Reimplementing PyTorch Cycle GANs for Art Style to Art Style Conversion

Shrinivas Venkatesan
venkat97@purdue.edu

Tejas Yalamanchili
tyalaman@purdue.edu

1. Introduction

Image-to-image translation is a fundamental problem in computer vision, aiming to learn the mapping between an input image and an output image using a training set of aligned image pairs. However, for many tasks, specifically artistic style transfer, paired training data will not be available. We cannot easily obtain a photograph of a specific landscape and a corresponding painting of that exact scene by a long-deceased artist. The introduction of Cycle-Consistent Adversarial Networks (CycleGAN) [5] represented a breakthrough in this domain, allowing translation between domains without the need for paired examples.

Despite the success of CycleGAN, significant challenges remain in the realm of artistic style transfer. Standard approaches often struggle to capture high-frequency details, such as specific artistic textures and brush strokes, resulting in outputs that look like color-filtered photographs rather than genuine paintings. Furthermore, the reverse task (translating art into realistic photographs, *Art2Photo*) presents a distinct and more difficult challenge. In this direction, the model must "hallucinate" realism, inferring photorealistic details from abstract or impressionist representations where such information does not explicitly exist.

In this work, we propose a re-implementation and extension of the CycleGAN architecture tailored specifically for bi-directional art-style translation. Our motivation is twofold. First, we aim to refine the generation of artistic textures to allow users to visualize their own photographs as if painted by masters such as Monet or Van Gogh. Second, we seek to refine the current performance of *Art2Photo* translation by enforcing stronger structural and perceptual constraints.

While the original CycleGAN (the baseline from the paper) relies primarily on Least Squares GAN (LSGAN) loss and cycle-consistency loss, we hypothesize that this objective function is insufficient for high-fidelity artistic rendering. Therefore, we explore a modified objective landscape. We replace the standard adversarial component with Hinge loss to improve training stability. Furthermore, to better capture the "style" and "content" distinct from pixel-level accuracy, we incorporate perceptual content loss (LPIPS/VGG), Gram-matrix style loss, and discriminator feature matching loss.

Our goal is to produce a model that creates high-quality, lightweight style transfer results, balancing the preservation of content structure with the synthesis of realistic artistic textures.

In summary, our contributions are as follows:

- We provide a PyTorch implementation of CycleGAN specialized for *Photo2Art* and *Art2Photo* domains.
- We introduce a composite loss function that integrates Hinge loss, Perceptual loss (LPIPS), and Gram-matrix constraints to improve style fidelity and training stability over the baseline LSGAN approach.
- We evaluate the difficulty of the *Art2Photo* task, analyzing the model's ability to hallucinate realism from stylized inputs.
- We present qualitative and quantitative comparisons demonstrating how these additional regularization terms influence content preservation and visual realism.

(a) Photo2Art Results (b) Art2Photo Results

Figure 1. Side-by-side comparison. (a) Photo translated to Art. (b) Art translated to Photo.

2. Related Works

Unpaired Image-to-Image Translation. The core of our investigation is grounded in the breakdown of reliance on paired training data. Early successes in image translation, such as Pix2Pix, required perfectly aligned image pairs (e.g., a specific edge map and its corresponding photo). However, obtaining such pairs for artistic domains is fundamentally impossible; one cannot capture a photograph of a scene as it appeared to a long-deceased artist.

Our work directly builds upon the seminal CycleGAN framework [5], which introduced cycle-consistency loss to learn mappings between domains X and Y without paired examples. By enforcing a bijection where $F(G(x)) \approx x$, the model ensures that content is preserved during the style transfer process. As demonstrated in the original paper, this allows for the translation of natural scenes into the styles of Monet, Van Gogh, Ukiyoe, and Cezanne. We adopt this foundational architecture but pivot the focus toward optimizing the objective function specifically for the asymmetric challenges of Art2Photo and Photo2Art tasks.

