VICTORIAN400: COLORIZING VICTORIAN ILLUSTRATIONS

HOYEOL KIM

Abstract *This article introduces the Victorian400 dataset, created to colorize black-and-white illustrations from nineteenth-century British novels with deep learning models. The Victorian400 dataset is a set of colour illustrations painted using nineteenth-century palettes which have been shared as open sources. This article examines the process of creating and curating the Victorian400 dataset and the validity of the Victorian400 dataset by looking into the results of the test set with the trained set based on conditional generative adversarial networks.*

Keywords: *Victorian400*, colorization, Victorian illustration, pix2pix, cGAN, Charles Dickens

I. INTRODUCTION

If we view data as a social and cultural artefact, datasets can be viewed as snapshots of the social and cultural climate of a time period. Tanya Clement and Amelia Acker assert that '[d]ata is not just given', but that culture is formed and changed by data.[1] Data is inclusive of history, life, arts and literature, and reflects contemporary culture, society and economy. Data, however, must be properly trimmed for usage in computation. 'Data curation' includes 'the act of discovering a data source(s) of interest, [then] cleaning and transforming the new data'.[2] The importance of data curation has been discussed by digital humanists, especially regarding the need to curate data in order to avoid faulty, biased or distorted results and reflect diverse and global data.[3] When data scientists look into data, they focus more on whether the data is sufficiently well curated to be trained for models.[4] Data scientists usually spend a large portion of their time curating datasets, as well-curated datasets are crucial to achieving

International Journal of Humanities and Arts Computing 15.1-2 (2021): 186–202
DOI: 10.3366/ijhac.2021.0269

www.euppublishing.com/ijhac

state-of-the-art results in the deep learning field. Datasets curated by digital humanists, however, have been mainly created for other humanists in the form of resources such as digital archives, rather than for machine learning/data scientists. Publicly shared humanities datasets for machine learning/data scientists on Kaggle and GitHub are mostly created by non-humanists. Therefore, in order for datasets to be tailored specifically to humanities-related deep learning tasks, I believe it would be beneficial for digital humanists, with their knowledge of both fields, to create, curate, analyse and share their own deep learning datasets.[5]

According to Goodfellow et al., ‘generative adversarial networks (GANs) are a class of methods for learning generative models based on game theory’.[6] Since GANs, which have proved to have an astonishing ability to create realistic high-resolution images, were first introduced by Ian Goodfellow in 2014,[7] they have been applied to a variety of imaging processes to predict the colour of images, transfer colours for a style, or produce higher resolutions in the conditional setting based on either conditional or unconditional GANs, such as image-to-image translation,[8] text-to-image synthesis[9] and image super-resolution.[10] Famously, the *Portrait of Edmond de Belamy*, the artificial intelligence artwork created by Obvious, a group of artists consisting of three French students who borrowed from Robbie Barrat’s code based on GANs, was sold for $432,500 in 2018, although there are technical defects in the *Belamy* portrait such as texture quality and low-resolution issues. In addition, there has been progress in colorizing photos and videos based on GAN-derived models, such as pix2pix,[11] DeOldify[12] and Paints Chainer.[13]

There have been attempts to colorize images by hand in order to improve the imaginational response in readers, as well as provide anticipation and interest, but hand-coloured results vary depending on the style of the illustrator and such methods are time- and cost-ineffective. In contrast, deep learning–based colorization is cheap and fast, in addition to creating consistent results based on models trained with datasets. There are a variety of possible deep learning colorizing projects in the digital humanities; colorizing black-and-white Victorian images such as those found in the works of Charles Dickens, Thomas Carlyle and John Ruskin would allow viewers to connect emotionally with the figures.[14] The colorization of black-and-white war photographs, as another example, could open up avenues for modern viewers to more easily conceptualize the past. For the project outlined in this article, I decided to tackle the colorization of Victorian illustrations, since as yet there has been no attempt to colorize illustrations from the nineteenth century using deep learning. Most illustrations from the Victorian era were printed in black and white due to the higher printing cost of colour illustrations.[15] For example, the works of Charles Dickens, who helped pioneer Victorian illustrations by actively including illustrations in his fiction, include only four hand-coloured illustrations, all printed for his story *A Christmas Carol* (1843). Colorizing the illustrations found in Dickens's works has the potential to enhance readers' understanding of the text and open up new interpretive possibilities. However, as yet no datasets of nineteenth-century illustrations have been made available for deep-learning-based colorization. As someone who has been trained for research and work in the fields of both the humanities and computer science, I have created, curated and publicly shared the *Victorian400* dataset for the deep learning colorization of illustrations from the Victorian era. The *Victorian400* dataset is a collection of Victorian colour illustrations which provides an opportunity for deep learning learners to run code easily without high-performance devices,[16] helps machine learning/data scientists to test and develop deep learning models,[17] and, as a pedagogical tool, shows the possibilities of improving students' ability to better understand the intersections of text and image.[18] In this article, I introduce the *Victorian400* dataset, reveal how I decided what to include and exclude in the process of curating the dataset, and examine the results of the test set with the trained set to see whether the *Victorian400* dataset produces reasonable results.[19]

2. *VICTORIAN400* DATASET

The *Victorian400* dataset, an open source shared on Kaggle[20] and GitHub,[21] is a nineteenth-century illustration dataset consisting of 400 illustrated images (Fig. 1). It has been downloaded, experimented with, and shared by data scientists and the deep learning community.

