IJCERT
Original Research Article

International Journal of Computer Engineering in Research Trends
e-ISSN: 2349-7084 | Volume 13, Issue No: 3, July–September 2026, Pp: 1–12

Low-Overhead Adaptive Multi-Scale Structure-Preserving Refinement for Pretrained Generative Photo Cartoonization

Kavali Manisha[1] **and A. Soujanya**[2]

[1]M.Tech Scholar, Department of Computer Science & Engineering, CVR College of Engineering, Hyderabad, Telangana, India; Email: Kavalimanisha47@gmail.com

[2]Assistant Professor, Department of Computer Science & Engineering, CVR College of Engineering, Hyderabad, Telangana, India; Email: soujanya052022@gmail.com

*Corresponding author: kavalimanisha47@gmail.com.

Abstract— Photo cartoonization transforms photographic content into a simplified artistic representation while attempting to preserve scene geometry, salient contours, and recognizable object structure. Pretrained generative models provide an efficient route to visually convincing stylization; however, local boundaries can be weakened during transformation, particularly in texture-rich scenes and regions containing thin structural details. This work presents a low-overhead adaptive multi-scale structure-preserving refinement framework that operates after pretrained cartoon generation without retraining the underlying model. Structural evidence is extracted from the original image using Gaussian filtering at three spatial scales, adaptive Canny edge estimation, confidence-map aggregation, and edge-density analysis. An adaptive fusion coefficient then regulates boundary reinforcement according to scene complexity. Local luminance anchoring is subsequently used to control excessive darkening, followed by luminance-channel contrast refinement in a perceptual color space. A reproducible pilot evaluation was conducted on 24 samples derived from 12 public images at a resolution of 512×512 pixels. The mean edge-preservation harmonic-mean score increased from 0.6240 to 0.6429, corresponding to a 3.02 percent relative improvement, with improvements observed in 18 of 24 samples. The complete refinement introduced approximately 0.93 percent additional runtime, while the structural similarity index measure changed marginally from 0.6230 to 0.6180. The findings indicate that targeted post-generation structural refinement can strengthen important boundaries at modest computational cost while retaining the visual character of the pretrained cartoonization backbone.

Keywords— photo cartoonization; generative image stylization; structure preservation; multi-scale edge analysis; adaptive fusion; training-free refinement

Article Info: Received Date: 12/06/2026; Revised: 19/08/2026; Accepted: 18/09/2026; Published: 30/09/2026

DOI: 10.22362/ijcert/2026/v13/i3/v13i3001

1. Introduction

Image stylization modifies the visual appearance of an image while retaining enough spatial and semantic information for the underlying scene to remain recognizable. Within this broader problem, photo cartoonization is particularly demanding because successful outputs must simplify texture, reorganize color distributions, emphasize characteristic boundaries, and maintain scene geometry. Recent analysis of the field shows that modern style-transfer research has progressed from optimization-based approaches toward feed-forward, adversarial, transformer-based, and large pretrained generative systems, while objective evaluation and content preservation remain unresolved concerns [1].

The tension between stylization and content retention is fundamental rather than incidental. Strong artistic transformation can weaken geometry, whereas overly restrictive content constraints can reduce the visible effect of the target style. Content self-supervision and style-contrastive learning have been used to address this conflict by encouraging geometric consistency while retaining style expressiveness [2]. This concern is directly relevant to cartoonization because degradation is often most visible along object boundaries, facial details, thin architectural components, and other high-frequency structures.

Style representation also affects the fidelity of the generated result. Global feature statistics are computationally convenient, yet they can overlook semantically distinct local style components. Soft multimodal transfer based on optimal transport has therefore been proposed to assign style components more flexibly while maintaining the spatial organization of the content image [3]. In a related direction, intrinsic-style distribution matching has been shown to reduce over-stylization and style leakage while preserving content structure in high-contrast images [4]. These findings suggest that locally controlled transformation is preferable to applying a uniform correction over the full image.

Controllability has consequently become an important design objective. Pixel-level control based on implicit neural representations enables the degree and location of stylization to be adjusted during test-time optimization [5]. Such flexibility is valuable, but iterative optimization is unnecessary when a pretrained cartoon generator already produces an acceptable artistic representation and only localized structural degradation requires correction.

Efficiency is equally important for practical systems. Fast style-transfer frameworks demonstrate that computationally economical design can improve the usability of artistic image transformation in interactive and resource-constrained settings [6]. Reusing learned generative representations can further

reduce the burden of retraining. For example, pretrained multi-path style-transfer networks have been fused to preserve facial identity and structure while producing diverse outputs [7]. This principle motivates the present work: the pretrained generator is kept fixed, and the proposed contribution is placed entirely after the generative stage.

Large pretrained models have broadened the range of attainable styles, but they do not remove the need for explicit structural reasoning. Style-prior methods built on large pretrained models continue to report trade-offs among content structure, generation quality, and inference cost [8]. Adversarial stylization has likewise introduced attention mechanisms and content-retention objectives to balance artistic appearance with semantic fidelity [9]. At the same time, non-neural patch matching based on optimal transport demonstrates that local image structure remains useful even when the primary stylization mechanism is already established [10].

