Attention-Enhanced CycleGAN for Unpaired Image-to-Image Translation

Wenqi Zheng

Department of Mathematics, University College London, London, United Kingdom
wenqi.zheng.23@ucl.ac.uk

Abstract. Cycle-consistent GANs are widely used for unpaired image-to-image translation, but they often over-translate textures in regions that should remain largely unchanged (e.g., background grass or sky in horse↔zebra). This failure mode is encouraged by discriminators that score the full image uniformly, which can reward the generator for spreading domain-specific cues beyond the intended foreground. The paper proposes a lightweight, plug-in attention head attached to the CycleGAN generator to predict a soft foreground attention mask. During training the paper uses this mask to (i) gate the discriminator's input and (ii) weight the adversarial loss so that discriminator feedback is concentrated on attended regions. A key stability issue in attention-guided adversarial training is that attention maps can drift late in training; to prevent this, the paper monitor attention-map changes across epochs and freeze the attention parameters once the maps stabilize. On the Horse↔Zebra benchmark, our method improves Kernel Inception Distance (KID; lower is better) compared with a vanilla CycleGAN baseline and yields visibly cleaner backgrounds with reduced texture bleeding.

Keywords: Unpaired translation; CycleGAN; Frozen attention; Mask-guided discriminator; Foreground-aware translation

1 Introduction

Unpaired image-to-image translation aims to learn mappings between two visual domains using unpaired samples [1]. CycleGAN-style methods enforce cycle consistency to regularize this ill-posed problem, enabling compelling translations without paired supervision. However, for many practical tasks the desired translation is not global: a small set of semantically meaningful regions should change, while the remainder should be preserved. In horse→zebra, for example, stripes should appear on the animal but not on grass, fences, or sky. In our baseline CycleGAN runs, the paper frequently observes 'texture bleeding': target-domain textures appear in the background or along object boundaries, even when the global structure is preserved. This paper targets that specific failure mode.

A common way to inject region awareness is to introduce an attention mechanism [2] that highlights discriminative regions and suppresses irrelevant ones. Prior unpaired translation models have used spatial attention in the generator or attention-guided discriminators [3] to focus the adversarial signal. These approaches can be effective,

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_100

but they introduce two challenges for a small, reproducible extension of CycleGAN: (1) additional attention losses or large attention subnetworks can complicate training [4], and (2) attention maps may remain unstable, causing late-stage shifts that degrade previously good translations. Our goal is a minimal modification that improves foreground focus without relying on external masks or heavy architectural changes.

The paper proposes Frozen-Attention CycleGAN (FA-CycleGAN): a CycleGAN variant that learns a soft attention mask from internal generator features, uses it to mask discriminator inputs and adversarial losses, and then freezes the attention parameters once the attention maps converge. Freezing converts attention into a fixed, learned spatial prior, reducing training instability while keeping the rest of the network trainable.

This paper introduces a lightweight attention head that predicts a soft foreground mask from the generator's deep feature space, providing an explicit mechanism for spatially selective adversarial learning. This attention signal is used to guide discriminator training by modulating both its input and loss, thereby concentrating on adversarial supervision on semantically relevant regions and improving gradient efficiency. From a stability perspective, the paper identifies late-stage attention drift as a source of training degradation and propose a convergence-based freezing criterion that fixes the attention head once its predictions stabilize, leading to more consistent adversarial dynamics. Extensive experiments on the Horse↔Zebra task demonstrate improved translation quality measured by KID, with ablation results confirming the stabilizing role of each proposed component.

2 Problem Definition

The paper considers two unpaired image domains. First is text notation, let X be the source domain and Y the target domain. The paper observes samples $x \sim p_X(x)$ (horses) and $y \sim p_Y(y)$(zebras). The paper learns generators$G: X \to Y \; and \; F: Y \to X$, with discriminators$D_Y$ and D_X . The standard CycleGAN objective is

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda_{cyc}\mathcal{L}_{cyc}(G, F) + \lambda_{id}\mathcal{L}_{id}(G, F) \quad (1)$$

Here $\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x\sim p_X(x)}[\| F(G(x)) - x \|_1] + \mathbb{E}_{y\sim p_Y(y)}[\| G(F(y)) - y \|_1]$. $\odot$ denotes element-wise (Hadamard) product. X represents source domain (e.g., horses), with samples $x \sim p_X(x)$. Y represents target domain (e.g., zebras), with samples $y \sim p_Y(y)$

Unpaired translation aims to learn two mappings:

$$G_{X\to Y}: X \to Y, G_{Y\to X}: Y \to X, \quad (2)$$

such that $G_{X\to Y}(x)$ is indistinguishable from real images in Y, and $G_{Y\to X}(y)$ is indistinguishable from real images in X, while preserving content.