Regarding the curation of the 400 images for the *Victorian400* dataset, I outlined four criteria: 1) the images should be from the Victorian era; 2) the painting style of the images should not be too idiosyncratic (the drawings of Aubrey Beardsley, for example, do not represent illustrations of the Victorian age); 3) the images should derive from a variety of illustrators and not focused on a single illustrator, and 4) the drawing style of the images should be close to that of the illustrations commonly found in Victorian novels.[22]

The *Victorian400* dataset was originally created to colorize Charles Dickens's black-and-white illustrations for a pedagogical purpose. While researching into the area of colorizing illustrations, I found deep learning datasets for pictures, shoes, animals and anime, but was unable to find nineteenth-century illustration datasets created for deep learning.[23] Therefore, I had to create my own dataset, to include illustrations with similar drawing styles to those found in Dickens's fiction, in order to carry out imaginative colorizations of the Dickens illustrations. The scope of this dataset had to be expanded to all nineteenth-century illustrations due to the fact that most of the illustrations in Dickens's books exist only in black and white. I aimed to find colour illustrations that were similar in style to the black-and-white illustrations drawn by George Cruikshank, Hablot Knight Browne (Phiz) and John Leech for Dickens's works.

Figure 1. The *Victorian400* dataset.

To do this, I first coded web crawling tools to scrape together around 3,000 images from websites such as *The Victorian Web* and *The Charles Dickens Page*. Web scraping technology is beneficial to data scientists, but it requires a great deal of time to curate the scraped data. According to Aurélien Géron, 'If your training data is full of errors, outliers, and noise (e.g., due to poor-quality measurements), it will make it harder for the system to detect the underlying patterns, so your system is less likely to perform well.'[24] An extensive amount of experimentation is required to obtain reasonable results for training data. In general, the errors of training data such as overfitting and underfitting stem from poor-quality datasets, datasets that are too small, or models that are too complicated. In order to have credible data for deep learning, data must be curated based on specific and consistent criteria. Accordingly, using my pool of collected images, I spent the majority of my time in this project selecting appropriate images for the *Victorian400* dataset and curating the files in order to improve the accuracy of the colorization results. For example, I removed images that were clearly outliers, specifically those which included a preponderance of red or green, such as pictures of Santa Claus. I found that Victorian illustrations commonly feature red and green, but when, for example, the dataset included a series of illustrations coloured with a lot of green, as in scenes dominated by trees, the test results generated images with too much green, sometimes colouring in green what should not be green, for instance skin. Whenever the dataset was unbalanced, no matter how much data I tested with, the validation results were poor and biased.

The *Victorian400* dataset contains three different folders: original, resized and grey. Images in the 'original' folder were curated for significant parts such as faces and bodies, which were not to be cut in the process of resizing. In the 'resized' folder were images resized to 256×256 for the process of deep learning. Based on these resized images, I created the 'grey' folder, which includes black-and-white images converted from the resized colour images.

3. EXPERIMENTS

For colorizing illustrations from the nineteenth century, I chose the pix2pix model built based on cGANs by Isola et al. The pix2pix model performs automatic graphic operations on photographs based on cGANs by learning from datasets. The pix2pix model proved able to solve a number of issues when translating an input image into a corresponding output image by showing test results with a variety of datasets.[25] The GAN model suggested by Goodfellow et al. has the generator and the discriminator learn and compete with each other in order to generate the best outputs, which are referred to as adversarial nets.[26] Likewise, the conditional GANs that pix2pix deploys consist of two main poles, the generator and the discriminator. In pix2pix, the generator colorizes input images by transforming them into output images through several steps using a series of encoders and decoders. A series of encoders, which include convolution and activation, helps the generator compress input images throughout the layers, and a series of decoders, which contain deconvolution and activation, decompresses them. However, the pix2pix model deploys a 'U-Net' instead of simply using an encoder-decoder for the generator in order to improve performance when predicting the colours of input images.[27]

The difference between an encoder-decoder and U-Net is that the U-Net has added skip connections. The skip connections help each path achieve better localization by combining context from the connection of encoder layers (the downsampling path) and decoder layers (the upsampling path). For example, the size of feature maps in the first layer of the encoder (256×256×3) is the same as in the last layer of the decoder. Each layer of the encoder is merged with the