1.1 Problem Motivation

Pretrained photo-cartoonization models offer a practical compromise between visual quality and deployment effort. Once a generator has learned a satisfactory artistic domain, a source photograph can be transformed without collecting a new paired dataset or optimizing a full adversarial model. Nevertheless, stylization may attenuate thin contours, merge adjacent regions, or suppress visually important boundaries. Retraining the generator with additional structural losses is possible, but it requires training data, hardware resources, hyperparameter selection, and a new validation cycle. More importantly, fine-tuning can alter the visual characteristics of a style that is already satisfactory.

A direct post-processing alternative is also problematic. Uniformly superimposing an edge map on the generated cartoon can create unnaturally dark contours, amplify fine texture, and disturb smooth color regions. Effective refinement therefore requires two decisions: which boundaries are sufficiently reliable to reinforce, and how strongly the correction should be applied for a given scene.

1.2 Research Gap and Objective

Published work has extensively examined feature-statistics matching, contrastive learning, optimal-transport style modeling, controllable stylization, pretrained style pathways, and large-model priors. Comparatively less attention has been given to a lightweight, training-free post-generation module that can be attached to an existing cartoonization model and selectively restore source-image structure without changing the generator parameters.

The objective of this study is therefore not to design another cartoon generator. Instead, it investigates whether reliable structural information in the source photograph can be extracted at multiple spatial scales and reintroduced into an already stylized image at low additional cost. The proposed *Adaptive Multi-Scale Structure-Preserving Fusion (AMSPF)* framework separates artistic transformation from structural correction and keeps the pretrained generator completely frozen.

1.3 Contributions

The main contributions are as follows. First, a two-branch framework is developed in which pretrained generative stylization and source-image structural analysis are separated. Second, multi-scale edge evidence is aggregated into a confidence representation so that reinforcement depends on structures that remain stable across smoothing scales. Third, scene edge density controls the fusion coefficient rather than relying on a fixed contour weight. Fourth, structural reinforcement is coupled with local luminance anchoring and controlled $L^*a^*b^*$-domain refinement to reduce unwanted darkening. Finally, the framework is evaluated in a reproducible paired pilot experiment using structural similarity, edge preservation, gradient consistency, color behavior, runtime, confidence intervals, and a non-parametric paired significance test.

The remainder of the paper is organized as follows. Section 2 reviews structure-preserving and controllable style-transfer research. Section 3 presents the proposed framework and algorithm. Section 4 describes the experimental protocol and metrics. Section 5 reports the quantitative, qualitative, and runtime results. Section 6 discusses implications, limitations, and threats to validity, and Section 7 concludes the study.

2. Related Work

Research on image stylization has evolved from texture-oriented transfer toward learned representations that explicitly separate content, style, geometry, and perceptual structure. The central challenge is no longer merely to produce visibly stylized imagery; contemporary systems must also preserve object identity, local contours, spatial consistency, and fine-scale information while remaining computationally practical. These concerns are especially important in photo cartoonization because strong simplification of texture and color can suppress thin boundaries or merge neighboring regions.

2.1 Structure Preservation in Image Stylization

Elad and Milanfar formulated style transfer from the perspective of texture synthesis and emphasized the need to keep selected content regions intact while synthesizing stylistic texture [11]. Cheng *et al.* later introduced an explicit structure representation containing global depth information and local edge details, demonstrating that object boundaries can be disrupted when style transfer relies only on recognition-oriented high-level features [12].

A broad review of neural style transfer identified content preservation, controllability, speed, style representation, and evaluation as recurring design problems [13]. Frequency-aware processing provides one route to protecting fine detail: wavelet decomposition has been used to recover high-frequency information and reduce structure distortion in stylized images [14]. Structure-guided refinement networks have likewise been integrated with style-transfer models to maintain hierarchical content and boundary information [15]. These approaches support the premise that explicit structural processing can complement the main stylization operation.

2.2 Localized and Domain-Specific Stylization

Localized preservation becomes especially important when the input contains faces or other semantically sensitive regions. Multiscale headshot transfer has shown that local contrast and illumination should be handled regionally rather than through a global appearance transform [16]. Domain-calibrated portrait translation further separates content adaptation, geometric expansion, and texture transfer so that few-shot portrait stylization can retain high-fidelity content [17].

Attention and saliency have also been used to preserve visually meaningful content during cross-domain stylization. A

saliency-regularized and attended generative adversarial network for Chinese ink-wash transfer exemplifies this strategy [18]. Multi-style transformation based on classifier-predicted style scores provides a complementary form of control by conditioning a generative model on smoothed attribute representations [19]. Both lines of work reinforce the value of adaptive rather than uniform transformation.

2.3 Spatial and Temporal Consistency

Consistency requirements extend beyond single images. Artistic video transfer has demonstrated that independent frame-wise stylization can produce instability under motion and occlusion, motivating explicit temporal constraints [20]. Consistency training and self-attention have subsequently been used to make arbitrary style transfer more robust to small perturbations [21]. A channelwise analysis of temporal consistency further shows that misalignment between content and style channels can contribute to flicker and structural instability in stylized video [22]. Although the present work concerns still images, these studies illustrate the broader principle that desirable appearance does not automatically guarantee stable structure.

2.4 Decoupled and Pretrained Style Representations

Decoupling content from style provides another route to controlled stylization. Explicit filterbank learning represents styles through dedicated convolutional filter banks while a shared autoencoder carries content information, allowing style processing to remain modular [23]. More recent generalized style-transfer systems use dedicated style-aware encoders and large learned priors to obtain broad style coverage without test-time retraining [24]. Separation of geometry and appearance has also been demonstrated in neural radiance-field visualization, where color and non-photorealistic rendering can change while the reconstructed scene geometry remains stable [25].