Loss Functions and Stability in GANs. Generative Adversarial Networks are notoriously difficult to train due to convergence instability and mode collapse. Since our approach relies on experimentation with alternative objectives, we draw insights from [1], which provides a comprehensive analysis of loss function efficacy in image-to-image translation. Their work highlights that while standard Min-Max loss is effective, it is often prone to vanishing gradients. We aim to leverage their findings regarding alternative formulations (specifically Hinge loss and Least Squares loss) to stabilize the adversarial game.

Furthermore, we investigate the impact of non-adversarial regularization terms. While standard GAN losses focus on pixel-level distributions, they often fail to capture high-level perceptual quality. This motivates our integration of style-specific losses (such as Gram matrices) to better enforce the transfer of artistic texture without compromising the underlying content structure.

Domain Adaptation and Semantic Consistency. The challenge of translating between art and realism is analogous to domain adaptation, where the goal is to generalize across distributions. We examine [3], which details the CyCADA framework. The authors highlight a critical insight: combining pixel-level translation with feature-level adaptation helps overcome both low-level appearance differences (texture, color) and higher-level semantic gaps.

Crucially, they introduce semantic consistency losses to ensure that the "meaning" of the image remains invariant during translation, a car in a painting should remain a car in the generated photo. This approach demonstrates that domain-focused training can significantly improve downstream performance. This is directly relevant to our motivation, as we seek to prevent the model from hallucinating artifacts that distort the semantic reality of the input image during the Art2Photo conversion.

Perceptual Metrics and Evaluation. Evaluating generative models, particularly for artistic synthesis, remains an open challenge as standard metrics like MSE (Mean Squared Error) correlate poorly with human perception. To accurately assess our improved model's effectiveness, we will leverage two primary metrics: the Fréchet Inception Distance (FID) [2] and Learned Perceptual Image Patch Similarity (LPIPS) [4].

We utilize FID specifically to quantify the realism of our *Art2Photo* translations. As noted by Heusel et al. [2], FID improves upon the Inception Score by comparing the statistics of generated samples against real samples in the feature space of an Inception-v3 network. This is crucial for our task because there is no ground truth photo for a given painting; the model must "hallucinate" photorealistic details (textures, lighting) that do not exist in the source art. FID allows us to measure how closely the distribution of these hallucinated details matches the manifold of real-world photography. Beyond the metric itself, this framework allows us to rigorously benchmark our proposed objective functions. We use FID to determine if our switch from the baseline LSGAN loss to our improved Hinge and perceptual loss formulations actually results in a statistically significant improvement in the realism of the generated textures.

Complementing this, we employ LPIPS to measure content preservation, as proposed by Zhang et al. [4]. The authors demonstrate the "unreasonable effectiveness" of deep features, showing that internal activations of networks like VGG or AlexNet correlate much better with human perceptual judgment than traditional metrics like L2 or PSNR. In our context, LPIPS is vital for ensuring that while the style changes, the structural content remains perceptually unchanged. It penalizes the model if it alters the composition (e.g., removing a tree) but remains robust to the necessary stylistic changes in color and texture that naive pixel-metrics would penalize.

3. Approach

3.1. Baseline Model

3.1.1. Architecture

The baseline model is a standard CycleGAN implementation with ResNet generators and PatchGAN discriminators. We work with two image domains X and Y and instantiate:

Two generators: $G_{XY} : X \to Y$ and $F_{YX} : Y \to X$

Two discriminators: D_X for domain X and D_Y for domain Y

ResNet Generator:

Each generator is a fully convolutional encoder and decoder network:

Input Block: Reflection padding of 3 pixels, a 7×7 convolution with stride 1, followed by InstanceNorm and ReLU.

Downsampling: Two strided 3×3 convolutional layers with a stride of 2 and padding of 1. Each layer doubles the number of feature channels (64 to 128 to 256) with InstanceNorm and ReLU after each convolution. This produces a low-resolution, high-dimensional feature map.

Residual Bottleneck: 9 ResNet blocks at the bottleneck resolution, where each block consists of reflection padding of 1 pixel, a 3×3 convolution layer followed by norm and ReLU, and another 3×3 convolution layer followed by norm. A residual skip connection allows these blocks to learn complex transformations while preserving information through the skip paths.

