Origami Model using Neural Style Transfer and CycleGAN

Yuina Iseki [1] Changju Yuan [2] Antra Nakhasi [3]

[1]Department of Computer Science [2]Department of Civil and Environmental Engineering [3]Department of Management Science and Engineering, Stanford University

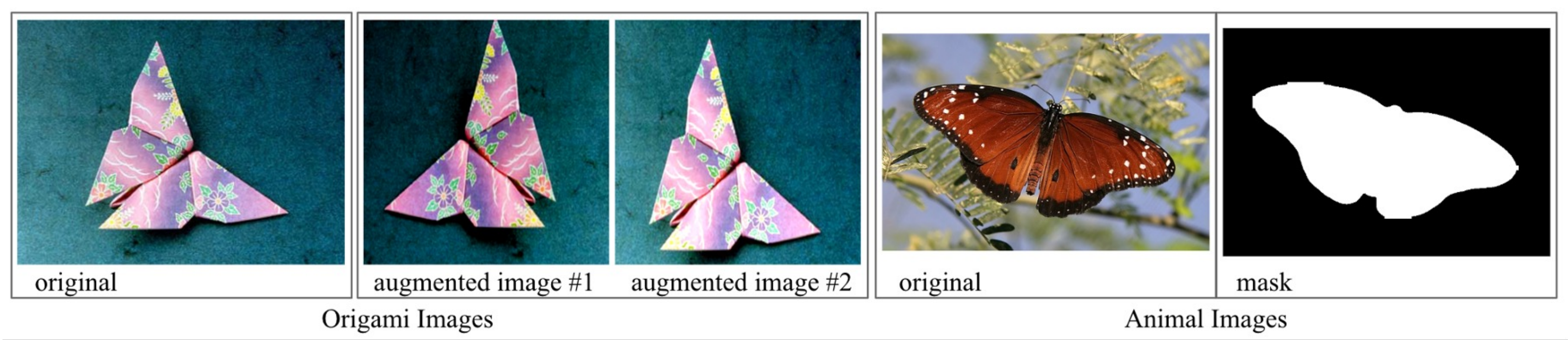
Introduction

We transform animal photos into origami-style images to explore bio-inspired foldable structure design. We implemented three models: Vanilla NST, feed-forward NST and CycleGAN.

Key Findings: Vanilla NST achieved the best origami-like results on position-aligned, segmented butterfly images, while feed-forward NST remained stable but texture-limited, and CycleGAN struggled without geometric guidance. Combining style transfer with spatial alignment and geometric priors is crucial for faithful origami modeling.

Data

Figure 1. Augmented origami images (left) and an animal image with its segmentation mask (right)



- Dataset composition: ~61,200 unpaired images (56,814 animals from ImageNet and 4,387 origami from Kaggle); 80/10/10 train/val/test split
- Normalization: ImageNet mean–std normalization for NST, and $[-1, 1]$ scaling for CycleGAN.
- Augmentation: Flips, rotations, random crops, color jitter, and mild perspective distortion.
- Segmentation: YOLOv8 to isolate animal subjects.

Vanilla Neural Style Transfer (NST)

Optimizes pixel values with VGG-19 content/style features from multiple convolutional layers.

Loss Function:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{content}} + \beta\mathcal{L}_{\text{style}} + \gamma\mathcal{L}_{\text{tv}}$$

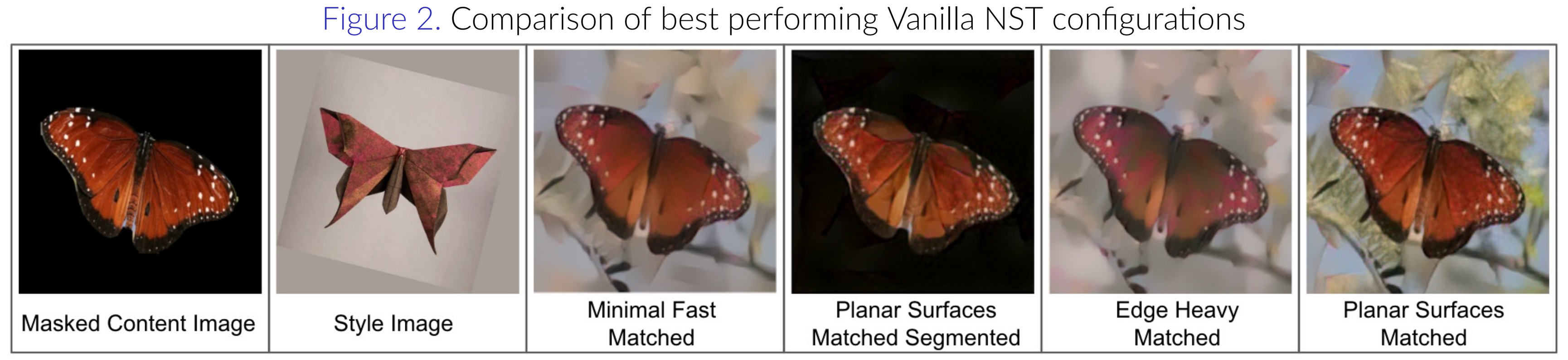
Optimization: Adam (lr = 0.003, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), $\gamma = 0$.

