

Real-to-Cartoon Image Translation with CycleGAN

AML 2025 Group 19

Sergei Fedorchenko, Aleksandr Efremov, Maxim Emelianov



Motivation & Use Case

Automatic cartoonization of photos is useful for:

- Creative industries (animation, graphic design)
- Social media filters
- Identity anonymization

Manual cartoonization is time-consuming.

Can we learn the mapping from unpaired photo-cartoon datasets?

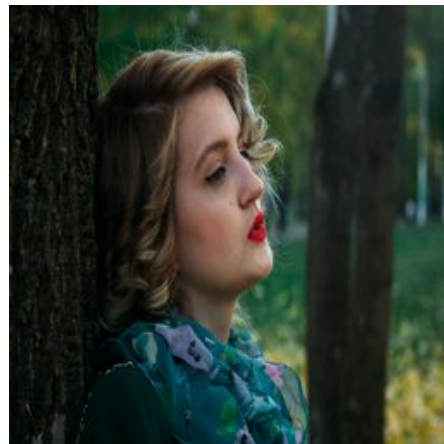


Problem Statement

Goal: Convert real photos into a cartoon-like style using unpaired datasets.

Challenges:

- No pixel-wise supervision (unpaired domains)
- Avoid mode collapse and preserve content structure



Related Work

- **CycleGAN (Zhu et al. 2017)**: unpaired image-to-image translation using cycle-consistency loss.
- **U-GAT-IT (Kim et al. 2020)**: attention-guided translation with stronger stylization.
- **SwinIR (Liang et al. 2021)**: transformer-based backbone for image restoration – we test its use as a generator backbone.

Dataset

Source: Web-crawled photos and cartoons.

Preprocessing:

- Resized to 256×256
- Colors normalized to match the model

Domains:

- Real: 2667 human photographs from Kaggle dataset
<https://www.kaggle.com/datasets/tapakah68/supervisely-filtered-segmentation-person-dataset>
- Cartoon: 412 frames from the soviet cartoon “Трое из Простоквашино” 1978



Evaluation Metrics

- **FID Score** (Fréchet Inception Distance): lower is better.
- **Qualitative Inspection**: Does the output resemble hand-drawn cartoons?
- **Training Stability**: Monitor losses (G, D, cycle, identity).

Our Approach

Start from statistical baseline, then modify generator with:

- ChainGAN approach
- SwinIR, UNet and ResNet backbones
- Test additional losses

Statistical Baseline: Color Transfer

Approach: Match color statistics (mean, std) between real photos and cartoon images

- **train_B** — cartoon-style images
- For each of the 3 channels (R, G, B), computed:
mean_B, std_B
- Did the same for **train_A** (real photos):
mean_A, std_A
- Transformed each image from **test_A** using the formula:
- FID \approx 300 and 368 without the transfer

$$\frac{pixel_{old} - mean_{old}}{std_{old}} \cdot std_{new} + mean_{new}$$



SwinIR: A Transformer Backbone for Image Restoration

Architecture Highlights:

- Hierarchical **Swin Transformer blocks** with **shifted window attention**
- Captures local and non-local context efficiently
- Better than CNNs in fidelity, **but costly to train**

⚠ Practical Limitation in Our Case:

- **Training time was a major bottleneck**
 - Training even **5–10 epochs** took days
 - Needed **100+ epochs** to converge fully
- Transformer-based backbones like SwinIR offer quality gains, but:

Not viable under time/budget constraints in our project

Before



After

ResNet9: Fast and Simple Backbone

- **ResNet9** is a compact ResNet variant:
 - 9 convolutional layers
 - Residual connections to maintain gradient flow
 - Popular for lightweight GAN applications
 - For our project it is:
 - **Much faster** training than transformers
 - **Stable convergence** in early epochs
 - Low memory usage and decent GPU utilization
 - Easy to plug into CycleGAN-style architecture
- FID \approx 238

