INM705: Deep Learning for Image Analysis

Title: **Unpaired Image-to-Image Translation using a CLIP-enhanced CycleGAN for Monet Style Transfer**

Due Date: 27 April 2025

Authors:

Sisipho Zinja (240047842)
Git Repo



CITY UNIVERSITY LONDON

1. Introduction

Image-to-image translation is the process of learning a mapping between images from a source domain to a target domain. This fundamental task in computer vision has applications ranging from artistic style transfers to medical imaging. While many deep learning techniques rely on paired training data, where each source image has a corresponding target image, we often find that such datasets are hard to find. Cycle-Consistent Adversarial Networks (CycleGAN) offer a better approach for unpaired image-to-image translation, learning the mapping without the need for direct one-to-one examples (Zhu et al., 2017).

The standard CycleGAN framework learns a mapping $G: X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution of Y using an adversarial loss. This is coupled with an inverse mapping $F: Y \rightarrow X$ and a crucial cycle consistency loss, that enforces $F(G(X)) \approx X$ and $G(F(Y)) \approx Y$. This cycle consistency constrains the unpaired image domains and ensures that the learned mappings are meaningful inverses of each other. A major drawback of training CycleGANs involves a delicate balance between generators and discriminators, which lead to instability, and the reliance on pixel-level and adversarial losses may not always capture high-level semantics effectively.

The aim of this project is to explore an enhancement to the standard CycleGAN framework by leveraging the semantic understanding capabilities of a pre-trained Contrastive Language-Image Pre-trained model (CLIP). CLIP has shown that large-scale contrastive pretraining does produce models with strong zero-shot and semantics alignment capabilities (Radford et al., 2021). We propose replacing the traditional discriminator networks with a loss function derived from CLIP to simplify the architecture and improve semantic guidance. Specifically, we apply this CLIP-enhanced CycleGAN to the task of translating paintings by Claude Monet into photorealistic images, utilizing the publicly available monet2photo dataset.

2. Background and Related Work

CycleGAN is trained using three types of losses:

a. Adversarial Loss encourages the generated images to be indistinguishable from real images in the target domain.
$$L_{GAN}(G_A \rightarrow B, D_B, A, B) = E_b \sim pdata(b)[(D_B(b) - 1)2] + E_a \sim pdata(a)[(D_B(G_A \rightarrow B(a)))2]$$

b. Cycle Consistency Loss ensures that an image translated to the other domain and back returns to its original form:
$$L_{cycle}(G_A \rightarrow B, G_B \rightarrow A) = E_a[\|G_B \rightarrow A(G_A \rightarrow B(a)) - a\|1] + E_b[\|G_A \rightarrow B(G_B \rightarrow A(b)) - b\|1]$$

c. Identity Loss encourages the generator to preserve the input when it is already in the target domain: $L_{identity}(G_A \rightarrow B, B) = E_b[\|G_A \rightarrow B(b) - b\|1]$

CLIP on the other hand is composed of two independent encoders:

a. Image Encoder: A convolutional neural network (ResNet) or vision transformer (ViT) that maps an input image into a feature embedding space.
b. Text Encoder: A transformer model that encodes natural language text (e.g., captions) into the same embedding space.

Both encoders are trained jointly but independently, the similarities between image and text is measured using cosine similarity which we than use to ensure the style transfer preserves semantics.

3. Methodology

Dataset Selection and Preprocessing:

The public monet2photo dataset contains two unpaired sets of 8232 total images:
Domain A: Paintings by Claude Monet (256x256 pixels) normalized to the range [-1, 1].
Domain B: Landscape photographs (256x256 pixels) normalized to the range [-1, 1].

Overall Architecture:

We adapt the standard architecture of CycleGANs by maintaining the two generator structure that maps between two domains (G_{A2B} mapping domain A to B and G_{B2A} mapping B to A) and eliminating adversarial discriminators entirely. We introduce a CLIP-based semantic loss that measures the cosine similarity between the generated images and predefined domain-specific text prompts (e.g., "a painting by Monet", "a photo of a landscape.") The generator is then trained to maximize this similarity, in addition to maintaining cycle consistency and identity preservation. We calculate the total loss by combining the cycle consistency loss, identity loss and CLIP semantic loss, each weighted appropriately. The main idea is to enforce the generated images to align semantically with textual descriptions of the target domain.