Table 1. Successful Style Layer Configurations.

Variant	Style Layers	Effect
planar_surfaces	conv3_1, conv4_1	Flat, paper-like surfaces
edge_heavy	conv1_1, conv2_1	Sharp folds, strong edges
geometric_emphasis	conv2_1–conv4_1	Angular, faceted structure

Results: Position-matched, segmented images significantly outperform unmatched inputs, with planar_surfaces achieving the highest fidelity (SSIM = 0.677, PSNR = 25.77 dB) by reducing background interference and ensuring spatial correspondence.Preprocessing via segmentation and spatial alignment is critical for preserving origami's planar geometry.

Figure 2. Comparison of best performing Vanilla NST configurations



Feed-Forward NST

Generator produces stylized output in one forward pass using perceptual VGG losses.

Architecture: 3 conv layers → 5 residual blocks → 2 upsampling layers → output conv

Residual block types:

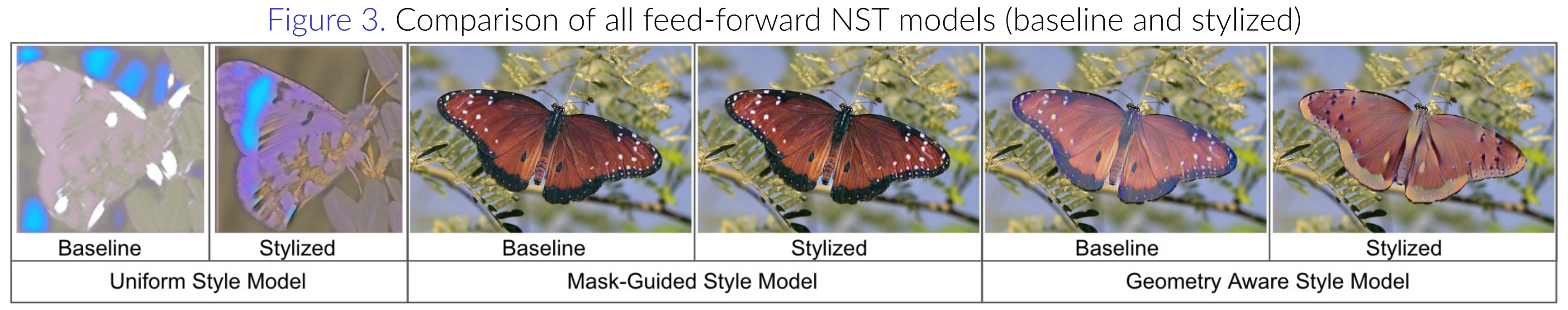
- *Baseline:* standard ResBlock (InstanceNorm + ReLU).
- *Stylized-residual:* adds 1×1 conv + sigmoid gating.

Table 2. Training Variants for Feed-Forward NST

Attribute	Uniform Style	Mask-Guided	Geometry-Aware
Input	RGB (3ch)	RGB + Mask (4ch)	RGB + Mask (4ch)
Style Layers	Gatys full set	Gatys full set (fixed Gram)	conv2_1–conv4_1
Style Weights	1.0–0.1	1.0–0.1	[1.5, 1.5, 1.0]
Learning Rate	10^{-3}	10^{-4}	10^{-4}
Batch Size	6	4	4
TV Weight	default	default	10^{-6}

Results: The Uniform Style Model produces washed-out outputs with weak structure, the Mask-Guided model preserves shape almost perfectly with the strongest metrics (SSIM ≈ 0.86), and the Geometry-Aware model introduces light angular, fold-like textures but sacrifices fidelity with lower SSIM and PSNR.

Figure 3. Comparison of all feed-forward NST models (baseline and stylized)



CycleGAN

Bidirectional mappings between animal and origami domains using paired generators $G : X \to Y$ and $F : Y \to X$, trained with adversarial, cycle-consistency, and identity constraints.

