Coloring Manga Images with Deep Learning: A Comparative Study

Gurman Bhullar
Department of Computer Science
University of Toronto
gbhullar@cs.toronto.edu

Naveen Thangavelu
Department of Computer Science
University of Toronto
naveentvelu@cs.toronto.edu

Rajesh Marudhachalam
Department of Computer Science
University of Toronto
rajesh1804@cs.toronto.edu

1 Introduction

Automatic colorization is a research area with great potential in applications such as B&W photos reconstruction, re-coloring images, grey scale drawing augmentation, etc., We investigated a sub-category of colorization: automatic colorizations in Manga (Japanese comics). Even today, many of the manga are published in black and white. The main reason being, it's hard to color the manga given the tight publishing deadlines and production costs. However, with recent advances in generative models, such as Dall-E, we can effectively leverage them to automatically color the manga. This can be beneficial in providing the user with more pleasant and comprehensive reading experience, and potentially resulting in a larger fan base. Colorization is a popular image-to-image translation problem, which can be solved by the generative models such as Conditional GANs and Diffusion models [8].

We use Generative Adversarial Networks (GANs) and Diffusion models. GANs can be trained on large datasets of images to generate high-quality and realistic results. GANs [4] have been widely used to generate high-quality images by training on large datasets of images. However, recent advances in diffusion models [3]have shown that they can generate even higher quality images and outperform GANs in image synthesis. This study explores the potential of these models for manga colorisation.

2 Related Work

Several studies have explored the use of deep learning models to color images, with a particular emphasis on Manga. These models have included Generative Adversarial Networks (GANs) [6], Pix2Pix [9], and neural style transfer [5].

CycleGANs are a type of GAN (Generative Adversarial Network) that can learn to translate images from one domain to another without paired training data [10]. Unlike traditional GANs, which require paired training data (i.e., images from both domains), CycleGANs can learn to map images from one domain to another in an unsupervised manner. The key idea behind CycleGANs is to use a cycle-consistency loss to ensure that the generated images are consistent with the input images. The cycle-consistency loss encourages the generator to learn a mapping from domain A to domain B and back to domain A again, such that the generated image is similar to the original image.

Pix2pix [9] refers to a type of GAN that facilitates the translation of images to other images, given that there is a paired training dataset. This model can acquire the ability to produce output images from input images that conform to a particular target distribution. Besides learning how to map input images to output images, these networks can also learn how to use a loss function to train the mapping. As a result, this method can be utilized to solve problems that would typically require different loss formulations. This technique is useful for various tasks, such as synthesizing pictures from label maps, colorizing images, and reconstructing objects from edge maps, among others.

While many text-to-image difussion models possess an inpainting functionality, effectively producing the desired output based on a textual description can prove to be challenging in practice. This is because users are either required to provide a complete description of the output image, or manually generate a mask to pinpoint which area(s) of the image should be modified, and then detail the intended modifications. However, InstructPix2Pix [2], which is a modification of Stable Diffusion - a conditional diffusion model that edits images from written instructions - introduces a novel approach to image modification using text instructions that are easy to comprehend. Rather than describing the desired output image or providing a mask, users can now simply provide intuitive instructions on how the input image should be modified, freeing them from the task of having to create a mask, and enabling mask-free editing.

Wang et al. [12] suggest when data for paired training is limited, fine-tuning a large image diffusion model is a more effective approach for image translation tasks than training a model from scratch. As such, the instructpix2pix training procedure adopts the weights of a pre-trained Stable Diffusion checkpoint to benefit from its extensive text-to-image generation capabilities. In this process, all available weights of the diffusion model are initialized from the pre-trained checkpoints, while weights associated with the new input channels are initialized to zero. To achieve this, Brooks et al. [2] re-purposed the text conditioning mechanism, originally designed for captions, to instead take the text edit instruction as input.

3 Experiments

3.1 Dataset

Publicly available Kaggle Animes image dataset [11] consisting of 58000 RBG portrait images of anime characters is used, out of which 2500 randomly sampled images were used for the training of the deep learning based and GANs models, and 106 randomly sampled images were used for the purpose of model evaluation. All these images were converted to grayscale before being used for training and evaluation.