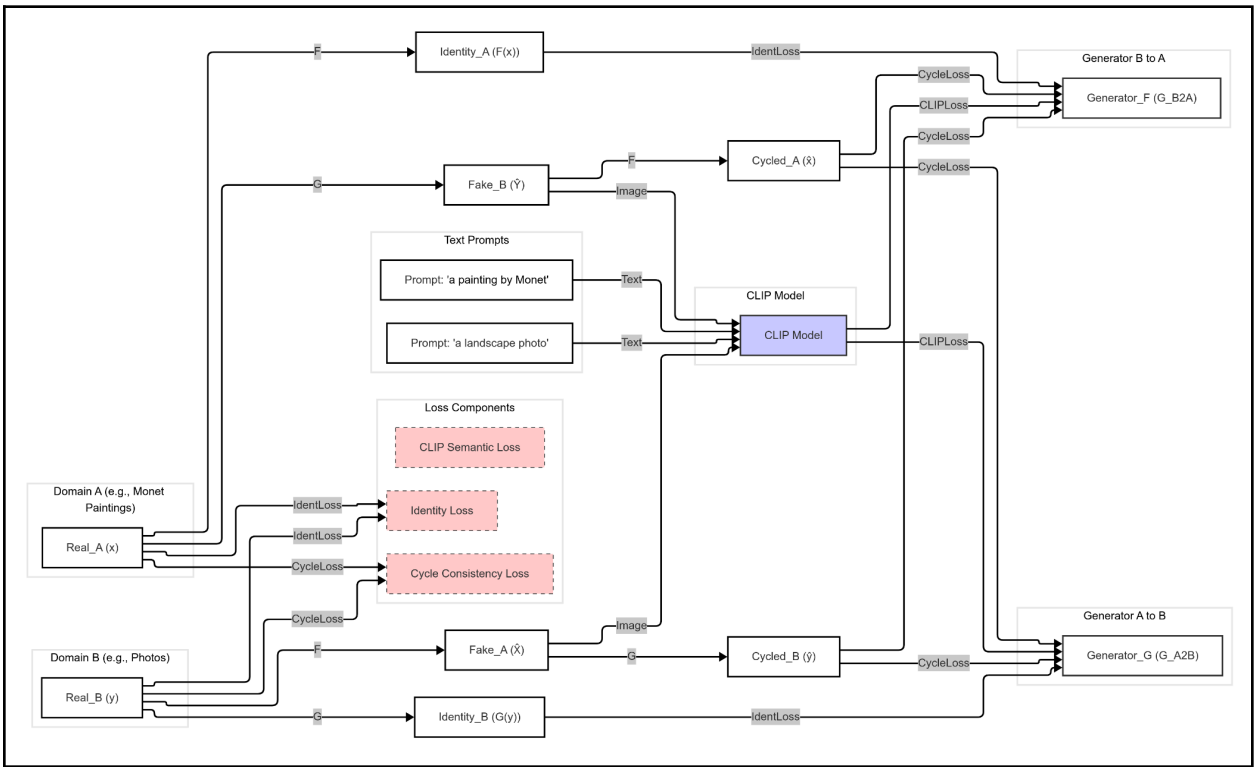


Figure 1: Robust CLIP-enhanced CycleGAN architecture with no discriminators.

Generator Network:

The generator network follows the ResNet-based architecture introduced in the original CycleGAN paper, consisting of an initial 7×7 convolutional block, two downsampling layers, nine residual blocks, two upsampling layers, and a final output layer. Instance normalization is applied after each convolution to stabilize training, and ReLU activations are used throughout except for the final layer, which uses a tanh activation to scale outputs to [-1,1]. The residual blocks allow the network to learn complex transformations while preserving the content structure of the input image. Compared to the original CycleGAN generator, minor simplifications were made, such as consistently using instance normalization and omitting dropout. This architecture provides a strong balance between expressive power and training stability for unpaired image-to-image translation tasks.