CycleGAN trains with (i) adversarial losses and (ii) a cycle-consistency constraint:

$$G_{Y\to X}\big(G_{X\to Y}(x)\big) \approx x, G_{X\to Y}\big(G_{Y\to X}(y)\big) \approx y \quad (3)$$

An optional identity term encourages minimal changes when the input is already in the target style:

$$G_{X\to Y}(y) \approx y, G_{Y\to X}(x) \approx x \quad (4)$$

A key limitation for spatially localised tasks is that the adversarial objective does not explicitly distinguish between foreground and background. As a result, the model may expend capacity modifying regions that do not require translation, producing texture leakage and spatial artefacts. This motivates adding spatial guidance while keeping the overall CycleGAN training objective unchanged.

3 Proposed Method: Attention-Enhanced CycleGAN

3.1 Architecture Overview

Starting from a standard CycleGAN with two generators $(G_{X\to Y}, G_{Y\to X})$ and two discriminators (D_X, D_Y). Our method introduces two minimal modifications. The paper first augments each generator with a lightweight spatial attention branch, enabling it to predict not only a translated image but also a soft attention map that identifies regions where stylistic transformation is necessary, thereby explicitly modeling the spatial extent of translation. To further improve adversarial supervision, the paper then adopts a masked discriminator training strategy once the attention maps have stabilized: the attention branch is frozen and the resulting binarized masks are used to gate discriminator inputs, restricting adversarial feedback to semantically relevant regions and reducing interference from unchanged background areas. Importantly, this approach does not rely on ground-truth segmentation [5,6], bounding boxes, or any additional supervision, and is deliberately designed to remain lightweight to preserve the computational efficiency and training stability of the original CycleGAN baseline.

3.2 Spatial Attention Module

A shallow attention branch is attached to the generator's last residual block output (feature map of size $C \times H/4 \times W/4$). The branch consists of: Conv 3×3, $C \to 64$, ReLU, Conv 3×3, $64 \to 32$, ReLU.

For 256×256inputs, the shared generator feature map is 64 channels at 64×64.

The attention head is

$$\mathrm{Conv}_{3\times3}(64 \to 64) + \mathrm{ReLU}, \mathrm{Conv}_{3\times3}(64 \to 32) + \mathrm{ReLU}, \mathrm{Conv}_{1\times1}(32 \to 1) + \mathrm{Sigmoid}, \quad (5)$$

followed by bilinear upsampling to $256 \times 256$to obtain

$$A \in [0,1]^{256\times256} \quad (6)$$

The paper broadcast Aacross RGB channels.

$$\hat{y}_{\mathrm{raw}} = G(x) \quad (7)$$

The final output is

$$\hat{y} = A \odot \hat{y}_{\mathrm{raw}} + (1 - A) \odot x \quad (8)$$

The final generator output is:

$$\hat{y} = A \odot \tilde{y} + (1 - A) \odot x \quad (9)$$

This encourages the generator to concentrate on changes on high-attention regions while preserving background appearance. To avoid degenerate attention maps (all-zeros or all-ones), the paper optionally adds a weak total-variation penalty:

$$\mathcal{L}_{tv}(A) = \sum_{i,j} | A_{i+1,j} - A_{i,j} | + | A_{i,j+1} - A_{i,j} | \quad (10)$$

Weighted by $\lambda_{tv} = 0.01$. In practice, the paper observed smoother and more compact masks without noticeably harming translation fidelity.