The progression summarized in Table 1 shows that prior work commonly improves structure by modifying the training objective, redesigning the generator, or adding learned attention or refinement modules. The present study takes a different position: it assumes that the generator already produces an acceptable cartoon style and treats local boundary loss as a deterministic post-generation correction problem.

3. Proposed AnimeGANv2-AMSPF Framework

The proposed framework separates artistic transformation from structural preservation. The pretrained generator is treated as a fixed style-producing backbone, while a deterministic second branch extracts structural evidence directly from the source photograph. Both branches meet in the proposed AMSPF module, where boundary reinforcement is followed by luminance anchoring and controlled color-space refinement. Figure 1 presents the complete processing flow.

3.1 Input Representation and Preprocessing

Let the original color image be $I \in \mathbb{R}^{H\times W\times 3}$. For the principal experiment, the image is converted to red-green-blue (RGB) representation and resized to 512×512 pixels. Equation 1 defines the resized representation.

$$I_R = \mathscr{R}(I; 512, 512), \tag{1}$$

where $\mathscr{R}(\cdot)$ denotes resizing. Normalization to the generator input interval is performed only for the generative branch according to Equation 2.

$$I_N = 2\left(\frac{I_R}{255}\right) - 1. \tag{2}$$

The unnormalized I_R is retained for structural analysis so that boundary estimation remains independent of generator-specific normalization.

3.2 Frozen Generative Cartoonization Branch

The first branch maps I_N to an initial cartoon using the pretrained AnimeGANv2 Paprika generator. No model parameters are updated in this study. If $G_\theta(\cdot)$ denotes the frozen generator, the initial cartoon is given by Equation 3.

$$C_0 = \text{clip}\left(\frac{G_\theta(I_N)+1}{2}, 0, 1\right). \tag{3}$$

The generator used in the executed pilot contained 2,143,552 parameters. All parameters remained frozen; no discriminator, optimizer, learning-rate schedule, backward pass, or epoch-based training procedure was used.

3.3 Multi-Scale Structural Extraction

A single edge detector can be sensitive to noise and fine texture, while strong smoothing can remove narrow but meaningful boundaries. The structural branch therefore evaluates the source image after three degrees of Gaussian smoothing. First, the RGB image is converted to grayscale using Equation 4.

$$I_g(x,y) = 0.299R(x,y) + 0.587G(x,y) + 0.114B(x,y). \tag{4}$$

For each kernel size $k \in \{3,5,7\}$, Equation 5 defines the smoothed image.

$$B_k = G_k * I_g, \qquad k \in \{3,5,7\}, \tag{5}$$

where G_k is a Gaussian kernel of size $k \times k$ and $*$ denotes convolution.

3.4 Adaptive Edge Detection and Confidence Estimation

To accommodate different illumination and contrast levels, Canny thresholds are derived from the median intensity m_k of each B_k. Equation 6 defines the lower and upper thresholds.

$$T_k^L = \max(0, 0.60m_k), \qquad T_k^H = \min(255, 1.40m_k). \tag{6}$$

The scale-specific edge map is then calculated according to Equation 7.

$$E_k = \mathscr{C}(B_k, T_k^L, T_k^H), \qquad k \in \{3,5,7\}, \tag{7}$$

where $\mathscr{C}(\cdot)$ denotes Canny edge detection. The three edge responses are averaged and stabilized with a 3×3 Gaussian kernel as defined in Equation 8.

$$E = \text{clip}\left[G_3 * \left(\frac{E_3 + E_5 + E_7}{3}\right), 0, 1\right]. \tag{8}$$

A location that remains an edge across several smoothing levels therefore receives stronger confidence than an isolated scale-sensitive response.

Table 1. Representative related work and its relevance to structure-preserving photo stylization.

Ref.	Main emphasis	Preservation strategy	Limitation relative to the present objective
[11]	Texture-synthesis transfer	Region-constrained content retention	Iterative non-neural formulation.
[12]	Structure-preserving neural transfer	Depth and edge representation	Structure integrated into the stylization model.
[14]	Wavelet-attentive transfer	High-frequency detail recovery	Requires a learned transfer framework.
[15]	Structure-guided transfer	Refine network and structure-aware loss	Refinement coupled to model training.
[17]	Few-shot portrait stylization	Domain calibration and local texture translation	Requires target-style examples and training.
[18]	Ink-wash stylization	Saliency and attention	Domain-specific adversarial learning.
[21]	Consistent arbitrary transfer	Self-attention and consistency regularization	Consistency learned during training.
[23]	Explicit style filter banks	Decoupled content/style processing	Style representation remains learned.
[24]	Pretrained generalized transfer	Style-aware encoding and content fusion	Relies on a large generative prior.
[25]	Radiance-field stylization	Geometry/appearance separation	Designed for volumetric visualization.

Table 2. Configuration of the AMSPF refinement module.