3.2 Models

In this work we experiemnt with, Cycle GAN [15], Pix2Pix GAN [9] and InstructPix2Pix diffusion [2] models.

3.3 Methodology

The methodology used for the CycleGAN involved training the model from scratch and using the open-source implementation available on GitHub to obtain results. For training, a total of 2500 images were provided to the model, and the inference was performed on 106 images. The CycleGAN was trained for 10 epochs. The network used for image generation contains three convolutions, several residual blocks, two fractionally-strided convolutions, and one convolution for mapping features to RGB. Instance normalization is used, and a PatchGAN discriminator is employed to classify whether 70x70 overlapping image patches are real or fake. The loss function used is $Loss_{GANs}$ [15], which replaces the negative log likelihood with least square loss. The Adam solver is used with a batch size of 1, a learning rate of 0.0002, and a learning decay rate of 50.

The Pix2Pix GAN was also trained from scratch using the same set of data used for the CycleGAN described for a total of 20 epochs. The model was trained using a U-NET256 generator (256 convolutional filters in the first layer), PatchGAN discriminator (70 * 70 patch size), and vanilla GAN loss (the cross-entropy objective used in the original GAN paper [7]).

The instructpix2pix pretrained API [1] from the Huggingface was used in our study, and 50 output images were generated for each input image, with every 10 images produced using a different set of hyperparameters controlling the similarity with the input image and consistency with the edit instruction (Refer to Appendix A). As anticipated, the model was able to generate aesthetically pleasing colorized images for every input image, with at least one set of hyperparameters yielding satisfactory results. However, the lack of fine-tuning on the anime dataset means that these generated images may not be fully representative of the ground truth, for instance a certain set of hyperparameters produced outputs with significant structural differences than those in the input image. Interestingly, it was observed that when the input grayscale image contained a higher number of darker pixels, such as due to dark backgrounds or dark attire, the diffusion model struggled to produce properly colored images and generated outputs with an overall brownish tint, regardless of the hyperparameter set used.

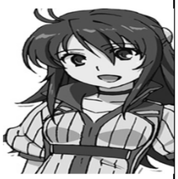
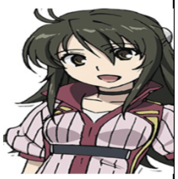
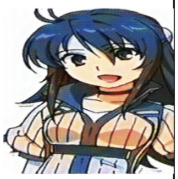
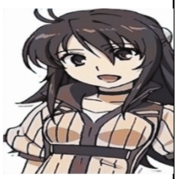
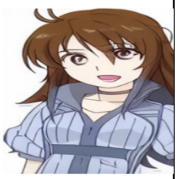
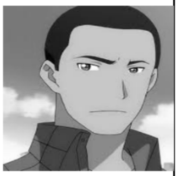
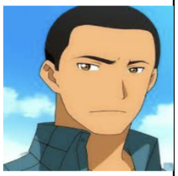
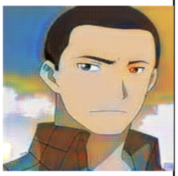
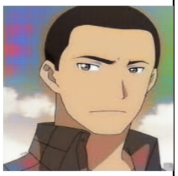
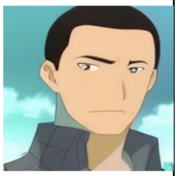
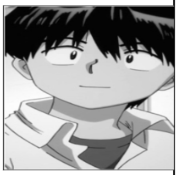
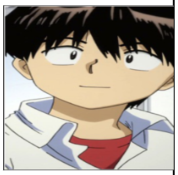
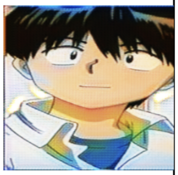
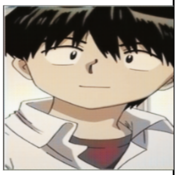
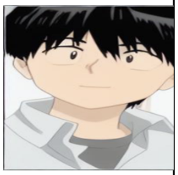
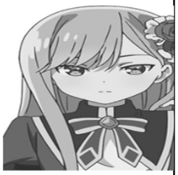
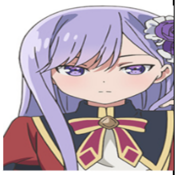
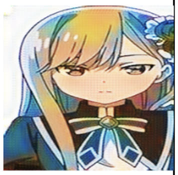
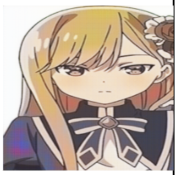
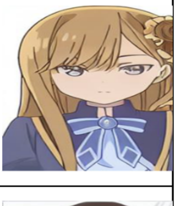
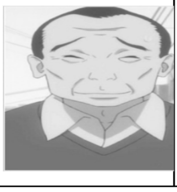
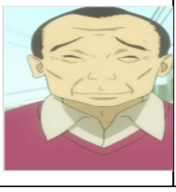
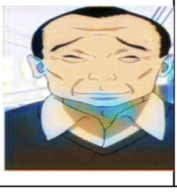
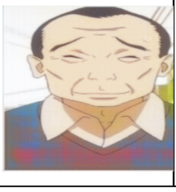
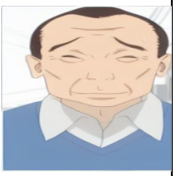
S. No	GrayScale Image	Ground truth Image	CycleGANs Output	Pix2Pix Output	InstructPix2 Pix Output
Image 1					
Image 2					
Image 3					
Image 4					
Image 5					