Architecture: Generators: encoder → 9 residual blocks → decoder (tanh). Discriminators: 70×70 receptive field with 4 convolutional layers (LeakyReLU 0.2).

Optimization: Adam (lr$_G$=2×10^{-4}, lr$_D$=1×10^{-4}), batch size 4.

Vanilla CycleGAN Objective:

$$\mathcal{L} = \mathcal{L}_{\text{GAN}}(G, D_Y) + \mathcal{L}_{\text{GAN}}(F, D_X) + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G, F) + \lambda_{\text{id}}\mathcal{L}_{\text{id}}(G, F), \quad \lambda_{\text{cyc}} = 10,\ \lambda_{\text{id}} = 5$$

Perceptual CycleGAN Objective:

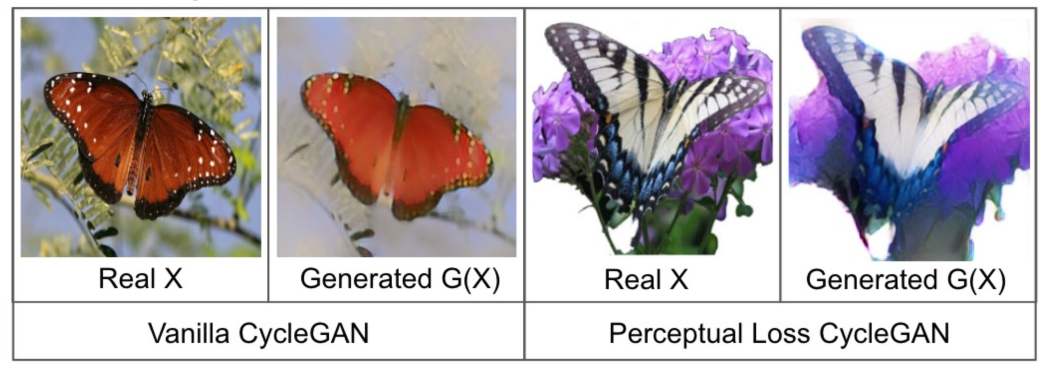
$$\mathcal{L} = \mathcal{L}_{\text{GAN}}(G, F, D_Y, D_X) + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}^{\text{total}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}}$$

L is split into L_G and L_F during the gradient process.

Results:

- Masked Vanilla CycleGAN: preserves high-level structure and segmentation boundaries (SSIM 0.49, PSNR 13.9 dB), though stylization remains weak (Gram distance 1.18×10^8).
- Perceptual CycleGAN: produces stronger artistic texture and color but sacrifices structural fidelity (SSIM 0.417, PSNR 9.93 dB) with larger Gram distance.
- Increasing stylistic emphasis consistently reduces geometric clarity, reflecting a trade-off between abstraction and reconstruction quality.

Figure 4. Comparison of both CycleGAN models



Discussion & Conclusion

- Vanilla NST achieves the best balance of geometry and texture, with PSNR = 23.93 and SSIM = 0.677, outperforming feed-forward NST in overall transformation quality.
- The strongest feed-forward result (Mask-Guided baseline) reaches SSIM = 0.864 but only PSNR = 17.89 - higher structural similarity does not translate to better texture accuracy.
- CycleGAN variants perform worse in geometric fidelity, with Masked Vanilla at PSNR = 13.90 and SSIM = 0.499, and Perceptual CycleGAN dropping further to SSIM = 0.417.
- Across all models, feed-forward NST and CycleGAN produce stronger stylization but fail to match Vanilla NST's combined preservation of shape and surface appearance.

Table 3. Comparison of all Model Variants

Model	Variant	PSNR↑	SSIM↑
Vanilla NST	planar_surfaces (matched)	23.93	0.677
	planar_surfaces (matched + segmented)	**25.77**	0.367
Feed-Forward NST	Mask-Guided (baseline)	17.89	**0.864**
	Geometry-Aware (baseline)	11.56	0.516
CycleGAN	Masked Vanilla	13.90	0.499
	Perceptual CycleGAN	9.93	0.417

Future Work

- Develop hybrid architectures that encode geometric constraints such as planarity losses and crease-aware edge detection.
- Collect a paired, pose-aligned dataset of animals and origami models to reduce the structured domain gap.
- Incorporate evaluation metrics that capture fold quality, angular alignment, and planar consistency beyond pixel similarity.

References

[1] Gatys, L. A., Ecker, A. S., Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

[2] Johnson, J., Alahi, A., Fei-Fei, L. (2016, September). Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision (pp. 694-711). Cham: Springer International Publishing.

[3] 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 (pp. 2223-2232).