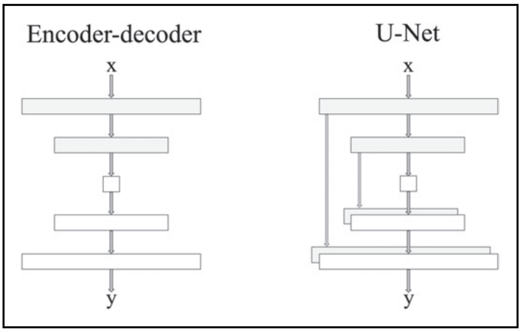

Figure 2. Simple depiction of differences between an encoder-decoder and U-Net.

corresponding layer of the decoder. U-Net, a convolutional network for image segmentation, is faster than a sliding-window approach,[28] which is a brute-force solution and often considered inefficient. A sliding-window approach, which predicts the label of each pixel by providing a local patch from input images, has two drawbacks: slowness and a 'trade-off between localization accuracy and the use of context'.[29] In contrast, U-Net makes it possible to get both local and contextual information quickly. In addition, the overlap-tile strategy of U-Net allows 'the seamless segmentation of arbitrarily large images' by extrapolating missing context through input images.[30] U-Net is an appropriate approach to run the *Victorian400* dataset as U-Net deploys massive data augmentation in order to train a model with smaller datasets.

The relationship between the generator and the discriminator can be compared to the one between counterfeiters and the police: counterfeiters (the generator) produce 'fake currency and use it without detection', and the police 'detect the counterfeit currency'.[31] Within GAN models, the generator and discriminator compete with and learn from each other to bring about enhanced results. The discriminator role is to classify whether outputs from the generator are real, based on target images, and to generate a probability for outputs being categorized successfully through image-to-image translation tasks. The pix2pix model deploys a PatchGAN, which is 'a discriminator architecture' that 'only penalizes structure at the scale of patches' instead of through a deep convolutional neural network, which is commonly used for traditional GAN models.[32] A smaller PatchGAN with fewer parameters proved to run faster when applied to 'arbitrarily large images'; the 70×70 PatchGAN was found to be most effective for image-to-image translation tasks.

In terms of validation, data scientists usually check outputs and add or delete data while running deep learning models with the datasets. There are no standards for validating datasets. Similarly, there is no agreement on which evaluation measures are the best for GANs, although there have been studies that propose measuring GAN-derived models based on quantitative and qualitative methods.[33] When evaluating GAN models, qualitative measures might 'favor models that concentrate on limited sections of the data' such as overfitting, memorizing or low diversity, whereas quantitative measures 'may not directly correspond to how humans perceive and judge generated images' despite being less subjective.[34] To validate the *Victorian400* dataset, I draw upon both quantitative (Figs 3, 4 and 5) and qualitative (Figs 6, 7 and 9) measures.

I tested the pix2pix model with the *Victorian400* dataset to verify the possibilities of colorizing black-and-white illustrations from the nineteenth century. The *Victorian400* dataset consists of 400 images, and each image has 196,608 features ($256\times256\times3$ dimensions) as each image contains 256×256 pixels with RGB (Red, Green, Blue) values from 0 to 255. In Figures 3 and 4, both the generator loss and the discriminator become more stable as the number of epochs increased. The generator and the discriminator are trained with datasets aimed at having lower losses as training iterations progress. Although the discriminator loss seems inconsistent in Figure 4, the loss changes between 1 and 1e-4, which is a very small range. The generator loss at epochs = 200 is 7.439, which is high compared to the discriminator loss at epochs = 200, which scores 1.5325e-3. From epochs≈15, the discriminator loss drops drastically, while the generator loss increases, which signifies that the discriminator is beating the generator. The characteristics of GANs are such that the generator and the discriminator improve the accuracy of outputs by competing with each other. However, the pix2pix model with the *Victorian400* dataset reveals that the discriminator generally wins over the generator. GANs are based on a zero-sum game: the discriminator is stronger than the generator in this training model with the *Victorian400* dataset. In other words, the generator finds it difficult to fool the discriminator, which is the limit of GAN models: if the generator wins over the discriminator, or vice versa, it will be difficult to properly train a model with datasets due to overfitting from the unbalance between the generator and the discriminator.[35] Although the pix2pix model was not created with the aim of predicting illustration colours, Figures 5 and 6 reveal that the pix2pix model generates reasonable outputs when trained with the *Victorian400* dataset.

The L1 loss is the mean absolute pixel difference between generated outputs and target images in the training process. In Figure 5, the L1 loss gradually diminishes, which signifies that the model has been trained properly with the *Victorian400* dataset. From epochs≈160, the L1 loss no longer decreases but the L1 loss trajectory fluctuates around 0.05 between epochs≈160 and epochs≈240. Based on Figure 5, I concluded that the epochs between 160 and 240 would generate the best outputs. In Figure 6, a generated output was recorded for every 20 epochs. There are pixel scratches on the bottom of the output until 140 epochs,

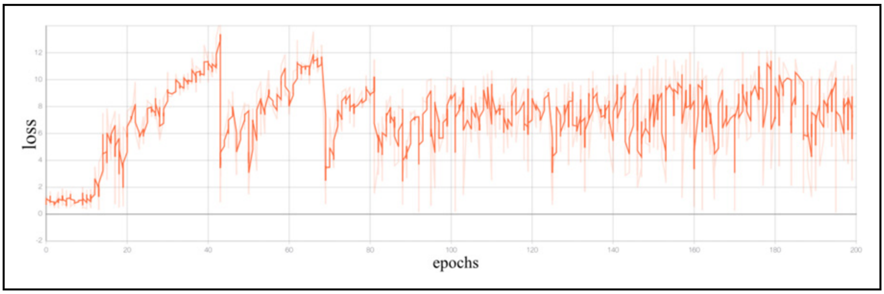

Figure 3. The generator loss of pix2pix with the *Victorian400* dataset.