Table 1: Colored Images Generated by Various Models

4 Results

Table 1 shows the images that exhibits satisfactory colouring of images. The choice of images depend on the different backgrounds, face zooms, age factor among anime characters and the level of complexity for image colouring.

The metrics parameters used for the purpose of evaluation are Fréchet Inception Distance (FID) score [14] and Structural Similarity Index (SSIM) [13]. Both FID and SSIM scores are used as objective measures of image quality, the scores obtained by various various models are described in Table 2.

FID score measures the distance between the feature representations of real and generated images. It calculates the distance between the mean and covariance of the feature representations, using a pre-trained Inception-v3 model. The lower the FID score, the closer the generated images are to the real images in terms of their features.

$$FID = \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1\Sigma_2\right)^{1/2}\right) \tag{1}$$

where μ_1 and μ_2 are the mean vectors of the feature representations of real and generated images respectively, and Σ_1 and Σ_2 are the covariance matrices of the feature representations of real and generated images respectively.

SSIM score, on the other hand, measures the structural similarity between the real and generated images. It calculates the similarity based on luminance, contrast, and structure. The closer the SSIM score is to 1, the higher the structural similarity between the two images.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{2}$$

where x and y are the original and generated images, μ_x and μ_y are the mean intensities of x and y respectively, σ_x and σ_y are the standard deviations of x and y respectively, σ_{xy} is the covariance between x and y, and c_1 and c_2 are constants to avoid division by zero.

Model	FID Score	SSIM Score	Fool Percentage
CycleGANs	92.236	0.8512	14%
Pix2Pix	61.205	0.9553	42%
InstructPix2Pix	72.568	0.7945	50%

Table 2: Evaluation Metrics

In addition to utilizing mathematical evaluation metrics, a human evaluation was also carried out on a set of 40 images. This set consisted of 10 randomly selected images from the test set of ground truth and images produced by the three models under consideration. Each image was presented individually to human evaluators who were asked to label it as "real" or "fake", based on whether the image resembled a real or fake anime character with coloring. The results of this evaluation have been tabulated in Table 2, as fool percentage defined by the ratio of the number of fake images classified as real to the total number of fake images, averaged across all human evaluators. We ensured the credibility of the evaluation by verifying that all human evaluators were able to correctly classify over 65% of the images as either real or fake.