CLIP Integration and Semantic Loss:

We utilize a pre-trained CLIP model (e.g., ViT-B/32) with frozen weights from OpenAI. For the A-to-B translation (Monet to Photo), the generated image $\hat{y} = G_{A2B}(x)$ is passed through CLIP's image encoder. At the same time, a text prompt describing the target domain B (e.g., "a photo of a landscape") is passed through CLIP's text encoder. The CLIP semantic loss, L_{CLIP}, is then calculated based on the cosine similarity between the image and text embeddings. A common formulation is to minimize the negative cosine similarity or maximize the similarity directly, encouraging the generated image to align semantically with the target description. The same process is applied for the B-to-A translation using generator G_{B2A} and a prompt for domain A (e.g., "a painting by Monet").

Loss Functions:

The total loss guiding the training of both generators (G_{A2B} and G_{B2A}) is a weighted sum of:

1. **Cycle Consistency Loss (L_{CYC}):** This loss ensures that translating an image to the target domain and back recovers the original image. It is typically implemented using an L1 norm: $L_{cyc}(G,F) = E_{x\sim X}[|F(G(x)) - x|1] + Ey\sim Y[|G(F(y)) - y|1]$
2. **CLIP Semantic Loss (L_{CLIP}):** This loss encourages the generated images to match the semantic description of the target domain, as measured by CLIP. Let I_E be CLIP's image encoder, T_E be the text encoder, p_A be the text prompt for domain A, and p_B be the text prompt for domain B. The loss can be formulated using negative cosine similarity as: $L_{CLIP}(G,F) = - Ex\sim X[cos(I_E(G(x)), T_E(p_B))] - Ey\sim Y[cos(I_E(F(y)), T_E(p_A))]$
3. **Identity Loss ($L_{identity}$):** This loss regularizes the generators to be close to identity mappings when applied to images already in their target domain. It helps preserve color. $Lidentity(G,F) = Ey\sim Y[|G(y) - y|1] + Ex\sim X[|F(x) - x|1]$

Total Loss: $L_{total} = \lambda_{cyc}L_{cyc} + \lambda_{CLIP}L_{CLIP} + \lambda_{identity}L_{identity}$ $where\ \lambda_{cyc}, \lambda_{CLIP}, \lambda_{identity}$ are hyperparameters weighting the contribution of each loss term.

4. Training

The CLIP-enhanced CycleGAN model was trained using the Adam optimizer with betas set to (0.5, 0.999) and an initial learning rate of 0.0002. The batch size was set to 2 due to limited GPU resources. The loss weights were configured as $\lambda cyc = 10,\ \lambda clip = 1.0,\ and\ \lambda identity = 0.5$. Training was conducted for a total of 100 epochs, learning the entirety of the dataset.
For comparison purposes, a baseline standard CycleGAN (with discriminators and adversarial loss, but without CLIP loss) was also trained under similar configurations over 40 epochs. All training runs, including loss curves for each component and periodically generated image samples, were logged using Weights & Biases (W&B) for monitoring and analysis.

5. Evaluation

Due to the unpaired nature of the data, evaluation primarily relies on qualitative visual inspection. Generated images from both translation directions (Monet → Photo and Photo → Monet) were examined for perceptual quality, realism (for photo generation), stylistic consistency (for Monet generation), and semantic correctness relative to the target domain. We looked for successful translations of key features as well as common failure modes like artifacts or incomplete translations.

a. Quantitative Comparison:

Model	**FID Score**
Baseline CycleGAN (40 epochs)	147.02531319913103
CLIP-Enhanced CycleGAN (100 epochs)	181.97669618094943

Table1: FID Scores for Monet-to-Photo ($A \rightarrow B$) Translation

To quantitatively assess the similarity between the generated images and real images, the Fréchet Inception Distance (FID) was calculated. The scores compare the outputs from each model against the original images in our testB shown in Table 1 above. Lower scores indicate that the generated image is closer to the real distribution. From our results, the Baseline CycleGAN model achieved a lower FID score of 147.03 after 40 epochs compared to the CLIP-Enhanced model's score of 181.98 after 100 epochs. The baseline models outputs were quantitatively closer to the distribution of the real images. This outcome contrasts with some qualitative observations where the CLIP model shows better semantic alignment. The significant difference in training duration (40 vs 100 epochs) might be causing this discrepancy; the CLIP model might require further tuning or longer training to potentially improve its FID score, or perhaps the CLIP loss, while enforcing semantics, pushes the image distribution away from the specific features captured by FID in this particular setup. It might be interesting to investigate how CLIP loss weighting and training duration impact quantitative measures like FID.

