1 Methodology

1.1 Overview

We propose a biologically-preserving unsupervised image-to-image translation framework tailored for the underwater domain, where visibility varies across turbidity, depth, and lighting. Our framework builds upon CycleGAN, but introduces critical novel mechanisms to ensure species identity and perceptual realism are preserved during style transfer between underwater domains.

Our central innovation is the **Domain-Aware Adaptive Identity Loss (DA^{2}IL)**, which injects anatomical awareness into the translation process, guided by saliency and high-frequency structure priors.

1.2 Architecture Design

The core of our framework is a dual-generator, dual-discriminator architecture inspired by CycleGAN [**?**], with substantial modifications to enhance semantic consistency and anatomical fidelity in the domain of underwater species transfer. Our modifications are designed specifically to address challenges such as poor visibility, domain gap, occlusion, and the need for biologically plausible outputs.

1.2.1 Generator Network

The generator is structured as an encoder–bottleneck–decoder architecture with residual connections and skip pathways. Its objective is to learn a mapping $G : X \to Y$, where X and Y are image domains (e.g., underwater A and underwater B). The inverse mapping $F : Y \to X$ is also learned simultaneously.

Each generator G (and symmetrically F) is composed of the following:

1) Initial Convolutional Layer (Feature Extraction) We begin with a 7×7 convolutional layer with stride 1, reflection padding, and instance normalization:

$$\mathbf{f}_1 = ReLU(IN(Conv_{7\times7}(\mathbf{x})))$$

This wide receptive field helps in capturing global texture and color context, which is crucial for learning domain-specific styles such as lighting gradients and water color shifts.

2) Downsampling Layers (Spatial Compression) We employ two 3×3 convolutional layers with stride 2 to halve the spatial resolution at each step, increasing the depth of features:

$$\mathbf{f}_{i+1} = ReLU(IN(Conv^{s=2}_{3\times3}(\mathbf{f}_i)))$$

These layers capture increasingly abstract features such as object shape, inter-region contrast, and large-scale pose information.

3) Residual Bottleneck (Feature Transformation) The compressed features are passed through six residual blocks. Each block contains two 3 × 3 convolutional layers, instance normalization, and ReLU activation. The residual structure enables stable learning of identity-preserving transformations:

$$\mathbf{f}_{res} = \mathbf{f}_i + \mathcal{F}(\mathbf{f}_i, \theta)$$

where $\mathcal{F}$ denotes the residual function (a stack of two conv–norm–ReLU layers). These blocks model transformations such as texture morphing, lighting harmonization, and detail refinement, while retaining spatial identity.

4) Upsampling Layers (Image Reconstruction) The decoder mirrors the encoder, using transposed convolutions (a.k.a. deconvolution) with stride 2 to gradually restore the original image size:

$$\mathbf{f}_{i-1} = ReLU(IN(ConvTranspose_{3\times3}^{s=2}(\mathbf{f}_i)))$$

These layers perform semantic upsampling, re-projecting the abstracted features into spatially accurate image content. They allow the model to preserve biologically important details such as fin edge geometry or eye shape.

5) Final Output Layer The final output is produced via a 7 × 7 convolution followed by a tanh activation to bring pixel values into $[-1, 1]$:

$$\hat{\mathbf{y}} = \tanh(Conv_{7\times7}(\mathbf{f}_{out}))$$

This ensures stable gradient flow and naturalistic appearance.

6) Skip Connections To retain low-level spatial information lost during downsampling, we introduce skip connections (as in U-Net [**?**]) from encoder layers to corresponding decoder layers. This preserves boundaries and fine-scale anatomical features crucial for underwater species.

1.2.2 Discriminator Network

The discriminator D_X is implemented as a 70×70 PatchGAN [**?**], which classifies each overlapping patch of the input image as real or fake. This form of local adversarial training is well-suited to texture-rich underwater imagery.

Mathematically, the PatchGAN discriminator operates as:

$$D_X(\mathbf{x}) \in R^{H' \times W'} \rightarrow patch - wiseprobabilities$$

Each element of the output grid corresponds to a 70 × 70 receptive field in the input image. This encourages the generator to produce realistic local structure (e.g., fish scales, water caustics), instead of just matching global image statistics.