Upsampling: Two transposed convolutional layers mirroring the encoder. Each is a 3×3 `convTranspose2d` with a stride of 2, padding of 1, and output padding of 1. This divides the number of channels by 2 at each step (256 to 128 to 64), followed by InstanceNorm and ReLU.

Output Block: Reflection padding of 3 pixels, a final 7×7 convolution mapping back to `out_channels`, followed by a tanh activation.

All conv and norm layers are initialized with an `init_weights` function which applies a normal initialization to convolution weights and unit-mean initialization to normalization scales, helping stabilize early training.

PatchGAN Discriminator:

Each discriminator D_X and D_Y is a PatchGAN that produces a spatial map of logits instead of a single scalar.

Input Block: 4×4 convolution layer with a stride of 2 and padding of 1, followed by LeakyReLU.

Downsampling Stack: 3 layers of convolution blocks. For layers 1 to 2, the model uses: a 4×4 convolution with stride 2 and padding 1, InstanceNorm, followed by LeakyReLU. The number of channels doubles with each layer.

Final Convolutions: One more 4×4 convolution block with a stride of 1 and padding of 1, followed by norm and LeakyReLU, followed by a last 4×4 convolution to get 1 channel.

The result is an $H' \times W'$ real vs. fake score map, where each element corresponds to a 70×70 image patch. All discriminator layers are also initialized with `init_weights`.

The Baseline model simply ties these components together in a cycle structure:

$x \in X$ is mapped to $y' = G_{XY}(x)$ and reconstructed as $x' = F_{YX}(y')$.

$y \in Y$ is mapped to $x' = F_{YX}(y)$ and reconstructed as $y' = G_{XY}(x')$.

Therefore, the architecture is a symmetric two-domain CycleGAN: two identical ResNet generators, two identical PatchGAN discriminators, and a forward pass that forms $X \to Y \to X$ and $Y \to X \to Y$ cycles on which the baseline LSGAN, cycle, and identity losses are applied.

3.1.2. Objective Function

The Baseline model keeps the standard training objective with: LSGAN + Cycle Consistency Loss + Identity Loss. The total objective function can be written as:

$$L_G = L_{adversarial}^{LSGAN} + \lambda_{cycle} L_{cycle} + \lambda_{identity} L_{identity}$$

Below we summarize each component of this baseline multi-loss objective:

- **Adversarial Loss (LSGAN):**

  In the Baseline model, we use the Least-Squares GAN (LSGAN) objective as the adversarial loss instead of the hinge loss. For each discriminator $D \in \{D_X, D_Y\}$, given real samples x and generated samples x', the discriminator loss is:

  $$L_{discriminator}(D) = \frac{\mathbb{E}_x[(D(x)-1)^2]}{2} + \frac{\mathbb{E}_{x'}[(D(x'))^2]}{2}$$

  Here, real images are encouraged to have the discriminator output close to 1 and fake images close to 0. The corresponding generator adversarial loss is:

  $$L_{adversarial}(Generator) = \frac{\mathbb{E}_{x'}[(D(x')-1)^2]}{2}$$

  Therefore, generators are trained so that their fake samples are classified as "real" by the discriminators. In practice, we sum this loss over both discriminators D_X, D_Y for the discriminator objective, and over both generators G and F for the generator objective.

  The intuition behind LSGAN is that it treats the discriminators as a regressor to continuous labels—1 if real and 0 if fake—penalizing $(D(x) - \text{label})^2$. This provides smoother, non-saturating gradients than the original log-GAN loss (BCE) and stabilizes training. However, because the quadratic penalty strongly pulls predictions towards the exact labels, the resulting image can be slightly softer or less sharp. This motivates switching to the hinge loss in the improved model to obtain crisper textures and more pronounced high-frequency details.
- **Cycle and Identity Losses:**

  We use the cycle and identity losses to preserve structure and colors.

Cycle Consistency:

$$L_{cycle} = \mathbb{E}_x[||F_{YX}(G_{XY}(x)) - x||_1] + \mathbb{E}_y[||G_{XY}(F_{YX}(y)) - y||_1]$$

Identity:

$$L_{identity} = \mathbb{E}_x[||F_{YX}(x) - x||_1] + \mathbb{E}_y[||G_{XY}(y) - y||_1]$$

These terms encourage G_{XY} and F_{YX} to be approximately invertible and to act close to the identity on images already in the target domain. Therefore, in practice, they prevent large geometric distortions and unnecessary color shifts.