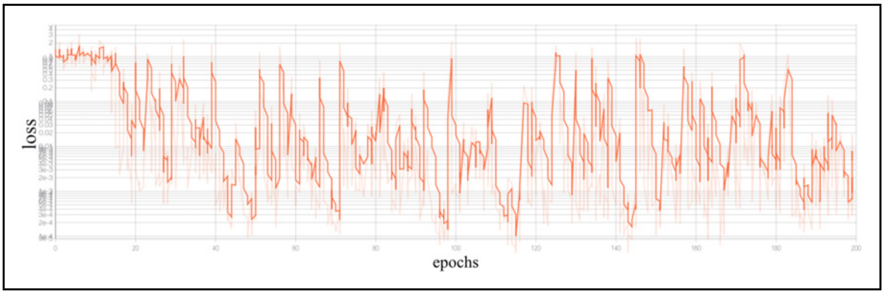

Figure 4. The discriminator loss of pix2pix with the *Victorian400* dataset.

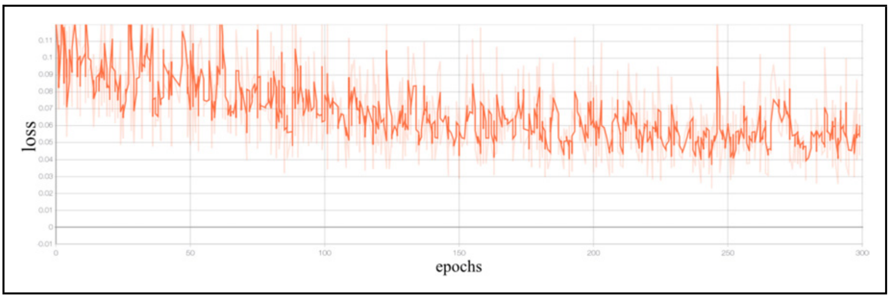

Figure 5. The L1 loss of pix2pix with the *Victorian400* dataset.

but from 160 epochs, the pix2pix model generates reasonable results without many improvements. Similarly, there are no remarkable improvements to the L1 loss in Figure 5 from around 160 epochs. Based on both the quantitative and qualitative methods (Figs 5 and 6), I decided to train the pix2pix model with the *Victorian400* dataset for 200 epochs to predict colours for black-and-white illustrations from the Victorian era.

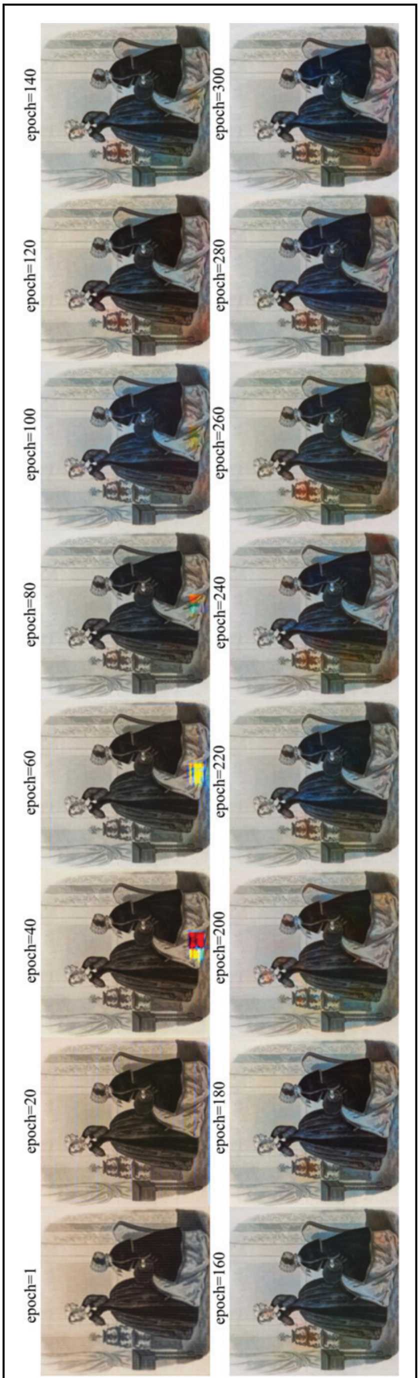

Figure 6. Test results of pix2pix with the *Victorian400* dataset based on every 20 epochs.