Parameter	Setting
Primary resolution	512×512
Gaussian kernels	$3\times3, 5\times5, 7\times7$
Canny multipliers	0.60, 1.40
Confidence smoothing	3×3 Gaussian
Density threshold	0.20
Fusion bounds	$0.10 \le \alpha \le 0.24$
Luminance blend β	0.10
Ratio limits	[0.82, 1.18]
Contrast-limited histogram clip limit	1.25
Local histogram tile grid	8×8
Luminance blend	0.75/0.25

3.5 Edge Density and Adaptive Fusion Weight

The confidence map indicates where stable structures occur but does not specify how strongly they should be reinforced. Scene edge density is estimated using Equation 9.

$$\rho = \frac{1}{HW}\sum_{x=1}^{W}\sum_{y=1}^{H}\mathbf{1}[E(x,y) > 0.20], \tag{9}$$

where $\mathbf{1}[\cdot]$ is an indicator function. The adaptive coefficient is then computed with Equation 10.

$$\alpha = \text{clip}(0.26 - 0.35\rho, 0.10, 0.24). \tag{10}$$

This inverse relationship makes refinement conservative in texture-dense scenes and permits stronger boundary reinforcement in structurally sparse scenes. Table 2 summarizes the fixed configuration used in the experiment.

3.6 Adaptive Structural Reinforcement

The initial cartoon, confidence map, and adaptive coefficient are combined using Equation 11.

$$C_1 = C_0 \odot (1 - \alpha E), \tag{11}$$

where $\odot$ denotes element-wise multiplication. Pixels outside reliable boundaries remain nearly unchanged, while stronger edge locations receive controlled darkening.

3.7 Local Luminance Anchoring

Because Equation 11 can darken reinforced boundaries, the next stage limits tonal distortion. Original and cartoon luminance are calculated with Equation 12.

$$\begin{aligned} L_I &= 0.299R_I + 0.587G_I + 0.114B_I, \\ L_C &= 0.299R_{C_1} + 0.587G_{C_1} + 0.114B_{C_1}. \end{aligned} \tag{12}$$

The bounded local ratio in Equation 13 prevents extreme correction.

$$R_L = \text{clip}\left(\frac{L_I}{L_C + \varepsilon}, 0.82, 1.18\right), \qquad \varepsilon = 10^{-4}. \tag{13}$$

The luminance-adjusted candidate is given by Equation 14.

$$C_A = \text{clip}(C_1 \odot R_L, 0, 1). \tag{14}$$

Rather than replacing C_1 completely, the correction is blended only around confident structures using Equation 15.

$$C_2 = C_1 \odot (1 - \beta E) + C_A \odot (\beta E), \qquad \beta = 0.10. \tag{15}$$

3.8 Lab-Domain Contrast and Color Refinement

The anchored image is transformed into the $L^*a^*b^*$ color representation so that luminance can be refined without directly amplifying chromatic channels. Equation 16 denotes this conversion.

$$(L, a, b) = \mathscr{T}_{RGB\to Lab}(C_2). \tag{16}$$

Contrast-limited adaptive histogram equalization (CLAHE) is applied only to L. The enhanced and original luminance channels are blended using Equation 17.

$$L_F = 0.75L + 0.25\,\text{CLAHE}(L). \tag{17}$$

The final refined cartoon is obtained by Equation 18.

$$C_{\text{final}} = \mathscr{T}_{Lab\to RGB}(L_F, a, b). \tag{18}$$

Algorithm 1 summarizes the complete inference procedure before its formal listing.

3.9 Computational Characteristics

For fixed Gaussian kernels, filtering, edge extraction, confidence aggregation, pixel-wise fusion, color conversion, and local contrast processing all scale linearly with image area. Equation 19 therefore characterizes the additional time complexity.