3.2. Improved Model

3.2.1. Architecture

The improved model keeps the same architecture as our baseline CycleGAN implementation and only modifies the training objective. We use two ResNet generators $G_{XY} : X \to Y$ and $F_{YX} : Y \to X$ with 2 downsampling layers, 9 residual blocks, and 2 upsampling layers, followed by a tanh output. Each generator uses instance normalization and residual connections to stabilize training and preserve low-level structure.

For the Discriminator part, we use two PatchGAN discriminators D_X and D_Y that output a grid of real/fake scores over local image patches instead of a single scalar, encouraging realistic high-frequency details. The only change in the improved model is the objective function used to train these networks. The network capacity and topology remain identical to the baseline.

3.2.2. Improved Objective

We replaced the baseline LSGAN adversarial loss with a Hinge adversarial loss and, in addition to the cycle and identity losses of the baseline objective, added several perceptual and feature-based terms. The total generator loss is:

$$L_G = L_{adv} + \lambda_{cycle}L_{cycle} + \lambda_{identity}L_{identity} + \lambda_{content}L_{content} + \lambda_{style}L_{style} + \lambda_{feat}L_{feat}$$

Below we summarize each component of the multi-loss objective described above:

- **Adversarial Loss (Hinge):**

  The main difference between the Baseline model and the improved version is that we use a Hinge loss instead of the LSGAN loss. Therefore, for each discriminator $D \in \{D_X, D_Y\}$, given real samples x and generated samples x', the loss is:

  $$L_{disc}(D) = \mathbb{E}_x[\max(0, 1 - D(x))] + \mathbb{E}_{x'}[\max(0, 1 + D(x'))]$$

  The corresponding generator loss is:

  $$L_{adversarial}(Generator) = -\mathbb{E}_{x'}[D(x')]$$

  We sum the loss for each discriminator to get the final adversarial loss for the discriminator, and sum the adversarial loss for each generator G and F to get the final generator adversarial loss. The intuition is to get a hinge margin that encourages $D(real) \geq 1$ and $D(fake) \leq -1$, preventing discriminator saturation and giving the generators strong, stable gradients. This leads to sharper shapes and more realistic textures than the baseline GAN loss.
- **Cycle and Identity Losses:**

  We retain the standard CycleGAN cycle and identity losses to preserve structure and colors, identical to the baseline equations provided above.
- **VGG Perceptual Content Loss:**

  To better preserve high-level content, we add a perceptual loss using a pre-trained VGG19 network. Let ϕ_l denote VGG features at a mid-level layer (e.g., ReLU layer 4).

  $$L_{content} = ||\phi_4(G_{XY}(x)) - \phi_4(x)||_1 + ||\phi_4(F_{YX}(y)) - \phi_4(y)||_1$$

  This loss encourages the translated images to keep the same semantic structure (object shapes, layout, and contours) as the inputs, even though their style changes. Compared to pure pixel L_1, the VGG features are more tolerant to small shifts but penalize large structural distortions.
- **Style Loss using Gram Matrices:**

  We explicitly match style using Gram matrices of VGG features. For features $\phi_l(z) \in \mathbb{R}^{C_l \times H_l \times W_l}$, we compute:

  $$Gram_l(z) = \frac{1}{C_l \cdot H_l \cdot W_l}(\phi_l(z) \cdot \phi_l(z)^T)$$

  The style loss matches Gram matrices of generated images to those of real target-domain images:

  $$L_{style} = \sum_l ||Gram_l(G_{XY}(x)) - Gram_l(y)||_1 + \sum_l ||Gram_l(F_{YX}(y)) - Gram_l(x)||_1$$

  Because Gram matrices capture channel-channel correlations rather than exact spatial positions, this term enforces similar texture, color statistics, and brush strokes while being agnostic to where they occur in the image.
- **Discriminator Feature-Matching Loss:**

  Finally, we add a feature-matching loss on intermediate activations of the discriminators. Let f_k^X and f_k^Y be features from layer k of D_X and D_Y. We minimize:

  $$L_{feat} = \sum_k (||f_k^X(F_{YX}(y)) - f_k^X(x)||_1 + ||f_k^Y(G_{XY}(x)) - f_k^Y(y)||_1)$$

With the real features detached so only the generators are updated, this encourages generated images to match the discriminator's internal feature statistics of real images, stabilizing training and reducing artifacts such as flickering or inconsistent textures.