5 Conclusion

In conclusion, the evaluation of fool percentage values reveals that the InstructPix2Pix diffusion model is the top-performing model for the task of Manga image colouring. Notably, while the InstructPix2Pix model obtained the best SSIM score, the Pix2Pix GAN model achieved the best FID score. However, human evaluation results were given precedence as there is no definitive colouring scheme for Manga images. It is worth mentioning that the prompts used for the diffusion model may not be generalizable for other use cases, such as generalized image colouring, hence, the obtained results are specific to the Manga image colouring task. In general, the diffusion model outperformed the GAN model for this specific task, perhaps due to its greater robustness to overfitting, lower sensitivity to the quality and quantity of training data, and higher capacity to generate high-quality images. Nevertheless, the selection of the best model should be based on an evaluation of multiple models and the requirements of the particular task.

6 Future Work

In terms of future work, our focus will be on improving the generalizability of the InstructPix2Pix model beyond Manga image colouring. This may entail exploring alternative prompts and training data to make the model adaptable to a wider range of image types and styles. Furthermore, we aim to investigate the potential of transfer learning techniques to enhance the performance of the InstructPix2Pix model. One possibility is to pre-train the model on a vast dataset of images such as ImageNet, and then fine-tune it on the Manga dataset to enhance the model's ability to learn and generalize to new images.

Another potential area for future work is the exploration of other generative models, such as Variational Autoencoders (VAEs) and Generative Flow Models, for the task of Manga image colouring. Comparing the performance of these models with the current ones used in this study could provide additional insights into the strengths and limitations of different generative models for this task.

References

[1] T. Brooks. Instructpix2pix: Learning to follow image editing instructions, 2023.

[2] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[3] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[4] R. Dhir, M. Ashok, S. Gite, et al. An overview of advances in image colorization using computer vision and deep learning techniques. *Rev. Comput. Eng. Res*, 7(2):86–95, 2020.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[8] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[9] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[10] B. Li, Y. Lu, W. Pang, and H. Xu. Image colorization using cyclegan with semantic and spatial rationality. *Multimedia Tools and Applications*, pages 1–15, 2023.

[11] Shanmukh. Anime names and images dataset, 2020.

[12] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[14] Y. Yu, W. Zhang, and Y. Deng. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 2021.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Appendices

A Hyperparameters & Prompts in InstructPix2Pix

A.1 Hyperparameters

In our experimentation, we varied several hyperparameters to determine their impact on the performance of our model. The following is a list of the hyperparameters and their respective values that were explored:

1. `image_guidance_scale` : 1, 1.25 (**BEST**), 1.5 (**BEST**), 1.75, and 2
2. `text_guidance_scale` : default (**BEST**), 7, 10
3. `inference_steps` : The number of inference steps was set to 20 during the experimentation phase.

Through the analysis of these hyperparameters, we were able to identify the optimal values for our model, which led to improved performance.

A.2 Prompts

During our experimentation, we tested various prompts to determine their effectiveness in accurately coloring images. It should be noted that the following prompts were used in combination with the BEST hyperparameters to generate the most favorable outcomes:

Prompt	**Observation**
Colorize this person to have natural skintones	This prompt produced results that were realistic and not overly animated, though it often altered the image structure.
turn it colorful	This prompt produced wildly random outputs and often failed to properly color skin tones.
make it a colorful professional headshot realistically animated	This prompt produced the best results among all of the prompts tested. It effectively maintained the original image structure while producing believable and accurately colored images.

Table 3: Image edit text prompts

Through the evaluation of these prompts in conjunction with our hyperparameter tuning, we were able to identify the optimal prompt to achieve our desired outcomes.

B Reproducing the results

The source code utilized to produce the results can be accessed in the master branch of the GitHub repository, which is accessible to the public at `https://github.com/rajesh1804/csc2516-coloring-manga-images`.

C Contributions

Authors are listed in alphabetical order. All members contributed equally throughout the duration of the project. All decisions related to the different parts of the project (like problem formulation, literature review, experimental design, and inferences) were taken collectively after discussion. Some specific contributions are as follows: Gurman took the initiative to implement the Cycle-GAN experiments and was also responsible for compiling the results. Naveen led the experimentation of Pix2Pix GAN, and was also responsible for setting up the automatic evaluation metrics and designing the human evaluation framework. Rajesh took the end-to-end initiative of experimenting, generating results and metric scores with the Instruct Pix2Pix Diffusion Model.