Towards a Novel Multi-Loss CycleGAN for Full-to-Flat Colour Transformation

Simba Shi

SketchX, Centre for Vision, Speech and Signal Processing, University of Surrey

ABSTRACT

Full colour to flat colour transformation tools are essential in digital art and design for simplifying detailed, gradient-rich images, useful in creating vector illustrations, comics, animations, and general graphic design that emphasise clarity for various stylistic and practical applications. However, such existing transformation tools are 1) not autonomous, and require much human operation, and 2) often inaccurate for images of different styles. With the prevalence of generative AI, in this paper, we propose a novel Generative Adversarial Network (GAN) AI pipeline for transforming full colour into flat colour, trained on a hand-picked unpaired dataset of full colour and flat colour images. Specifically, we employ a CycleGAN architecture with additional loss functions; aside from the standard GAN adversarial loss, discriminant loss and cycle consistency, we employ depth geometry loss with InceptionV3 network, sketch extractor loss with a pretrained line extractor model, and semantics loss with OpenAI's CLIP model. The resulting full-to-flat colour AI model requires no human input, and preserves inherent image semantics, depth, shape and textual information. The full code is available at: https://github.com/Coderrexe/full2flat

INTRODUCTION

Flat colouring is a crucial technique in digital art and design, transforming full-color images into simplified, stylised forms. This process enhances visual content by distilling complex scenes into clean, vector illustrations, aiding emotional conveyance without the distraction of intricate gradients. Practically, flat coloring allows for efficient processing and rendering, crucial for digital mediums where file size and loading times are important.

Despite its benefits, current flat colouring tools require extensive human intervention and lack accuracy across various image styles. Our deep learning approach introduces a novel **Generative Adversarial Network (GAN)** AI pipeline. Our dataset is **unpaired**, eliminating the costs and time associated with human labelling.

- The full colour dataset contains 2246 anime portrait images, hand picked from the Danbooru dataset containing 5 million total images.
- The flat colour dataset contains 2604 hand-picked cartoon images from 3 separate datasets on Kaggle; Cartoon Classification, Family Guy Dataset, and FiveThirtyEight Comic Characters Dataset.

Furthermore, we incorporate additional models and loss functions to further improve the performance. These include, in order of importance:

- **GAN sketch extractor** taken from *Chan, C., Durand, F., & Isola, P. (2022). Learning to generate line drawings that convey geometry and semantics* for preserving image outline.
- **Custom InceptionV3-based MiDaS** image depth prediction network designed to preserve geometric and depth information between the original and generated images.
- **Contrastive Language-Image Pretraining (CLIP)** model with ViT-B/32 architecture (Vision Transformer), for preserving semantic information during the image transformation process.

RELATED WORK

No current AI tool transforms full-color images into flat-color ones. However, line drawing generation from 3D geometry is well-studied, using geometric features, depth, and normal maps, or deep learning combined with geometry-based techniques. Most 2D-based techniques require supervised data, vector graphics, or ground truth strokes, and focus on conditional generation from paired photos. In contrast, our approach translates between different sketch domains and handles unpaired data, offering greater versatility for various applications.

METHODOLOGY

Figure 1: CycleGAN Generator Architecture

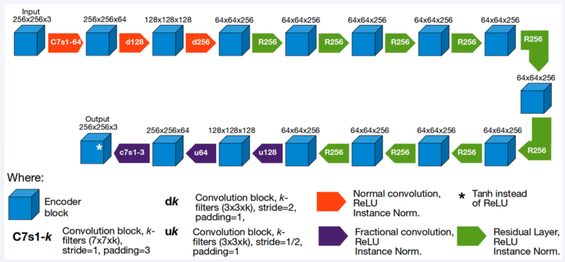


Figure 2: InceptionV3 Image Depth Prediction Network

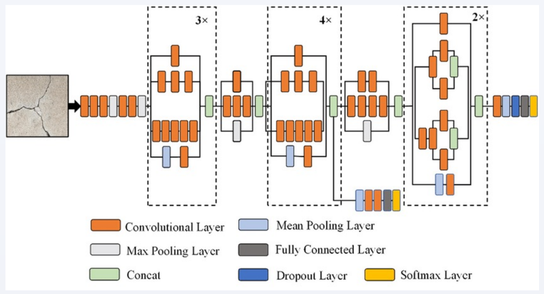


Figure 3: ViT-B/32 CLIP Model Architecture

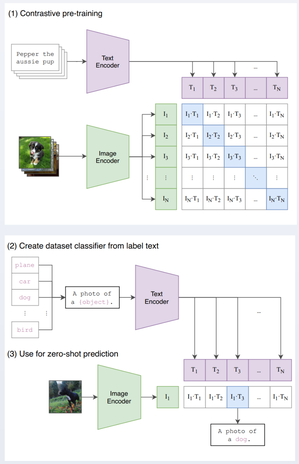


Figure 4: Full Model Architecture

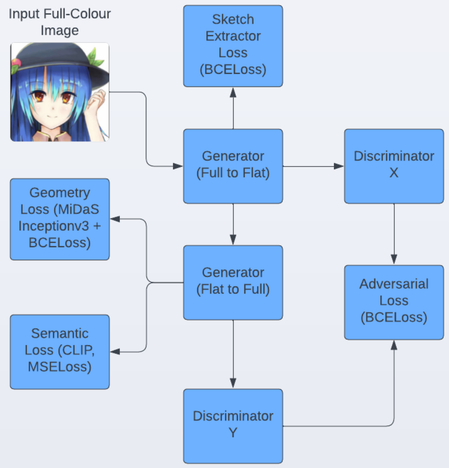
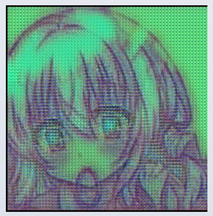


