our model overfit its training data, because our validation results, as pictured in Figure 5, featured some light hallucination and pixelated pink patches around clouds and other white elements.

One issue we encountered when training and evaluating our model was its sensitivity to the domain being worked on. Our model performed better when training on our dataset of people than it did when training on our landscapes dataset, in terms of our losses and average L1 and L2 differences between original images and their recreations. We suspect that this is because the landscape photos had fewer clearly defined objects and edges, as would be the case when comparing an image of two mountains against an image of several people. The additional complexity added to the images may have allowed the model to fine-tune its output more accurately, whereas with the landscape images it frequently "hallucinated" patches of green due to the lack of structure. We encountered a different issue, however, in experiments we performed on a dataset of 1866 images of Mickey Mouse cartoons we collected from 20 black and white episodes from the 1930s and 37 color episodes produced last year. On this dataset our model could only roughly identify edges and appeared to hallucinate colors at random. We suspect that this was due to the lack of consistency in the images in the color dataset, which depicted a variety of absurd and detailed cartoon scenarios. Should we have had more time and resources, we would have tried a dataset composed of less complicated cartoons. These experiments suggest the importance both of the domain being worked on, and the dataset being trained on within that domain.

One interesting detail about our quantitative results is that our average pixel-wise L1 and L2 losses computed between original and recreated photos was low for both our people dataset and our landscapes dataset. This is despite the landscape recreations occasionally having bright, unrealistic colors in them, such as splotches of purple. There seems to be a difference between comparing images pixel-wise and subjectively evaluating the plausibility of a generated image. Given more time and resources, we would have liked to quantify our human judgments of results by surveying multiple human participants.

6. Conclusion

We trained our model on datasets of people and landscapes, and discovered that cycleGANs is best at coloring objects with consistent colors, where there is only one plausible color to choose. Our results on the people datasets had two major flaws: overly desaturated and sepia images, or overly bright and incoherent images. These two failure cases seem to be opposites, making it difficult to solve for one without making the other worse. It may be interesting to explore whether a cycleGAN model would become more confused or be able to generalize better with a more diverse training set.

Future steps include exploring other color spaces, such as YUV, where only two channels, U and V, need to be predicted for colorization tasks. As a preliminary study, we tried using the default cycleGANs model to colorize images converted to the YUV color space. It tended to perform worse than RGB, especially by hallucinating white, yellow, and turquoise patches. Since our modified model was able to produce more plausible results with RGB than the default model, YUV outcomes may also improve with our modifications.

Acknowledgements

We would like to thank the staff of Stanford's CS 231N: Convolutional Neural Networks for Visual Recognition for their continued support and faith in us as developing deep learning researchers. We would also like to thank Jon Gauthier for his suggestions and knowledge of GANs.

References

[1] Z. Cheng, Q. Yang, and B. Sheng. Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 415–423, 2015.

[2] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 256–263. IEEE, 2009.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] R. Gupta, A. Chia, D. Rajan, E. S. Ng, and H. Zhiyong. Image colorization using similar images. In *Proceedings of the 20th ACM international conference on Multimedia*, page 369378, 2012.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] A. Hertzmann, C. Jacobs, N. Oliver, and B. C. D. Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001.

[7] S. Iizuka, S.-S. Edgar, and H. Ishikawa. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35, 2016.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[9] J. Johnson, A. Alahi, and F.-F. Li. Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision*, 2016.

[10] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. *European Conference on Computer Vision*, 2016.

[11] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. *European Conference on Computer Vision*, 2016.

[12] M.-Y. Liu, T. Breuel, , and J. Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017.

[13] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. page 469477, 2016.

[14] X. Liu et al. Intrinsic colorization. *ACM Transactions on Graphics (TOG)*, 27, 2008.

[15] X. Liu et al. Intrinsic colorization. *ACM Transactions on Graphics (TOG)*, 27, 2011.

[16] K. Plataniotis and A. N. Venetsanopoulos. *Color image processing and applications*. Springer Science & Business Media, 2013.

[17] O. Ronneberger, P. Fischer, , and T. Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015.

[18] T. Salimans et al. Improved techniques for training gans. *NIPS*, 2016.

[19] T. Welsh, M. Ashikhmin, and K. Mueller. Transferring color to greyscale images. *ACM Transactions on Graphics (TOG)*, 21:277–280, 2002.

[20] J. Zhang, J. Zhang, J. Lu, C. Shen, K. Curr, R. Phua, R. Neville, and E. Edmonds. Slnsw-uts: A historical image dataset for image multi-labeling and retrieval. In *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, pages 1–6. IEEE, 2016.

[21] R. Zhiang, P. Isola, and A. A. Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.