The *Victorian400* dataset was run with a test dataset that consisted of nine images. The L1 loss is around 0.044 at 200 epochs as a result of training the pix2pix model with the *Victorian400* dataset. Figure 7 shows the results of the test set after training with the *Victorian400* dataset. Despite the fact that the test results from Test1 to Test9 are not exactly the same as the original targets, the results provide natural-looking coloured images based on the trained set. Looking into Test1, Test5, Test6 and Test9 from the column of Figure 7, however, the objects have a red tint. As mentioned above, this is the result of the fact that red is often found in Victorian illustrations. For example, Figure 8 shows Victorian women wearing colourful, reddish clothes. While testing the pix2pix model with the *Victorian400* dataset, I found the illustrations were coloured with red, which was a case of overfitting due to the dataset being biased towards the colour red. I had to remove some images which had a preponderance of red in them in order to balance the colour palette. At the same time, I left a certain number of images using red so as not to disregard the colour preferences of the Victorian age, as I deemed this important to reflect contemporaneous colour trends. As the dataset significantly influences the test results of deep learning models, these Victorian colour trends were still revealed through Test1 and Test9. In addition, limitations remain due to the small size of the *Victorian400* dataset.

After verifying the *Victorian400* dataset with the test dataset, I used the trained model with the *Victorian400* dataset in order to predict the colours of black-and-white illustrations from Charles Dickens's *Bleak House* (serialized 1852–1853). The illustrations used for the test are from the Bradbury & Evans book edition (1853) of Dickens's *Bleak House*, for which the plates were created by Hablot Knight Browne (Phiz). Browne illustrated ten of Dickens's books including *Dombey and Son* (serialized 1846–1848), *David Copperfield* (serialized 1849–1850), and *Bleak House*. *Bleak House*, which includes 40 black-and-white illustrations with the title-page etched on steel, is a good subject for testing colorization due to the fact that there is a good variety of illustrated depictions of characters and settings. In order to evaluate the outputs generated by the pix2pix model with the *Victorian400* dataset, I divided them into three categories: not good, fair, and good. As for validating GAN models, 'visual examination of samples by human raters is one of the common and most intuitive ways'.[36]

In Figure 9, there are 12 illustrations generated by pix2pix trained with the *Victorian400* dataset as examples. Figure 9 can be divided into three categories based on the quality of the output:

1. Not Good: Test1, Test2
2. Fair: Test3, Test4, Test8
3. Good: Test5, Test6, Test7, Test9, Test10, Test11, Test12

Figure 7. The outputs of the pix2pix model with the *Victorian400* dataset.

Figure 8. Two images from the *Victorian400* dataset.

Test1 through Test6 are focused on characters, while Test7 through Test12 are focused on settings. Test1 reveals that the background and people in the background were not coloured. There are a number of blurry faces in the background, and walls were not sketched with enough detail for the deep learning model to detect them. Similarly, Test2 was not coloured thoroughly, with only two characters on the left painted in yellow and orange. In Test3 and Test4, characters were coloured based on the trained set, but the backgrounds were not coloured enough to show improvements on the black-and-white illustrations. Test3, Test4 and Test8 were categorized as fair since there were reasonable improvements in assigning the colours between the inputs and outputs. Test5 and Test6 were sufficiently coloured in terms of both the characters and the background, although the colours were not always appropriate. Test7 to Test12 are coloured based on black-and-white background illustrations, which Browne created by using the dark plate technique[37] in order to provide deep contrast for a higher quality of depth. The illustrations from Test7 to Test12 created with the dark plate technique turned out to be compatible with deep learning colorization by generating reasonable outputs due to the distinct contrast between darkness and highlights.

As proved by this experimentation, there is no single answer to the question of how to colorize black-and-white illustrations. Colorizing black-and-white illustrations is a double-edged sword since, while it may distort the original illustration, it nonetheless provides the opportunity for modern readers to expand their imagination and increase reading pleasure. To increase the accuracy of colorizing backgrounds, I would need a vast amount of coloured background