3.3 Masked Discriminator Training

The paper declares attention stable if

$$\Delta A(e) = \mathbb{E}_{x \in B} [\| A_e(x) - A_{e-1}(x) \|_1]$$
$$\Delta A(e) < 0.02 \quad (11)$$

For 5 consecutive epochs; the paper freezes the attention head at the first epoch satisfying this condition (epoch 25 in our runs). The paper uses $\tau = 0.5$to form

$$M = \mathbf{1}[A \geq \tau] \quad (12)$$

The discriminator input is

$$M \odot \text{image} \quad (13)$$

If the mask area is <1%of pixels, the paper falls back to full-image discriminator input for that batch. By monitoring the attention map change across epochs, using mean absolute difference:

$$\Delta_A(e) = \mathbb{E}_{x \sim p_X}[\| A_e(x) - A_{e-1}(x) \|_1] \quad (14)$$

The paper consider attention "stable" when $\Delta_A(e) < 0.02$for 5 consecutive epochs (illustrative), typically around epoch 20–30 on Horse↔Zebra. Setting $\tau = 0.5$ to obtain a binary mask $M = \mathbb{1}[A \geq \tau]$. The paper then feed masked images to the discriminator:

Hyperparameters LSGAN is used. The paper sets

$$\lambda_{\text{cyc}} = 10 \text{and} \lambda_{\text{id}} = 5 \quad (15)$$

The optimiser is Adam with

$$\text{lr} = 2 \times 10^{-4}, (\beta_1, \beta_2) = (0.5, 0.999) \quad (16)$$

and batch size $= 1$. The paper trains for 200 epochs (constant learning rate for epochs 1–100, then linear decay to 0 for epochs 101–200). The paper enables masked discriminator training after freezing attention (from epoch 26 onward). The paper applies a total-variation regulariser on $A$with weight

$$\lambda_{\text{TV}} = 0.01$$
$$y_{\text{mask}} = M \odot \hat{y} \quad (17)$$

This encourages discriminators to judge realism primarily on regions intended for translation, reducing background-driven artefacts.

3.4 Objective Function and Training Strategy

It is capable to use LSGAN adversarial losses, cycle-consistency $\lambda_{cyc} = 10$, identity $\lambda_{id} = 5$. Optimiser: Adam, $\text{lr} = 2 \times 10^{-4}$, $\beta_1 = 0.5, \beta_2 = 0.999$.

When epoch 1–25, it uses full-image discriminator and attention learns freely. When epoch 26–200, it uses freeze attention head and masked discriminator training enabled

4 Experiments

4.1 Dataset and Evaluation Metric

Experiments are conducted on the Horse↔Zebra unpaired image-to-image translation benchmark. The source domain X contains horse images and the target domain Y contains zebra images. The paper uses TrainA 1,000 images and TrainB 1,300 images for training, and TestA 120 images and TestB 140 images for testing.

All images are trained with 256×256crops. During training, the paper resizes the shorter side to 286 pixels while keeping the aspect ratio, randomly crop to 256×256, and apply horizontal flipping with probability 0.5. Inputs are scaled to the range $[-1,1]$to match $\tanh$outputs. At test time, the paper resizes to $256 \times 256$with no random crop and no flipping.

Kernel Inception Distance (KID) is reported in units of $\times 10^{-3}$, where lower values indicate that generated images are closer to the target distribution in Inception feature space. The paper computes KID on all test translations in each direction.
Adam with

$$\mathrm{lr} = 2 \times 10^{-4}, (\beta_1, \beta_2) = (0.5, 0.999) \tag{18}$$

Batch size $= 1$, trained for 200 epochs. The learning-rate schedule is constant for the first 100 epochs, then linearly decayed to 0. The paper uses no weight decay, no gradient clipping, and no mixed precision.

$$\lambda_{\mathrm{cyc}} = 10, \lambda_{\mathrm{id}} = 5 \tag{19}$$

The attention total-variation regularisation weight is

$$\lambda_{\mathrm{TV}} = 0.01 \tag{20}$$

The stabilisation threshold is $\Delta A < 0.02$for 5 consecutive epochs; the attention head is frozen at epoch 25. The mask threshold is $\tau = 0.5$, with a tiny-mask fallback when the mask area is $<1\%$.

4.2 Baselines and Implementation Details

The baseline is the original CycleGAN trained under the same optimisation, schedule, and preprocessing.

The paper uses a ResNet-9 generator with three downsampling stages, nine residual blocks, and two upsampling stages. The discriminator is a fully convolutional 70×70 PatchGAN. InstanceNorm is used in both generator and discriminator.

The spatial attention branch is attached after the last residual block, at a 64-channel 64×64 feature map for 256×256 inputs. The branch is Conv 3×3 from 64 to 64 with ReLU, then Conv 3×3 from 64 to 32 with ReLU, then Conv 1×1 from 32 to 1 with Sigmoid, followed by bilinear upsampling to 256×256. The attention map A lies in [0,1] and is shared across RGB channels

Training uses Adam with learning rate 2×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$, batch size 1, for 200 epochs. The learning rate is held constant for the first 100 epochs and linearly decayed to zero over the remaining 100 epochs. Weight decay is 0; gradient clipping is not used; mixed precision is disabled.