3.2.3. Effect of the Improved Objective

Overall, the architecture of the improved model is unchanged, but the new objective constrains the generators more tightly: the hinge adversarial loss and style loss push images towards the target domain's appearance while cycle, identity, and VGG content losses preserve geometry and scene structure. Feature matching further aligns generated and real images in the discriminator feature space. In our experiments, this combination produced clearer photo-to-art translations with sharper brush strokes and more realistic art-to-photo outputs while maintaining the structural integrity of the original artworks.

3.3. Training Loop

We train both the baseline CycleGAN and the improved model using the same core training loop, differing only in the loss function and the number of epochs.

3.3.1. Data Pipeline and Image Downsampling

We use the standard unpaired datasets in the CycleGAN format with two folders, `trainA` and `trainB`, for art and photo domains respectively. Our `ImageDataset` class collects all file paths in `trainA` and `trainB`, shuffles them with a fixed seed, and subsamples up to 1000 images per domain to keep training time to a feasible amount.

3.3.2. Image Transformation

Each image is transformed with two functions in our training loop code: `Resize()` and `CenterCrop`, and is converted to a tensor and normalized to $[-1, 1]$. Therefore, all training happens on 128×128 RGB crops.

This is smaller than the 256×256 resolution used in the original CycleGAN paper and is a deliberate trade-off to fit the ResNet generators and PatchGAN discriminators comfortably into our GPU time limit of 4 hours while still preserving enough spatial detail for style transfer. We also use a small batch size and 4 workers in the DataLoader to balance GPU and CPU utilization.

3.3.3. Optimizers, Learning Rate Schedule, Hyperparameters, and Updates

We use separate Adam optimizers for generators and discriminators (one for the generator pair of G_{XY} and F_{YX}, and one for the discriminator pair of D_X and D_Y).

For both baseline and improved models, we keep the common hyperparameters identical to the CycleGAN paper: $lr = 0.0002, \beta_1 = 0.5, \beta_2 = 0.999$.

The learning rate scheduling is handled by `LambdaLR`.

For the **baseline**, we train for 80 epochs with the hyperparameters from the paper and use a constant learning rate.

For the **improved model**, we train for 20 epochs with a fixed learning rate, followed by 20 epochs of linear decay to 0 (therefore, 40 total epochs). We keep all the shared hyperparameters the same as the baseline and only tune the new loss weights. For the hyperparameters related to the additional losses, we derived the values through small trial and error sweeps, selecting the combination that qualitatively gave the best trade-off between strong style transfer and content preservation.

Updates:

- **Generator Update:**

  Freeze the discriminator parameters, zero the generator gradients, and compute generator losses (either baseline or improved). Then backpropagate and step the generator optimizer.
- **Discriminator Update:**

  Unfreeze the discriminator parameters. Zero the discriminator gradients and compute the discriminator losses (either baseline or improved). Then backpropagate this total loss and step the discriminator optimizer.

We also cap the number of iterations for every epoch to 1000 to reduce the training time. We log losses every 50 iterations. Every 10 epochs, we save a full checkpoint and save a grid of sample images showing real-to-fake-to-reconstruction in both directions (photo-to-art and art-to-photo). Before saving, we denormalize the images back to $[0, 1]$ and resize to 128×128 to make samples visually comparable across epochs and between the baseline and improved models.