$$T_{\text{AMSPF}}(N) = \mathscr{O}(N), \qquad N = HW. \tag{19}$$

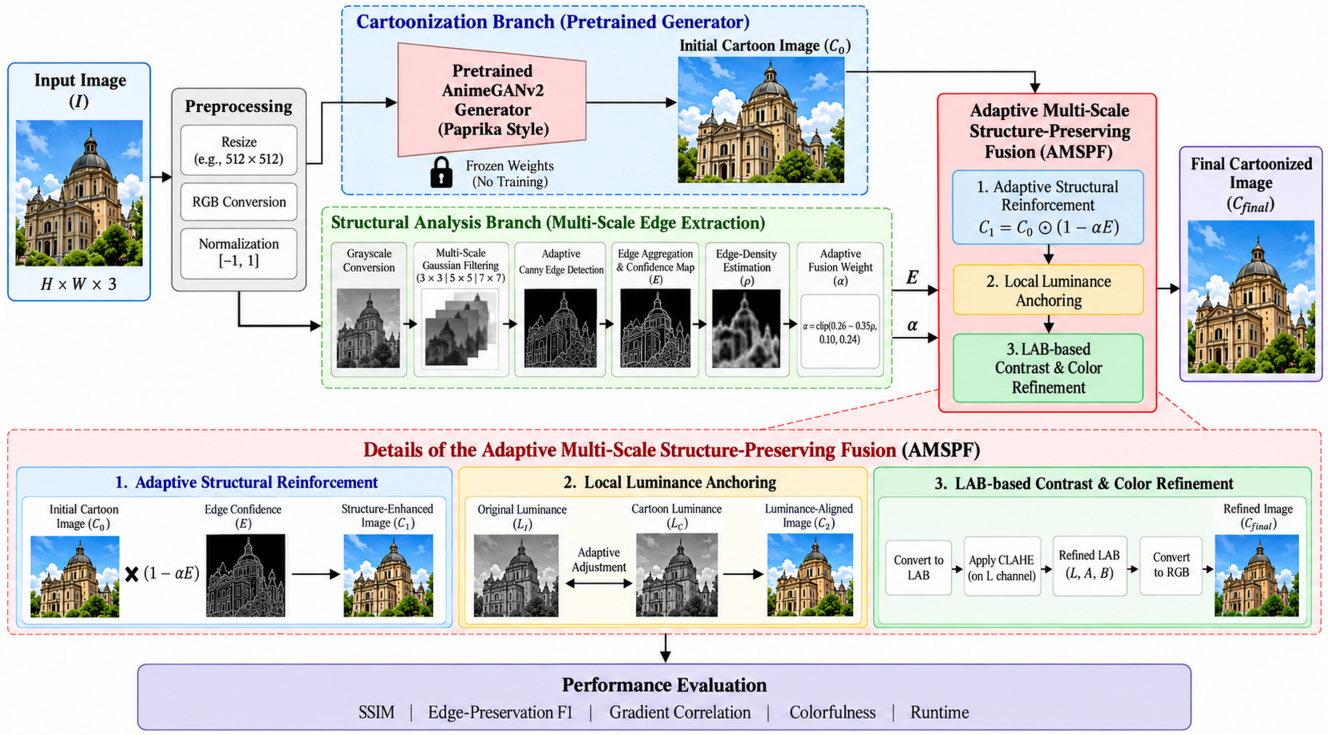

Figure 1. Overall architecture of the proposed AnimeGANv2-AMSPF framework. The upper branch produces the initial cartoon with a frozen generator, while the lower branch extracts multi-scale structural evidence from the source image. The adaptive fusion block performs structural reinforcement, local luminance anchoring, and $L^*a^*b^*$-domain refinement before evaluation.

Algorithm 1 Adaptive Multi-Scale Structure-Preserving Fusion

Require: RGB image I, frozen generator G_θ
Ensure: Refined cartoon C_{final}
Resize I to obtain I_R; normalize to I_N
Generate $C_0 \leftarrow G_\theta(I_N)$
Convert I_R to grayscale I_g
for $k \in \{3,5,7\}$ **do**
    Smooth I_g with G_k
    Compute median-based Canny thresholds
    Obtain scale edge map E_k
end for
Aggregate E_3, E_5, E_7 into confidence map E
Estimate ρ and compute adaptive α
Reinforce structure to obtain C_1
Apply bounded luminance anchoring to obtain C_2
Apply controlled $L^*a^*b^*$ luminance refinement
return C_{final}

The intermediate confidence and image maps also require linear memory, as stated in Equation 20.

$$M_{\text{AMSPF}}(N) = \mathcal{O}(N). \tag{20}$$

The module introduces zero trainable parameters; its cost is incurred only during inference.

3.10 Parameter Interpretation and Stability Considerations

The AMSPF parameters were selected to keep the refinement bounded, interpretable, and independent of image-specific optimization. The three Gaussian kernels serve complementary structural roles rather than acting as interchangeable repetitions of the same detector. The 3×3 branch retains narrow transitions that may disappear under heavier smoothing, the 5×5 branch provides an intermediate representation, and the 7×7 branch suppresses a larger proportion of high-frequency texture before edge detection. Averaging the three responses in Equation 8 therefore favors structures that survive changes in smoothing scale. This mechanism does not assign semantic meaning to an edge; instead, it treats cross-scale persistence as evidence that a boundary is less likely to be caused by isolated fine texture.

The adaptive coefficient in Equation 10 was deliberately constrained to a narrow interval. The upper bound prevents the multiplicative operation in Equation 11 from producing sketch-like darkening, while the lower bound retains a nonzero amount of structural reinforcement even in texture-rich scenes. Because ρ is expressed as a proportion of image pixels rather than an absolute edge count, the control variable remains comparable across the evaluated square resolutions. The linear decrease of α with ρ also keeps the response monotonic and directly interpretable: a denser structural map cannot cause stronger global reinforcement than a sparser one under the same rule.

The luminance stage provides a second bound on the intervention. The ratio in Equation 13 is limited to [0.82, 1.18], so local brightness compensation cannot arbitrarily reconstruct the source photograph. The small blending coefficient in Equation 15 further confines this correction to confident structural regions. Similarly, the final local-contrast operation gives greater weight to the unmodified luminance channel than to the contrast-limited result. These choices are intended to preserve the role separation of the framework: the generator determines style, whereas AMSPF makes a conservative structural correction.

No parameter in the proposed module is learned from the 24 evaluation outputs, and no per-image search is performed

Table 3. Composition of the compact pilot evaluation set.

Source	Visual characteristic	Variants
Astronaut	Portrait and clothing contours	2
Chelsea	Animal and fur texture	2
Coffee	Indoor object scene	2
Rocket	Outdoor geometric object	2
Camera	Human silhouette and detail	2
Horse	High-contrast silhouette	2
Moon	Low-color surface structure	2
Hubble deep field	Dense small-scale structure	2
Coins	Repeated circular boundaries	2
Page	Thin document-like structure	2
Brick	Repetitive geometric texture	2
Grass	Fine natural texture	2
Total	12 public source images	24

during inference. Consequently, two identical inputs processed under the same software configuration produce the same refinement map and output. This deterministic behavior is useful for reproducibility and for isolating the computational contribution of the post-processing stage. It also exposes an important limitation: fixed bounds may not be optimal for every photographic domain. The present study therefore evaluates the fixed configuration as implemented rather than claiming that the selected constants are universally optimal.