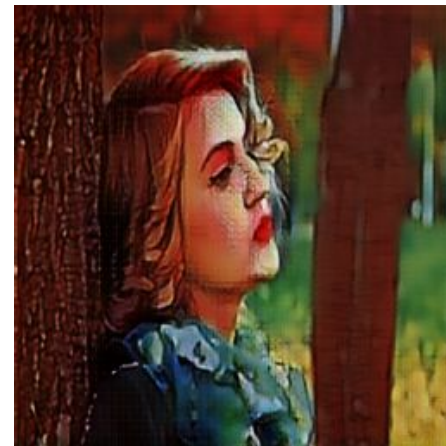
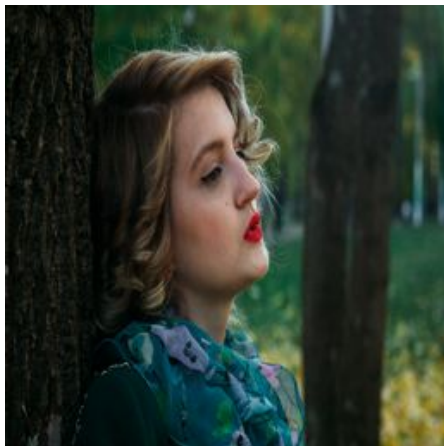
After



Before

U-Net + Improved Loss Performs Better

- We replaced the ResNet generator with a **U-Net** architecture.
- U-Net's skip connections help preserve spatial details, improving output sharpness.
- Combined with a tuned loss: **Cycle + Identity + GAN + Perceptual Loss**.
- Results are:
 - Visually more coherent.
 - **FID score improved** over ResNet9 and statistical baseline (FID \approx 200).
- U-Net proved to be **more sample-efficient**, achieving decent quality in fewer epochs.



Results Overview

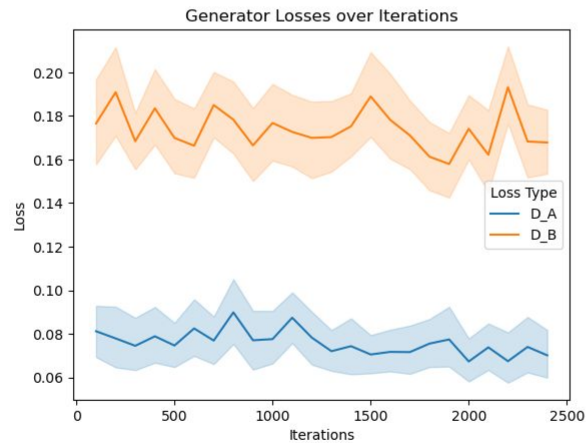
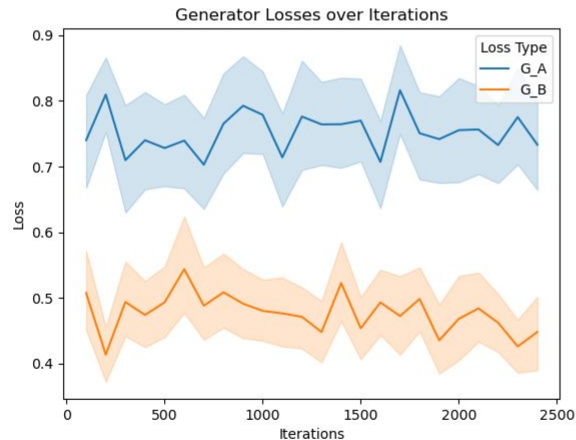
No changes	368
Naive color matching	300
CycleGAN + ResNet9	238
CycleGAN + Unet+ Identity and Perceptual losses	200

Results

FID: **200**

Outputs show consistent coloring, but:

- Style is not always “cartoonish”
- Shapes and expressions sometimes distorted



We did not reach OpenAI Ghibli quality



But maybe the real style translation was the friends we made along the way

Conclusion & Future Work

CycleGAN with tuned losses and backbones produces images with some degree of style-transfer.

FID still high → needs better domain-specific loss or attention mechanisms.

Acknowledgements & Q&A

- Thanks to the AML2025 TAs Deborah Noemie Jakobi, David Robert Reich, and Lena Jäger.
- Questions?