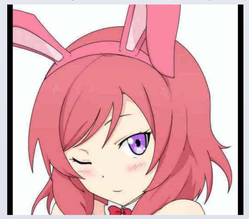
Figure 5: Unpaired Dataset Examples, Full Colour (Above), Flat Colour (Below)

RESULTS

Flat-colour generation (5/100 epochs)



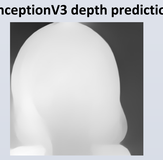
Flat-colour generation (100/100 epochs)



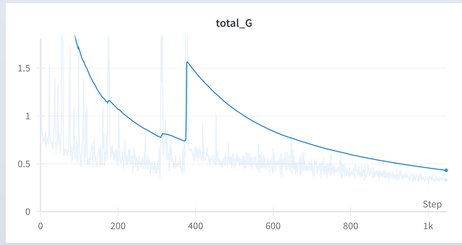
Original full-colour image

Sketch extractor model output

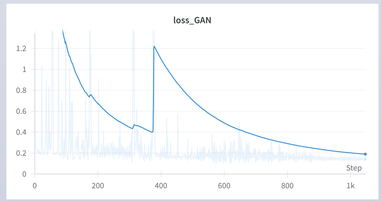
MiDaS InceptionV3 depth prediction output



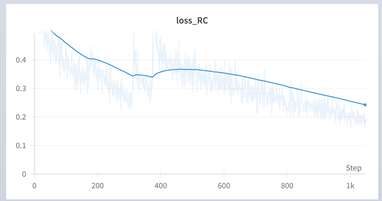
Total loss (weighted sum of all individual losses)



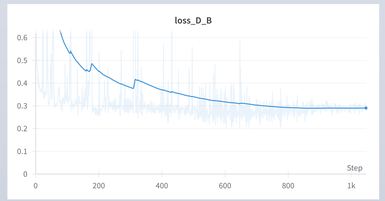
GAN generator adversarial loss



GAN cycle consistency loss



GAN discriminator loss



MODEL TRAINING

Our GAN model was trained on a single **NVIDIA GeForce GTX TITAN X**, for **100 epochs** (including 50 with learning rate decay) and **batch size 4**, utilising the **Adam optimiser** with momentum parameters (beta1=0.5, beta2=0.999). Each full model training took around 17 hours. The core GAN components, coded with PyTorch, include:

- **Generators (G and F):** Responsible for transforming images between full-colour and flat-colour domains.
- **Discriminators (D_X and D_Y):** Distinguish between real and generated images to enforce realism and help train the generators.
- **Adversarial Loss:** Uses the Binary Cross-Entropy Loss (BCELoss) to ensure generated images are indistinguishable from real images.
- **Cycle Consistency Loss:** Utilises L1 Loss to ensure that an image transformed to the target domain and back remains unchanged.
- **Sketch Extractor Loss:** Applies BCELoss with a pre-trained line extractor model to maintain image outlines and details.
- **Depth Loss:** BCELoss via a custom InceptionV3-based MiDaS depth prediction network to preserve geometric and depth information.
- **Semantic Loss:** Uses Mean Squared Error Loss (MSELoss) with OpenAI's CLIP model to retain semantic consistency between the original and generated images.

The **total loss** during each training step is a **weighted sum** of the individual losses, with the weights indicated by the formula:

$$L_{total} = 10L_{cycle} + 10L_{GAN} + 5L_{sketch} + L_{depth} + L_{CLIP}$$

DISCUSSION & CONCLUSION

We propose a novel GAN architecture that successfully transforms full-colour images into flat-colour images, preserving key features such as depth, geometry, and semantics. The integration of custom loss functions ensures high-quality and consistent outputs. The use of adversarial loss forces the generator to produce realistic images indistinguishable from real flat-colour images, while the cycle consistency loss ensures that an image transformed to the target domain and back retains its original characteristics. The sketch extractor and geometry losses further refine the output by maintaining the essential outlines and depth information, crucial for preserving the structural integrity of the images. Semantic loss, employing the CLIP model, ensures that the generated images retain meaningful content, enhancing the overall transformation quality.

A notable observation during training is a sudden spike in the GAN adversarial loss at around 35 epochs. This spike could be attributed to several factors, including sudden shifts in the discriminator's learning or changes in the data distribution. Such spikes are not uncommon in GAN training and often indicate a temporary instability as the generator and discriminator adjust to each other's improvements. Despite this, the model stabilises and continues to improve. This suggests that our multi-loss CycleGAN approach is robust and capable of overcoming temporary instabilities.

FUTURE WORK

Although this research shows promising results for the application of GANs for full-to-flat colour transformation, the dataset employed were hand-picked and consequently limited to the specific domains of anime portraits and flat cartoons. To ensure the model generalises across diverse image styles, a far larger and more comprehensive dataset must be collected. Further, integrating more advanced deep learning architectures, such as transformer and diffusion models, could further refine image transformation quality. Finally, expanding the model's capability to handle video inputs, preserving temporal coherence, and enhancing user interactivity through customisable style parameters are future areas to explore.