LSGAN is used for adversarial learning. The cycle-consistency weight is $\lambda_{\mathrm{cyc}} = 10$, and the identity weight is $\lambda_{\mathrm{id}} = 5$. A weak total-variation regulariser is applied to the attention map with weight 0.01.

Several typical failure modes are observed in practice. First, occlusion and motion blur may lead to fragmented attention masks, which in turn result in spatially inconsistent or patchy stylisation. Second, when multiple foreground objects appear in close proximity, they can be mistakenly merged into a single attention region [7], causing over-coverage and unintended stylisation of nearby background areas. In addition, very small or thin structures such as ears, legs, or tails are sometimes insufficiently attended to, producing incomplete or discontinuous stripe patterns. Finally, highly textured backgrounds (e.g. trees or fences) may attract spurious attention responses, introducing faint stylisation artefacts in non-target regions.

To alleviate these issues, several practical strategies can be adopted. Masking is applied only after the attention variation ΔA has stabilised, which helps reduce early-stage noise. The threshold τ can be tuned (e.g. within the range 0.4–0.6), and a sensitivity analysis may be reported to demonstrate robustness. The tiny-mask fallback is retained to preserve thin or small structures. In addition, a weak sparsity or area regulariser can be optionally introduced on the attention map A to discourage excessive foreground expansion. Finally, replacing hard thresholding with soft masking—by directly using A as a weighting map—may further reduce brittleness and improve continuity in challenging cases.

The paper monitor attention stability using $\Delta A(e)$, defined as the expected ℓ_1 change of the attention map between epoch e and $e-1$, computed over a fixed mini-batch of 16 images per domain. Attention is considered stable when

$$\Delta A(e) < 0.02 \tag{21}$$

for 5 consecutive epochs, and the paper freezes the attention branch at epoch 25, the first epoch satisfying the criterion. Masking uses threshold $\tau = 0.5$ to form a binary mask

$$M = \mathbf{1}[A \geq \tau] \tag{22}$$

and the discriminator receives

$$M \odot \text{image} \tag{23}$$

If the mask area is smaller than 1% of pixels, the paper fall back to full-image discriminator input for that batch.

Kernel Inception Distance (KID) is computed using TorchMetrics KID with Inception-V3 pooled 2048-dimensional features. For Horse→Zebra the paper generates 120 samples, and for Zebra→Horse the paper generates 140 samples, matching the test set sizes. KID is estimated with subset_size $= 50$ and num_subsets $= 100$. The paper runs three independent trainings with seeds 42, 43, and 44, and report mean ±standard deviation across runs, while within-run subset standard deviation is recorded separately.

4.3 Quantitative Results

Table 1 reports KID for both translation directions. Our method achieves lower KID than the CycleGAN baseline in both directions.

Table 1. KID scores ($\times 10^{-3}$) for Horse↔Zebra (mean ± std)

Model	Z→H	H→Z
CycleGAN	11.44 ± 0.38	10.25 ± 0.25
Ours	8.87 ± 0.26	6.93 ± 0.27

Qualitative comparisons are performed using the same test inputs for CycleGAN and the proposed approach. The baseline CycleGAN may introduce unintended texture changes in regions that do not require translation, and stripe-like patterns can bleed outside object boundaries. With attention-guided composition and masked discriminator training, edits are more concentrated on the foreground animal while background appearance is more often preserved, producing improved spatial coherence.

the paper also observes characteristic failure cases that motivate stabilisation and fallback rules. Under-coverage can occur on thin structures such as legs or tails, causing incomplete stylisation. Over-coverage can occur in cluttered scenes, where the attention expands into nearby background, producing faint artefacts. When attention is unstable early in training, enabling masking too early can make discriminator feedback inconsistent; freezing at epoch 25 under the defined stability criterion mitigates this, and the tiny-mask fallback prevents degenerate masked inputs.

4.4 Ablation Study

An ablation study is conducted to isolate the contributions of key components. Table 2 shows that removing cycle consistency substantially degrades KID, while removing either attention or masked discriminator training reduces performance relative to the full model, indicating that the two mechanisms contribute complementary gains.