b. Qualitative Comparison: It is clear that our CLIP-enhanced CycleGAN model outperforms the standard baseline CycleGAN model. Just comparing the original image, with that generated by the baseline model and CLIPCycleGAN we can clearly tell the stylistic and semantic differences between the two approaches.

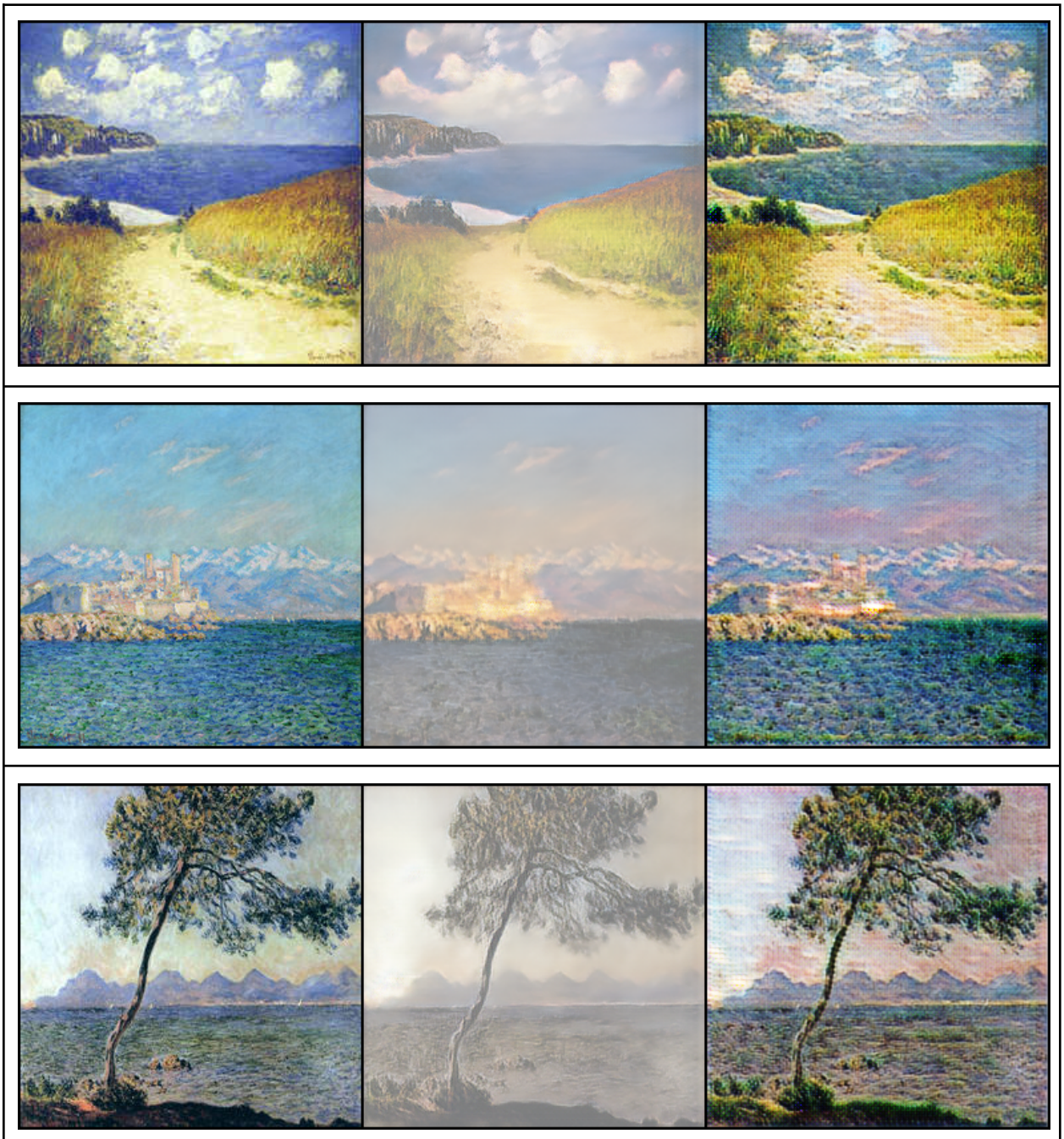


Figure 2: Side by side comparison of qualitative results showing the Input image (left), Baseline CycleGAN image (center), and our CLIP-enhanced CycleGAN (right).

6. Discussion

The CLIP-enhanced model uses a loss function that directly maximizes the similarity between the generated image and a specific text prompt (e.g., "a photo of a landscape"). This strongly guides the generator towards producing outputs that match that semantic description. While this improves semantic coherence it might push the generated images towards a narrower, CLIP-preferred interpretation of the prompt. This could subtly shift the overall distribution of generated images away from the true, diverse distribution of real landscape photos found in the dataset. For example, CLIP might have biases towards certain types of landscapes or lighting conditions based on its own training data. Since FID compares the entire distribution, if the CLIP-guided distribution, slightly shifted or was less diverse than the real distribution, the FID score could get worse, even if individual images look good semantically. We can claim that the CLIP model might be optimizing for semantic correctness over precise distributional matching in the feature space used by FID.

7. Conclusion

On exploring how FID works, we can see that the baseline CycleGAN has an advantage with an adversarial discriminator that is explicitly trained to make generated images indistinguishable from the real images. The baseline model has two discriminators that focus on realistic textures and local details to be able to spot which images are fake. It is clear that the baseline model, even with 40 epochs, is actually quite good at generating the low-level feature statistics of the real images resulting in a better FID score. Judging from the visual comparison alone, it might be fair to claim that the baseline is manipulating the FID metric by producing images that look statistically similar.
In addition, comparing a model trained at 40 epochs with one trained for 100 proves to be a challenge. FID scores usually don't improve with more training, it would be interesting to investigate further how to account for this.

8. Reflection