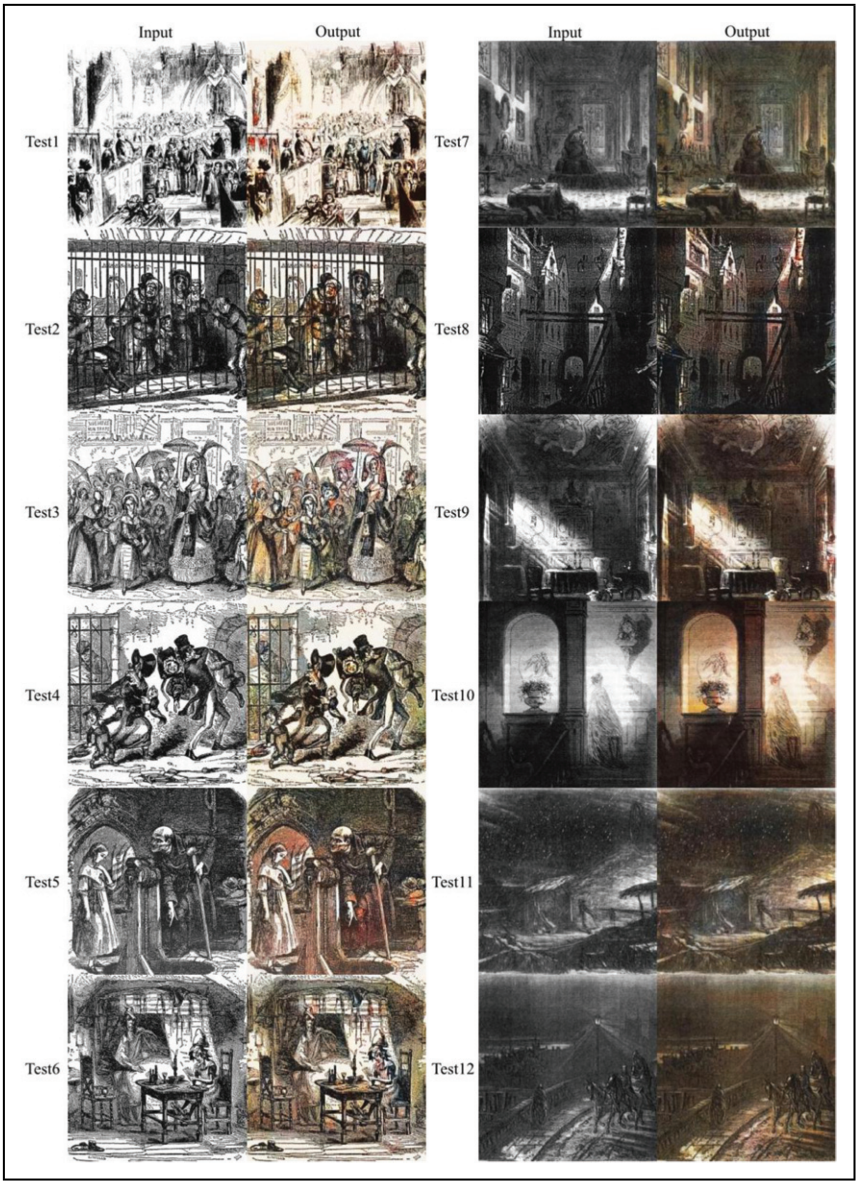

Figure 9. Test results for black-and-white illustrations from Charles Dickens's Bleak House.

images. In addition, it would require advanced deep learning layers dealing with a variety of different images. Without the support of grants, it would be impossible for an individual project like this one to create a large dataset and to develop more complicated and enhanced deep learning models. The typical cGAN training requires a large training set in order to score reasonable results. As an individual project, the *Victorian400* dataset is a compromise between time and cost. It is true that the *Victorian400* dataset is small compared to other datasets for GAN training. Nonetheless, it produces reasonable enough results from the test sets to attest to its quality.

3. CONCLUSION

The *Victorian400* dataset was created for data scientists and digital humanists who create, train and test colorization deep learning models. The *Victorian400* dataset has been publicly shared on websites, namely Kaggle and GitHub, for data scientists and digital humanists, and it has been downloaded and experimented with using a variety of deep learning models.[38] Undeniably, due to its size, there is still scope for improvements in the *Victorian400* dataset. Although some scholars might think the *Victorian400* dataset is small, it proved sufficient to produce reasonable results. In addition, small, high-quality datasets make it possible to run tests easily without spending too much time and money.[39] Data scientists spend a great portion of time curating datasets in order to make them valid and credible, in addition to spending funds to run deep learning models in cloud services such as Amazon EC2, Microsoft Azure and Google Colab. The *Victorian400* dataset will not only save a tremendous amount of time for digital humanists who experiment with Victorian illustrations, but will also contribute to the development of deep learning–based research in the digital humanities. Lastly, the *Victorian400* dataset will assist digital humanists in the creation of datasets for machine learning/data scientists; it is difficult to find credible humanities datasets on the Internet since most humanities datasets shared on Kaggle and GitHub are created by non-humanists.[40] It is the role of digital humanists not only to create and curate humanities datasets, but also to provide EDA results which show that humanities datasets created by digital humanists are credible enough to be used for deep learning tasks. This will enrich the scholarship in both digital humanities and data science as well as lead to the development of collaboration between digital humanists and data scientists.

END NOTES

[1] T. Clement and A. Acker, *Data cultures, culture as data*, special issue of *Cultural Analytics* (2019), https://culturalanalytics.org/2019/04/data-cultures-culture-as-data-special-issue-of-cultural-analytics/.

[2] M. Stonebraker et al., 'Data curation at scale: the data tamer system', *Innovative Data Systems Research* (2013), 1, https://cs.uwaterloo.ca/~ilyas/papers/StonebrakerCIDR2013.pdf.