Why PatchGAN? Global discriminators often overlook microstructures critical for marine biology (e.g., fin rays, scale granularity). PatchGAN enforces fidelity at the microtexture level, which complements our biologically motivated loss (DA^{2}IL) to maximize realism.

1.2.3 Summary of Parameters

Table 1: Network Design Summary

Component	**Layers**	**Key Functions**
Generator (G, F)	Conv-7×7, 2 Downsampling, 6 Residual Blocks, 2 Upsampling, Conv-7 × 7	Performs bidirectional domain translation using feature encoding, transformation, and reconstruction
Discriminator (D_X, D_Y)	5 Convolutional blocks (PatchGAN)	Performs patch-wise real/fake classification to enforce local realism
Skip Connections	Encoder–Decoder bridges	Preserves low-level spatial and anatomical features across layers

1.3 Design Rationale

Our architecture is designed around three biological priors:

- **Fine Edge Retention**: Critical for taxonomic identification in fish; enforced via skip connections and small receptive field convolutions.
- **Domain-Specific Texture Modeling**: Managed via residual blocks and PatchGAN discriminator.
- **Local-Global Style Balancing**: Achieved by combining instance normalization (style-invariant) and residual transformations (style-adaptive).

1.4 Loss Functions

The optimization objective of our proposed CycleGAN-based fish style transfer framework combines adversarial, cycle consistency, and a novel domain-aware identity loss (DA^{2}IL). The complete loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{DA^2IL}\mathcal{L}_{DA^2IL}$$

where λ_{cyc} and λ_{DA^2IL} are hyperparameters controlling the relative importance of each term.

1.4.1 Adversarial Loss

The adversarial loss ensures that the translated image $G(x)$ from domain $\mathcal{X}$ is indistinguishable from real images y in domain $\mathcal{Y}$. This is defined using a Least Squares GAN (LSGAN) objective for stability and better gradients:

$$\mathcal{L}_{GAN}(G, D_Y) = E_{y \sim p_{data}(y)}[(D_Y(y) - 1)^2] + E_{x \sim p_{data}(x)}[D_Y(G(x))^2]$$

A similar loss is used for F and D_X, forming the symmetric adversarial structure.

Motivation: This loss drives the generator to produce outputs that reside in the target domain distribution, enabling plausible water-style translation. Using LSGAN improves convergence and avoids vanishing gradients typical in standard GANs.

1.4.2 Cycle Consistency Loss

To prevent mode collapse and ensure meaningful mapping, we enforce a bijective cycle consistency constraint between domains $\mathcal{X}$ and $\mathcal{Y}$:

$$\mathcal{L}_{cyc}(G, F) = E_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1] + E_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]$$

Motivation: This loss encourages the network to learn mappings $G : \mathcal{X} \to \mathcal{Y}$ and $F : \mathcal{Y} \to \mathcal{X}$ that are approximately inverse to each other. It ensures semantic and structural preservation—vital for retaining fish morphology across style domains.

1.4.3 Domain-Aware Anatomical Identity Loss (DA^{2}IL)

We propose a novel loss to explicitly preserve anatomical features of the fish, such as body shape, fin orientation, and eye location, which are crucial for ecological monitoring applications. The DA^{2}IL leverages spatial attention over key body regions and penalizes distortions in the generator's output.

Let $\mathcal{A}(I)$ denote the spatial attention map over anatomical keypoints (e.g., derived from pose estimators or Sobel gradients), then:

$$\mathcal{L}_{DA^2IL}(x, G(x)) = \sum_i \mathcal{A}_i(x) \cdot \|x_i - G(x)_i\|_1$$

Motivation: Unlike traditional identity loss $\|G(y) - y\|_1$, our DA^{2}IL dynamically focuses on biologically relevant regions during training, reducing semantic drift during style transformation and enhancing species-level fidelity.

1.4.4 Final Objective

The complete optimization objective is then:

$$G^*, F^* = \arg \min_{G,F} \max_{D_X, D_Y} \mathcal{L}_{GAN}(G, D_Y) + \mathcal{L}_{GAN}(F, D_X) + \lambda_{cyc}\mathcal{L}_{cyc}(G, F) + \lambda_{DA^2IL}\mathcal{L}_{DA^2IL}$$

We empirically set $\lambda_{cyc} = 10$ and $\lambda_{DA^2IL} = 5$ after performing extensive ablation studies across multiple fish datasets and water domain variants.