Throughout this project, I had the opportunity to deeply engage with the challenges and intricacies of unpaired image-to-image translation. Initially, I was drawn to CycleGAN for its elegance in handling tasks without paired datasets. However, as I implemented and trained the baseline CycleGAN model, I quickly realized the practical difficulties of adversarial training. Balancing the generator and discriminator losses, tuning hyperparameters, and maintaining training stability required much more sensitivity and intuition than I had anticipated. I gained a deeper appreciation for why GANs, despite their powerful results, are considered notoriously difficult to train reliably.

In exploring ways to improve and simplify CycleGAN, I learned about the capabilities of contrastive models like CLIP. Integrating CLIP into my modified CycleGAN architecture opened up an entirely new perspective on how semantic understanding, learned from large-scale vision-language pretraining, could directly supervise generative models. Designing a system where a frozen CLIP model replaced the traditional discriminator challenged me to think creatively about loss functions, model guidance, and the broader meaning of "realism" in generated outputs. Instead of focusing only on pixel-level consistency, I learned to value and measure high-level semantic alignment.

This project also forced me to improve my practical machine learning skills. I developed a more disciplined workflow: modularizing code into clean training scripts, using tools like Weights & Biases for experiment tracking, and writing robust dataset loaders. I saw firsthand how important reproducibility and monitoring are, especially when training complex models over many epochs. I also became more comfortable using PyTorch at a deeper level, adjusting architectures, customizing loss functions, and troubleshooting training behavior.

Reflecting back, this project taught me not just technical knowledge about GANs, CycleGAN, and CLIP, but also broader lessons about the research process itself. I learned that improving a machine learning model is not only about bigger architectures or more data, but about thoughtful problem analysis, creative solutions, and disciplined experimentation. I leave this project with a stronger intuition for model design, a deeper respect for the complexity of generative models, and an excitement to continue exploring the intersection of vision and language models in the future.

References

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML).*

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*