[3] Christof Schöch has argued that data in the humanities are unique artifacts. Trevor Muñoz has articulated relationships between publishing and data curation. Miriam Posner has claimed that we should have high standards when creating 'data-based work that depicts people's lives'. Katie Rawson and Trevor Muñoz revealed how to properly clean data for humanities researchers. C. Schöch, 'Big? Smart? Clean? Messy? Data in the humanities', *Journal of Digital Humanities*, 2, no. 3 (2013), http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/; T. Muñoz, 'Data curation as publishing for the digital humanities', *Journal of Digital humanities*, 2, no. 3 (2013), http://journalofdigitalhumanities.org/2-3/data-curation-as-publishing-for-the-digital-humanities/; M. Posner, 'What's next? The radical unrealised potential of digital humanities', in M. Gold and L. Klein, eds, *Debates in the digital humanities 2016* (Minneapolis, 2016), 32–41; K. Rawson and T. Muñoz, 'Against cleaning', in M. Gold and L. Klein, eds, *Debates in the Digital Humanities 2019* (Minneapolis, 2019), 279–92.

[4] Data scientists usually perform EDA (Exploratory Data Analysis) to find how datasets can be used for deep learning and to see whether they are credible. One way to perform EDA is to use t-SNE (t-distributed Stochastic Neighbor Embedding) with random sampling to see clustering shapes and the distance between features. In general, dataset creators provide test results with specific information such as parameters, evaluation metrics and network architectures, in order to show credibility of the datasets, as shown in this article.

[5] To analyse humanities datasets is both to provide EDA and to further expand data analysis into digital projects and articles.

[6] I. Goodfellow et al., 'Generative adversarial nets', *Advances in Neural Information Processing Systems* (2014), 1, http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[7] Goodfellow et al., 'Generative adversarial nets', 1.

[8] P. Isola et al., 'Image-to-image translation with conditional adversarial networks', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017). https://openaccess.thecvf.com/content_cvpr_2017/papers/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.pdf.

[9] S. Reed et al., 'Generative adversarial text to image synthesis', *International Conference on Machine Learning* (2016). arXiv:1605.05396.

[10] C. Ledig et al., 'Photo-realistic single image super-resolution using a generative adversarial network', *Computer Vision and Pattern Recognition* (2017). https://openaccess.thecvf.com/content_cvpr_2017/papers/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.pdf.

[11] Isola et al., 'Image-to-image translation with conditional adversarial networks'.

[12] J. Antic, *DeOldify* (2019). GitHub repository, https://github.com/jantic/DeOldify.

[13] Y. Taizan, *Paints Chainer* (2017). GitHub repository, https://github.com/pfnet/PaintsChainer.

[14] Eight of Charles Dickens's photographs were colorized by Oliver Clyde and released by the Charles Dickens Museum in London ahead of the 150th anniversary of Dickens's death.

[15] There were hand-coloured additions of luxurious books until 'the last decades of the eighteenth century', but the emergence of printing methods such as chromolithography began to bring colour illustrations to the public from the 1830s. Colour printing was essential for 'children's books, religious works, and gift books'. Still, colour printing was costly during the Victorian era. See D. Allington et al., *The book in Britain: a historical introduction* (Hoboken, NJ, and Oxford, 2019), 293.

[16] The *Victorian400* dataset was used to introduce deep learning and Python tutorials on Pseudo Lab (https://pseudo-lab.github.io/Tutorial-Book-en/index.html) and Kaggle (https://www.kaggle.com/jiny333/tutorial-on-using-subplots-in-matplotlib), respectively.

[17] Several data scientists have contacted me through Kaggle to express their appreciation that I shared the dataset with the public.

[18] For two semesters, I used both hand-coloured and machine-coloured nineteenth-century illustrations in my English classes to see how different colouring influences readers and the possible usage of machine-colourized illustrations in facilitating students' learning and imagination in literature. Most of the students responded to the machine-colourized illustrations with surprise, curiosity and excitement. One made the observation that machine-coloured illustrations helped depict objects in a way that is difficult to achieve in hand-coloured illustrations. Another noted that the hand-coloured illustrations seemed too modern, whereas the machine-colored illustrations seemed more realistically Victorian.

[19] This article does not deal with the pedagogical side of deep learning colorization but focuses on the validation of the *Victorian400* dataset.

[20] https://www.kaggle.com/elibooklover/victorian400.

[21] https://github.com/elibooklover/Victorian400.

[22] See M. Hardie, *English coloured books* (London, 1906), which introduces English colour illustrations by a variety of artists including W. Savage, T. Rowlandson, T. S. Boys, David Roberts and Kate Greenaway. See P. Allingham, 'The technologies of nineteenth-century illustration: woodblock engraving, steel engraving, and other processes' at *The Victorian web* (http://www.victorianweb.org/art/illustration/tech1.html), which examines illustration printing techniques in the Victorian era. See V. Finlay, *Color: A natural history of the palette* (New York, 2004), which explores the histories of colours.