4. Results

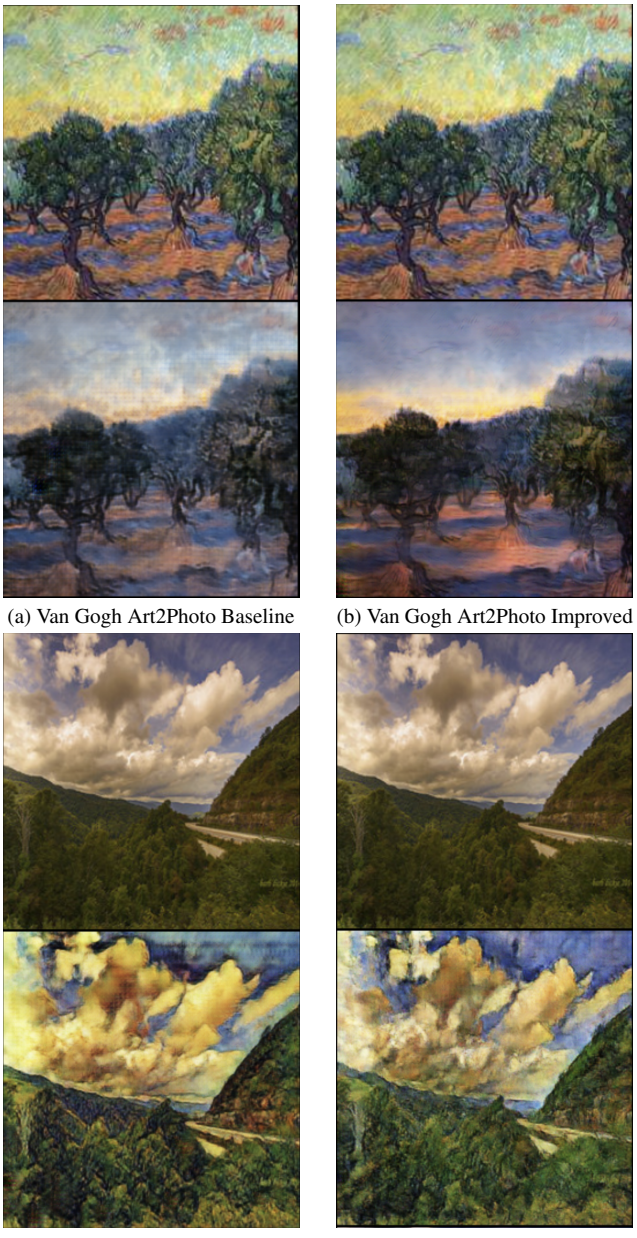
In alignment with our initial objectives, we successfully reimplemented the CycleGAN architecture and validated its performance against the baseline established by Zhu et al. Despite constraints on GPU computational power, we successfully trained both the standard baseline and our proposed improved model to convergence.

As anticipated at the outset of this study, modifying the training objective proved to enhance performance in artistic domains. While the baseline model demonstrated the fundamental ability to translate between visual domains, our improved architecture, utilizing a customized Hinge loss augmented with VGG feature matching and specific style losses, yielded significantly higher fidelity results.

For Photo-to-Art translation, the improved model confirmed our hypothesis regarding texture preservation. It produced clearer images that retained crucial artistic details, such as specific brush strokes, color palettes, and texture patterns, which the baseline frequently obscured.

For Art-to-Photo translation, the results aligned with our expectation that a domain-focused training paradigm would improve generalization. The improved model generated realistic photographic images while maintaining the semantic integrity and overall scene structure of the original artworks.

4.1. Output Images



(a) Van Gogh Art2Photo Baseline (b) Van Gogh Art2Photo Improved

(c) Van Gogh Photo2Art Baseline (d) Van Gogh Photo2Art Improved

Figure 2. **Art2Photo:** The improved model produces a more true to life color pallette, with the sunset looking more realistic.
Photo2Art: The baseline sky region demonstrates weaker stylistic consistency in brush stroke patterns.

(a) Ukiyoe Art2Photo Baseline (b) Ukiyoe Art2Photo Improved

(c) Ukiyoe Photo2Art Baseline (d) Ukiyoe Photo2Art Improved

Figure 3. **Art2Photo:** The baseline model hallucinates a seemingly sunset-looking background due to the red patch in the top left corner of the artwork. The improved model does not do this.
Photo2Art: The improved model generates a more ukiyoe–like image because it uses flatter, warmer colors and compressed depth.