4. Experimental Setup

The experimental protocol evaluates AMSPF independently from the pretrained cartoonization backbone. Each source image is processed by the frozen generator alone and by the same generator followed by AMSPF. Because the two configurations share the same input and generator output, measured differences can be attributed to the proposed refinement stage rather than to retraining or model variation.

4.1 Pilot Evaluation Set

The pilot set contains 12 public sample images distributed with the `scikit-image` library. The source collection covers portraits, animals, objects, natural scenes, scientific imagery, documents, silhouettes, and repeated textures. Two deterministic variants were created from each source, giving 24 evaluation samples. No image was used for model fitting because neither the pretrained generator nor AMSPF was trained in this study. Table 3 summarizes the source composition.

Figure 2 illustrates one deterministic variant from each of the 12 source categories so that the structural diversity of the compact pilot set is visible before the quantitative protocol is described.

4.2 Deterministic Sample Construction

The random seed was fixed to 42. For an image of size $H \times W$, the square crop size is defined by Equation 21.

$$s = \max\left[32, \lfloor u \min(H,W) \rfloor\right], \qquad u \sim \mathscr{U}(0.78, 1.00). \quad (21)$$

The crop origin was restricted so that the complete crop remained inside the source image, and horizontal flipping was applied with probability 0.5. Mild intensity perturbation created a second controlled visual variant according to Equation 22.

$$\begin{aligned} I_V &= \mathrm{clip}(aI_C + b, 0, 255), \\ a &\sim \mathscr{U}(0.88, 1.12), \qquad b \sim \mathscr{U}(-8, 8). \end{aligned} \quad (22)$$

These transformations were used only to diversify evaluation inputs; they were not training augmentations.

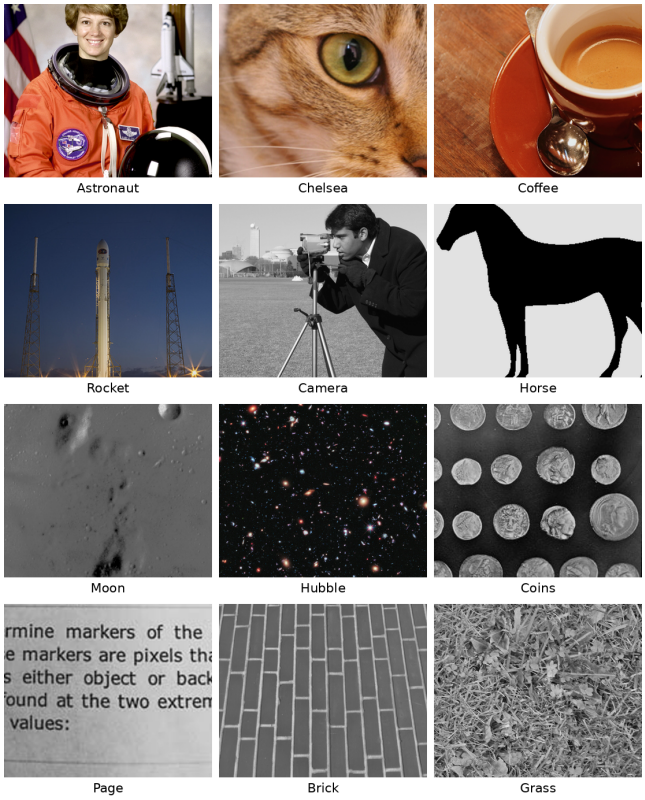

Figure 2. Representative inputs from the 12 public source categories used to construct the pilot evaluation set. A second deterministic crop/photometric variant was generated from each source for the 24-sample paired experiment.

4.3 Implementation Environment

The archived execution recorded PyTorch 2.11.0 on a central processing unit (CPU). The frozen AnimeGANv2 Paprika generator contained 2,143,552 parameters. The proposed refinement introduced no trainable parameters. Table 4 gives the principal configuration.

4.4 Evaluation Metrics

Cartoonization intentionally changes the source image, so a single global similarity measure is insufficient. Structural similarity index measure (SSIM), edge preservation, gradient correlation, colorfulness, and runtime were therefore evaluated together.

For local image regions x and y, SSIM is defined by Equation 23.

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (23)$$

The complementary stylization-strength indicator is defined in Equation 24.

$$S_{\mathrm{style}} = 1 - \mathrm{SSIM}(I, C). \quad (24)$$

The primary task-specific metric is the harmonic-mean edge score (F1). Source and cartoon edge maps were extracted with adaptive Canny thresholds, and a one-pixel dilation tolerance was used for small localization shifts. Equation 25 defines edge precision, recall, and F1.

$$\begin{aligned} P_e &= \frac{|E_C \cap D(E_I)|}{|E_C| + \varepsilon}, \quad R_e = \frac{|E_I \cap D(E_C)|}{|E_I| + \varepsilon}, \\ F1_e &= \frac{2P_eR_e}{P_e + R_e + \varepsilon}. \end{aligned} \quad (25)$$

Table 4. Experimental implementation and inference configuration.

Item	Configuration
Random seed	42
Execution device	CPU
Framework	PyTorch 2.11.0+cpu
Backbone/style	AnimeGANv2/Paprika
Generator parameters	2,143,552
Generator training	None
AMSPF trainable parameters	0
Source images	12
Evaluation samples	24
Primary resolution	512×512
Runtime resolutions	256, 384, 512
Evaluation design	Paired comparison

where $D(\cdot)$ denotes one-pixel dilation.