1.4.5 Proposed DA^{2}IL: Domain-Aware Adaptive Identity Loss

We introduce a novel biologically motivated loss function that adaptively emphasizes fish-relevant regions such as eyes, fins, and silhouettes — areas crucial for identity recognition in marine biology.

Let $y \in \mathcal{Y}$ be the source image, and $G(y)$ be the translated image. Define a saliency mask $\mathcal{M}(y)$ based on combined edge strength and local contrast (computed via Laplacian of Gaussian + histogram equalization). Then:

$$\mathcal{L}_{DA^2IL} = E_{y \sim p(y)} \left[\| \mathcal{M}(y) \odot (G(y) - y) \|_1 \right]$$

This formulation adaptively penalizes changes in semantically important areas. $\mathcal{M}(y)$ is computed per image, making it robust to different fish shapes and occlusion patterns.

1.5 Training Details

- **Optimizer:** Adam ($\alpha = 2 \times 10^{-4}, \beta_1 = 0.5, \beta_2 = 0.999$)
- **Scheduler:** Linear decay after 100 epochs
- **Batch Size:** 1 (stabilizes high-resolution GAN training)
- **Epochs:** 200
- **Image Size:** 256×256
- **Saliency Implementation:** Laplacian edge maps + histogram equalization $\rightarrow$ Gaussian blur $\rightarrow$ sigmoid threshold

1.6 Evaluation Metrics

We assess performance using both perceptual and signal-level metrics.

- **Fréchet Inception Distance (FID):** Measures distributional distance between generated and real images via Inception features. Lower is better. Our score: **58.7**.
- **Structural Similarity Index (SSIM):** Measures luminance, contrast, and structure similarity. Range [0,1]. Higher is better. Ours: **0.82**.
- **Peak Signal-to-Noise Ratio (PSNR):** Measures pixel-wise fidelity in decibels. Higher indicates better quality. Ours: **25.4 dB**.
- **Learned Perceptual Image Patch Similarity (LPIPS):** Measures feature distance using a deep network (AlexNet/VGG). Lower indicates better perceptual similarity. Ours: **0.184**.

These metrics collectively ensure both low-level and perceptual fidelity across domains.

2 Ablation Study

To quantify the individual contributions of our proposed components, we perform controlled experiments under multiple variants.

2.1 Experimental Variants

- **V1 – CycleGAN Baseline:** Uses adversarial + cycle consistency only.
- **V2 – +Identity Loss:** Adds traditional identity loss: $\|G(y) - y\|_1$
- **V3 – +Saliency Mask Only:** Replaces identity loss with saliency-weighted $\mathcal{L}_{DA^2IL}$ using fixed $\mathcal{M}(y)$
- **V4 – +DA^{2}IL (Ours):** Full model with adaptive $\mathcal{M}(y)$ computed dynamically per image

2.2 Quantitative Results

Table 2: Ablation Results on Underwater Fish Translation Dataset

Variant	**FID** ↓	**SSIM** ↑	**PSNR** ↑	**LPIPS** ↓
V1: CycleGAN	73.2	0.71	21.3	0.241
V2: + Identity	65.8	0.77	23.0	0.210
V3: + Saliency Mask	61.2	0.79	24.1	0.196
V4: + DA^{2}IL (Ours)	**58.7**	**0.82**	**25.4**	**0.184**

2.3 Qualitative Observations

Compared to V1 and V2, V4 consistently preserves fish fin structure, eye sharpness, and boundary contours. V2 improves low-level preservation but fails to emphasize salient anatomy. V3 validates the importance of region weighting, while V4 shows that adaptive masks outperform static heuristics.

2.4 Discussion

The introduction of DA^{2}IL significantly reduces semantic drift, a common GAN failure case where fish are misaligned or biologically implausible. This is evident in LPIPS and FID improvements. The gains are especially prominent in low-visibility scenarios (e.g., domain B), where structural integrity is harder to maintain. These results support the hypothesis that semantically guided regularization is crucial in biological domain transfer tasks.