4.2. LPIPS and FID Scores

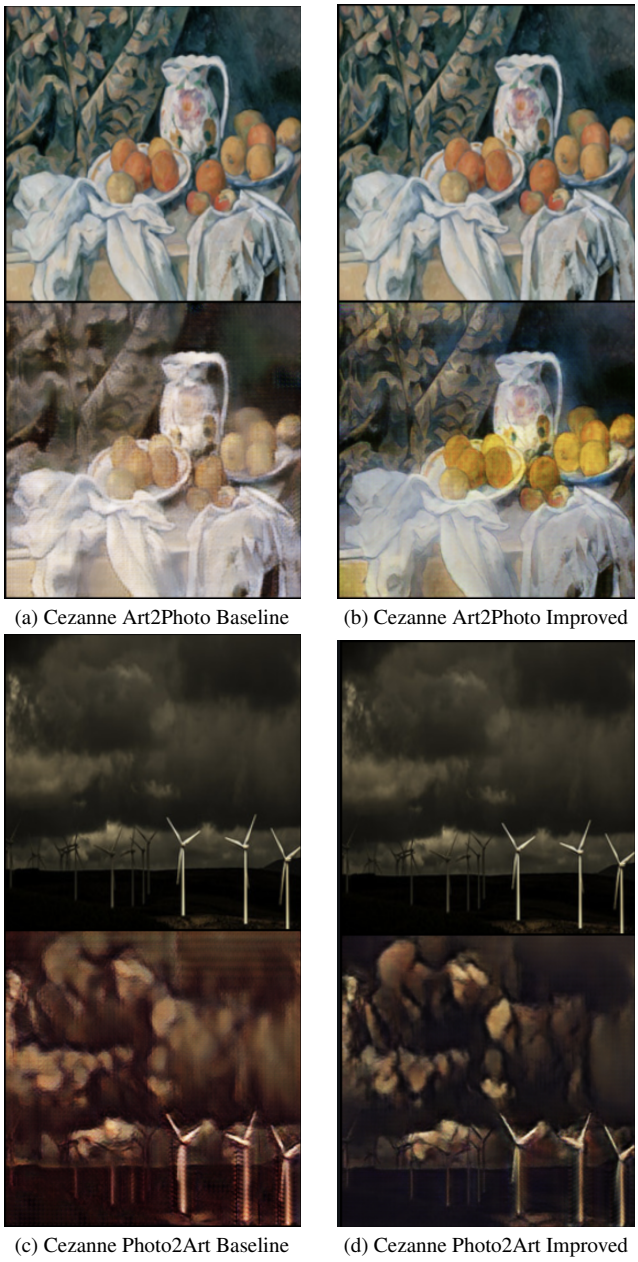
To quantitatively evaluate our generated images, we used two standard metrics: Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS).

FID (Fréchet Inception Distance): This metric assesses the similarity between the distribution of generated images

(a) Monet Art2Photo Baseline (b) Monet Art2Photo Improved

(c) Monet Photo2Art Baseline (d) Monet Photo2Art Improved

Figure 4. **Art2Photo:** The baseline model produces a much grainier image, where individual pixels are visible. The improved model generates a higher resolution image with better contrast.
Photo2Art: The improved model eliminates the blurriness found in the artwork generated by the baseline model. The colors are also significantly less muddled.



(a) Cezanne Art2Photo Baseline (b) Cezanne Art2Photo Improved

(c) Cezanne Photo2Art Baseline (d) Cezanne Photo2Art Improved

Figure 5. **Art2Photo:** The colors show much more contrast in the improved model, making the scene elements look much more realistic, especially the curtain. The colors look dull and washed out in the baseline.
Photo2Art: Excessive smears present in the converted artwork (baseline) give it an element of motion blur compared to the structured brushstrokes and solid construction of objects seen in the artwork generated by the improved model.

and the distribution of real images. A lower FID score indicates that the generated images are more statistically similar to the real dataset, implying higher realism and diversity. We calculated the Best FID for Art-to-Photo (A2P) translation to measure how well our model "hallucinates" realistic photographs from artworks.