[23] There are a number of Victorian databases/archives/platforms with Victorian illustrations: *Database of mid-Victorian illustration* by Julia Thomas et al. (http://www.dmvi.cardiff.ac.uk), *The illustration archive* by Julia Thomas et al. (https://illustrationarchive.cardiff.ac.uk/), *The Victorian web* by George Landow (http://www.victorianweb.org/), *The Rossetti archive* by Jerome McGann et al. (http://www.rossettiarchive.org/), *The Charles Dickens page* by David Perdue (https://www.charlesdickenspage.com/), *The George Eliot archive* by Beverley Rilett (http://www.georgeeliotarchive.org/), and *Collaborative organization for virtual education* by Dino Felluga et al. (https://editions.covecollective.org/). These sites are historically and pedagogically helpful with open access for researchers, instructors and students.

[24] A. Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems* (Sebastopol, CA, 2019).

[25] Since this article is not about the pix2pix model, which is a deep learning model for image-to-image translation, I focused on the process and results of training the *Victorian400* dataset with the pix2pix model. For more details on the pix2pix model, see P. Isola et al., 'image-to-image translation with conditional adversarial networks', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).

[26] Goodfellow et al., 'Generative adversarial nets'.

[27] Isola et al., 'Image-to-image translation with conditional adversarial networks', 3.

[28] D. Ciresan et al., 'Deep neural networks segment neuronal membranes in electron microscopy images', *Neural Information Processing Systems* (2012), http://papers.nips.cc/paper/4741-deep-neural-networks-segment-neuronal-membranes-in-electron-microscopy-images.pdf.

[29] O. Ronneberger, P. Fischer and T. Brox, 'U-net: convolutional networks for biomedical image segmentation', *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2015), 2, arXiv:1505.04597.

[30] Ronneberger, Fischer and Brox, 'U-net: convolutional networks for biomedical image segmentation', 3.

[31] Goodfellow et al., 'Generative adversarial nets', 1.

[32] Isola et al., 'Image-to-image translation with conditional adversarial networks', 4.

[33] M. Lucic et al., 'Are GANs created equal? A large-scale study', *NIPS* (2018), arXiv:1711.10337; K. Shmelkov, C. Schmid and K. Alahari, 'How good is my GAN?' *Proceedings of the European Conference on Computer Vision* (2018), https://openaccess.thecvf.com/content_ECCV_2018/papers/Konstantin_Shmelkov_How_good_is_ECCV_2018_paper.pdf; Q. Xu et al., 'An empirical study on evaluation metrics of generative adversarial networks', (2018), arXiv:1806.07755; S. Arora and Y. Zhang, 'Do GANs actually learn the distribution? An empirical study' (2017), arXiv:1706.08224; K. Kurach et al., 'The GAN landscape: Losses, architectures, regularization, and normalization', (2018), https://arxiv.org/pdf/1807.04720v1.pdf.

[34] A. Borji, 'Pros and cons of GAN evaluation measures', *Computer Vision and Image Understanding*, 179, (2019), 2, doi.org/10.1016/j.cviu.2018.10.009.

[35] There are several issues with GANs such as non-convergence, mode collapse and unstable gradients. This article does not deal with the limits of GANs. For discussions of the limits of GANs, see Lucic et al., 'Are GANs created equal?'; M. Arjovsky and L. Bottou, 'Towards principled methods for training generative adversarial networks', *International Conference on Learning Representations* (2017), and K. Li and J. Malik, 'Implicit maximum likelihood estimation' (2018), arXiv:1809.09087.

[36] Borji, 'Pros and cons of GAN evaluation measures', 30.

[37] Browne deployed the dark plate technique, which is a combination of engraving and etching, in order to convey thematic schemes in Dickens's works such as crime and hopelessness. This article does not delve into the interpretations of the illustrations in Dickens's novels but focuses on the validity of the *Victorian400* dataset.

[38] The DH community has been participating in an open data movement. Amy Earhart and Toniesha Taylor assert that allowing 'the public to view the type of work that we accomplish is powerful, particularly within the current environment of distrust of the academy'. From the perspective of a dataset creator, sharing work with the public yields feedback from a wider range of viewers that makes it possible to improve my work. The *Victorian400* dataset is currently ranked with a bronze medal on Kaggle. A. Earhart and T. Taylor, 'Pedagogies of race: Digital humanities in the age of Ferguson', in M. Gold and L. Klein, eds, *Debates in the digital humanities 2016* (Minneapolis, 2016), 259.

[39] It took me around 13 hours to train the *Victorian400* set once with 200 epochs. There were difficulties when testing with the *Victorian400* dataset due to a poor research environment. In order to trim the data and avoid overfitting, I had to train with a variety of new datasets, which took a tremendous amount of time. Later, a project grant was given to this project by CoDHR (Center of Digital Humanities Research), so I was able to test the dataset in a better environment.

[40] When choosing datasets, some users consider who has created the datasets and how popular the datasets are, whereas others draw on EDA for the datasets, as these provide objective criteria, although it can be difficult to determine the credibility of some datasets until they are used for specific tasks. It is thus important for digital humanists who create and share datasets to provide EDA and thus enhance the credibility of their shared datasets.