Table 2. Ablation results (KID $\times 10^{-3}$, mean ± std)

Model	Z→H	H→Z
Without Cycle Consistency	25.31 ± 0.72	22.84 ± 0.65
CycleGAN Baseline	11.44 ± 0.38	10.25 ± 0.25
Without Attention	10.92 ± 0.41	9.58 ± 0.33
Without Masked Disc.	9.73 ± 0.35	8.41 ± 0.29
Full Model (Ours)	8.87 ± 0.26	6.93 ± 0.27

Masked discriminator training improves performance over the baseline by focusing adversarial feedback on regions indicated by the attention map, while the attention branch further helps localise edits and reduce background corruption. Removing cycle consistency causes the largest drop, consistent with the role of cycle constraints in preserving content structure in unpaired translation.

5 Discussion

These results suggest that spatial guidance can meaningfully improve CycleGAN in scenarios where the domain shift is largely localised. The attention branch learns to highlight foreground regions that require texture changes, while masked discriminator training concentrates adversarial pressure on those regions, reducing background corruption.

A practical takeaway is that substantial gains are achievable without replacing CycleGAN's backbone or introducing heavy components such as transformers [8]. This can be valuable when training stability and computational efficiency matter.

Limitations. Our approach depends on the attention map becoming stable; if attention remains noisy or systematically mis-localises the foreground, the binary mask can suppress useful adversarial gradients. The threshold τis also a sensitive hyperparameter: a high threshold may under-cover the object, while a low threshold may reintroduce background pressure.

Further limitations [9,10] (concrete): The method assumes that the desired domain shift is spatially localised; for tasks where style should change globally, masking can hurt realism. Stability detection depends on a fixed reference batch B; if B is not representative, freezing may occur too early/late. Hard thresholding introduces sensitivity to τ and can create sharp mask boundaries; soft masking may be smoother. Finally, when attention is systematically biased (e.g., cluttered scenes), concentrating adversarial gradients may amplify local artefacts rather than remove them.

Future work. Promising directions include replacing hard thresholding with a differentiable masking mechanism, adding weak regularisation to prevent attention collapse, and testing across datasets where the foreground/background separation is less clear to assess robustness [11,12].

6 Conclusion

The paper presented a minimal, attention-enhanced extension of CycleGAN for unpaired image-to-image translation. By adding a lightweight generator attention branch and applying masked discriminator training after attention stabilisation, the method reduces unnecessary background changes and improves spatial coherence. On Horse↔Zebra, the paper achieves lower KID than the CycleGAN baseline in both directions, and ablation studies confirm that attention and masked adversarial supervision contribute complementary improvements.

References

1. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: ICCV (2017)
2. Lin, Y., Wang, Y., Li, Y., Gao, Y., Wang, Z., Khan, L.: Attention-Based Spatial Guidance for Image-to-Image Translation. In: WACV (2021)
3. Tang, H., et al.: AttentionGAN: Unpaired Image-to-Image Translation Using Attention-Guided Generative Adversarial Networks. IEEE Transactions on Neural Networks and Learning Systems (2021)
4. Zhao, J., et al.: Lightweight Domain-Attention GAN for Unpaired Image-to-Image Translation. Neurocomputing (2022)
5. Shi, D., et al.: Depth-Aware Unpaired Image-to-Image Translation for Adverse Weather Perception. Frontiers in Neurorobotics (2025)
6. Zhao, J., et al.: TVA-GAN: Attention-Guided Generative Adversarial Network for Thermal-to-Visible Image Transformations. TechRxiv (2023)

7. Tu, H., et al.: Multiscale Attention-GAN for Unsupervised Image-to-Image Translation. Applied Intelligence (2024)
8. Torbunov, D., et al.: UVCGAN: UNet Vision Transformer Cycle-Consistent GAN for Unpaired Image-to-Image Translation. In: WACV (2023)
9. Tu, H., et al.: Unpaired Image-to-Image Translation with Diffusion Models. Mathematics 12(20), 3178 (2024)
10. Su, X., Song, J., Meng, C., Ermon, S.: Dual Diffusion Implicit Bridges for Image-to-Image Translation. arXiv:2203.08382 (2022)
11. Xie, S., et al.: Unpaired Image-to-Image Translation with Density-Consistent Diffusion. In: NeurIPS (2022)
12. Kim, G., et al.: Diffusion Model Compression for Image-to-Image Translation. In: ACCV (2024)