Gradient agreement is measured from Sobel gradient magnitudes. After vectorization, Pearson correlation is computed using Equation 26.

$$r_g = \frac{\sum_i (m_{I,i} - \bar{m}_I)(m_{C,i} - \bar{m}_C)}{\sqrt{\sum_i (m_{I,i} - \bar{m}_I)^2}\sqrt{\sum_i (m_{C,i} - \bar{m}_C)^2}}. \tag{26}$$

Colorfulness is monitored with opponent red-green and yellow-blue components, as summarized in Equation 27.

$$C_f = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}. \tag{27}$$

4.5 Runtime and Statistical Protocol

Generator and refinement times were recorded separately. Total processing time is defined in Equation 28.

$$t_{\text{total}} = t_G + t_{\text{AMSPF}}. \tag{28}$$

The relative overhead is given by Equation 29.

$$O_{\text{AMSPF}} = \frac{t_{\text{AMSPF}}}{t_G} \times 100. \tag{29}$$

A separate eight-case runtime subset was evaluated at 256, 384, and 512 pixels.

All metric comparisons were paired because each baseline output and refined output originated from the same source variant. The per-sample difference is defined in Equation 30.

$$\Delta M_i = M_i^{\text{AMSPF}} - M_i^{\text{Base}}. \tag{30}$$

Mean paired differences were accompanied by 95% confidence intervals. Because the sample was small and normality could not be assumed, a two-sided Wilcoxon signed-rank test was also applied with $p < 0.05$ as the conventional significance threshold.

4.6 Reproducibility Boundaries of the Pilot Protocol

The experimental design was constructed to make the paired comparison reproducible, but it should not be confused with a conventional train/validation/test study. The proposed refinement has no trainable parameters, and the pretrained generator is never updated; therefore, no training split is required for the reported experiment. The 24 cases are evaluation instances derived deterministically from 12 underlying public source images. All aggregate statistics consequently describe this compact pilot collection and not a larger population of independent photographs.

The fixed seed controls crop selection, flipping, and mild photometric variation, while the same generated variant is supplied to both compared configurations. This paired construction removes variation that would arise if the baseline and refined pipelines received different crops or brightness perturbations. It also means that the second variant of a source image is related to the first and should not be treated as a fully independent photographic subject. For this reason, the paper reports both image-wise paired statistics and source-level descriptive behavior, and it avoids population-scale accuracy or state-of-the-art claims.

Runtime measurements are likewise hardware- and software-dependent. The reported values represent the recorded CPU execution environment and are most useful for comparing the additional cost of AMSPF with the generator under the same run conditions. Absolute latency on a graphics processor, mobile processor, or optimized deployment engine may differ substantially. The relative runtime measurements reported later are therefore interpreted as implementation-specific evidence about the additional cost of the deterministic refinement, rather than as universal throughput figures.

5. Results and Analysis

The analysis considers structural preservation, metric stability, scene-dependent behavior, qualitative appearance, and computational cost. Multi-angle feature-fusion research similarly highlights the need to inspect both local content preservation and global style behavior rather than relying on one metric [26]. Region-aware text-conditioned transfer also demonstrates that foreground and background structures can benefit from selective rather than uniform stylization [27].

5.1 Overall Quantitative Performance

The mean results over all 24 samples are reported in Table 5. The clearest change occurs in Edge-Preservation F1, which rises from 0.6240 to 0.6429. SSIM, gradient correlation, and colorfulness change only slightly, while the mean runtime increases by 52.84 ms per image.

The 3.02% relative Edge-Preservation F1 improvement is substantially larger than the changes observed in gradient correlation and colorfulness. This pattern suggests that AMSPF primarily affects explicit boundary retention rather than globally reorganizing texture or chromatic appearance. Blind quality assessment research likewise distinguishes content and distortion representations rather than assuming that one global similarity statistic fully characterizes perceptual quality [28]. Frequency-aware quality assessment provides complementary evidence that local and global frequency structure can carry information not captured by a single similarity measure [29].

Figure 3 visualizes the image-wise change in Edge-Preservation F1 before its appearance. The positive changes in Figure 3 are distributed across most of the pilot set rather than arising from one isolated example. The largest gain occurs for one brick variant, where repeated geometric boundaries are consistently detected across smoothing scales.

5.2 Paired Statistical Analysis

Table 6 reports the paired change, 95% confidence interval (CI), number of improved samples, and Wilcoxon result. The confidence interval for Edge-Preservation F1 remains positive, and the two-sided Wilcoxon test gives $p = 0.0105$. The SSIM and gradient-correlation differences are not statistically

Table 5. Overall quantitative comparison on the 24-sample pilot set. Values are mean ± standard deviation where available.