LPIPS (Learned Perceptual Image Patch Similarity): This metric measures the perceptual distance between image patches. Unlike pixel-level metrics (like MSE), LPIPS aligns more closely with human visual perception. Lower

LPIPS scores indicate that the generated images are perceptually closer to the target domain's characteristics. We evaluated the Best LPIPS for both Art-to-Photo (A2P) and Photo-to-Art (P2A) tasks.

Table 1 summarizes our quantitative results across four artistic styles: Cezanne, Monet, Ukiyoe, and Van Gogh.

Table 1. Quantitative comparison of Baseline vs. Improved models across four style datasets. Lower scores indicate better performance.

Style	Ver	Best FID (A2P)	Best LPIPS (A2P)	Best LPIPS (P2A)
Cezanne	Baseline	2.91	0.2701	0.2775
Cezanne	Improved	**2.12**	**0.2247**	**0.2605**
Monet	Baseline	2.21	0.2579	0.2352
Monet	Improved	**2.00**	**0.2228**	**0.1915**
Ukiyoe	Baseline	2.91	0.2577	0.2717
Ukiyoe	Improved	**2.85**	**0.2267**	**0.2679**
Van Gogh	Baseline	2.26	0.2635	0.2757
Van Gogh	Improved	**2.13**	**0.2187**	**0.2649**

As shown in Table 1, our improved model consistently outperforms the baseline across all metrics and datasets.

When aggregated across all four artistic domains, the proposed architecture demonstrates clear superiority. On average, our model improved **Art-to-Photo FID scores by 10.98%**, indicating a substantial leap in the statistical realism of generated photographs. Perceptual quality saw even greater gains, with **LPIPS (A2P) improving by an average of 14.92%**.

For the **Photo-to-Art (P2A)** task, we observed an average **LPIPS improvement of 7.50%**. While this margin is tighter than the A2P task, the consistent reduction in perceptual distance confirms that our style-specific losses helped the model better capture the distinct textures of the target artworks. The most notable improvement was seen in the Monet dataset (0.2352 to 0.1915), suggesting that our architecture is particularly effective at handling impressionistic textures.

5. Conclusion

GitHub Link:

https://github.com/YalmanchiliTejas/ArtStyleToArtStyle

In this work, we presented a comprehensive study on enhancing unpaired image-to-image translation specifically for bi-directional artistic style transfer. Motivated by the limitations of standard GAN objectives, which often fail to capture high-frequency artistic textures or hallucinate missing photorealistic details, we proposed a refined architecture incorporating a composite loss function.

By integrating Hinge loss for training stability with Perceptual (LPIPS/VGG) and Gram-matrix style losses, we successfully moved beyond the surface-level adaptations typical of the baseline CycleGAN. Our quantitative analysis confirms the efficacy of this approach, demonstrating an average improvement of 10.98% in FID and 14.92% in LPIPS for *Art2Photo* translation across four distinct artistic datasets. These metrics indicate that our model does not merely memorize broad domain features but effectively bridges the gap between abstract artistic representations and natural image statistics.

Qualitatively, our results address the core challenges outlined in our introduction. For the *Photo2Art* task, our model overcame the "color-filter" effect, generating outputs with genuine brush stroke textures and structural patterns characteristic of the target artists. Conversely, in the challenging *Art2Photo* domain, the model demonstrated a robust ability to infer photorealism, reconstructing plausible depth and lighting from flattened, stylized inputs.

Ultimately, this work highlights that while cycle-consistency is fundamental for unpaired translation, a domain-aware objective function is essential for high-fidelity artistic rendering. Future work may explore applying these improved loss landscapes to higher-resolution generation or video style transfer to further extend the preservation of temporal and semantic consistency.

References

[1] Alaa Abu-Srhan, Mohammad A.M. Abushariah, and Omar S. Al-Kadi. The effect of loss function on conditional generative adversarial networks. *Journal of King Saud University - Computer and Information Sciences*, 34(9):6977–6988, 2022. 2

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 2

[3] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation, 2017. 2

[4] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 2

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. 1, 2