Metric	AnimeGANv2	AnimeGANv2+AMSPF	Absolute change	Relative change
SSIM to input	0.6230 ± 0.1224	0.6180 ± 0.1179	-0.0050	-0.80%
Stylization strength	0.3770 ± 0.1224	0.3820 ± 0.1179	$+0.0050$	–
Edge-Preservation F1	0.6240 ± 0.1393	$\mathbf{0.6429 \pm 0.1391}$	$\mathbf{+0.0189}$	$\mathbf{+3.02\%}$
Gradient correlation	0.5336 ± 0.0890	0.5319 ± 0.0891	-0.0016	-0.30%
Colorfulness	31.771	31.651	-0.120	-0.38%
Runtime/image (ms)	5651.74	5704.58	$+52.84$	$+0.93\%$

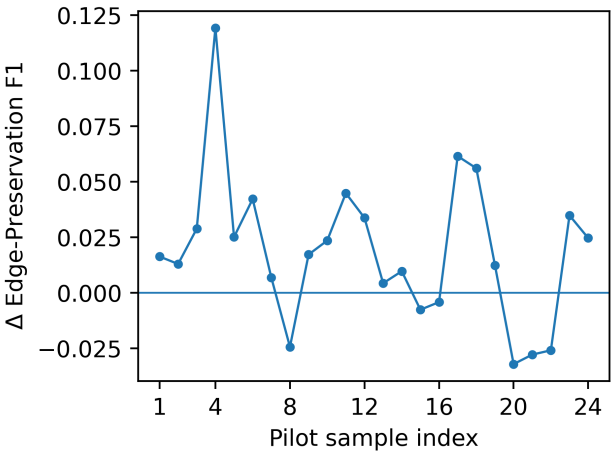

Figure 3. Per-image change in Edge-Preservation F1 after AMSPF refinement. Positive values indicate improved preservation of source-image boundaries.

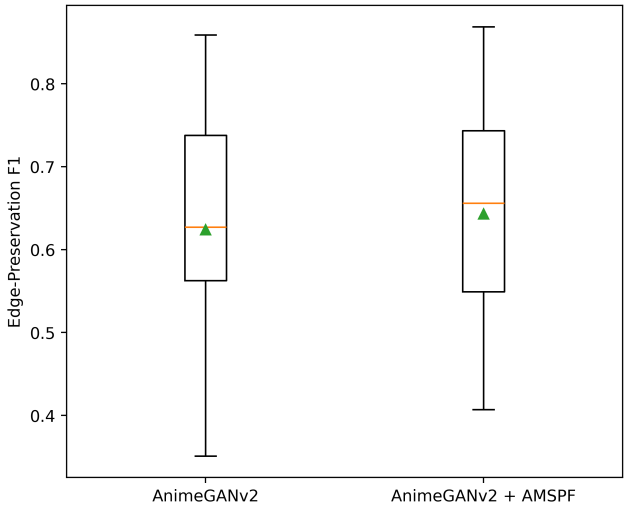

Figure 4. Distribution of Edge-Preservation F1 scores for the pre-trained generator and the generator followed by AMSPF. The triangular markers show the sample means.

Table 6. Paired statistical analysis of the principal structure-related metrics.

Metric	Mean Δ	95% CI	Improved	p
Edge F1	+0.0189	[0.0048, 0.0330]	18/24	0.0105
SSIM	−0.0050	[-0.0119, 0.0019]	8/24	0.0738
Gradient corr.	−0.0016	[-0.0058, 0.0026]	10/24	0.1875

distinguishable from zero at the 0.05 level.

The distributional shift is shown in Figure 4. Three-branch quality models have similarly demonstrated that complementary representations can be informative when one metric alone is insufficient [30].

5.3 Per-Source Structural Behavior

The two deterministic variants from each source were averaged to inspect whether the refinement behaved similarly across distinct structural conditions. The resulting source-wise Edge-Preservation F1 values are reported in Table 7.

Table 7 shows that the largest gains occur for brick, Hubble deep field, coins, camera, and rocket, whereas page, moon, Chelsea, and horse show small negative changes. The strong gains for repeated geometric and locally salient boundaries are consistent with multi-scale edge agreement. The negative cases demonstrate that the method is not uniformly beneficial for every content type. Because each source category contains only two derived variants, these values are descriptive rather than independent statistical groups.

The mixed behavior is consistent with the design objective. AMSPF does not attempt to maximize edge density; it reinforces structures that remain stable under multi-scale smoothing while reducing fusion strength in texture-dense scenes.

Table 7. Source-wise mean Edge-Preservation F1 over the two deterministic variants of each source image.

Source	Baseline	+AMSPF	Δ F1
Brick	0.5957	0.6698	+0.0741
Hubble deep field	0.3511	0.4099	+0.0588
Coins	0.7837	0.8230	+0.0393
Camera	0.6116	0.6453	+0.0337
Rocket	0.7567	0.7864	+0.0298
Coffee	0.6149	0.6353	+0.0204
Astronaut	0.6424	0.6571	+0.0147
Grass	0.8468	0.8538	+0.0070
Horse	0.7157	0.7098	-0.0059
Chelsea	0.5374	0.5287	-0.0087
Moon	0.4450	0.4352	-0.0098
Page	0.5877	0.5609	-0.0268

5.4 Qualitative Analysis

Representative input, baseline, and refined outputs are presented in Figure 5. The examples include a portrait, repetitive geometric texture, repeated object contours, and a dense astronomical scene.

The refinement primarily affects contour regions rather than reconstructing the entire image. Brick boundaries become more distinct, coin edges are strengthened, and portrait changes remain comparatively restrained. Diffusion-based style-transfer systems can produce highly detailed imagery but rely on substantially more complex generative mechanisms [31]. The present objective is narrower: to correct recoverable structure after a satisfactory cartoon has already been generated.

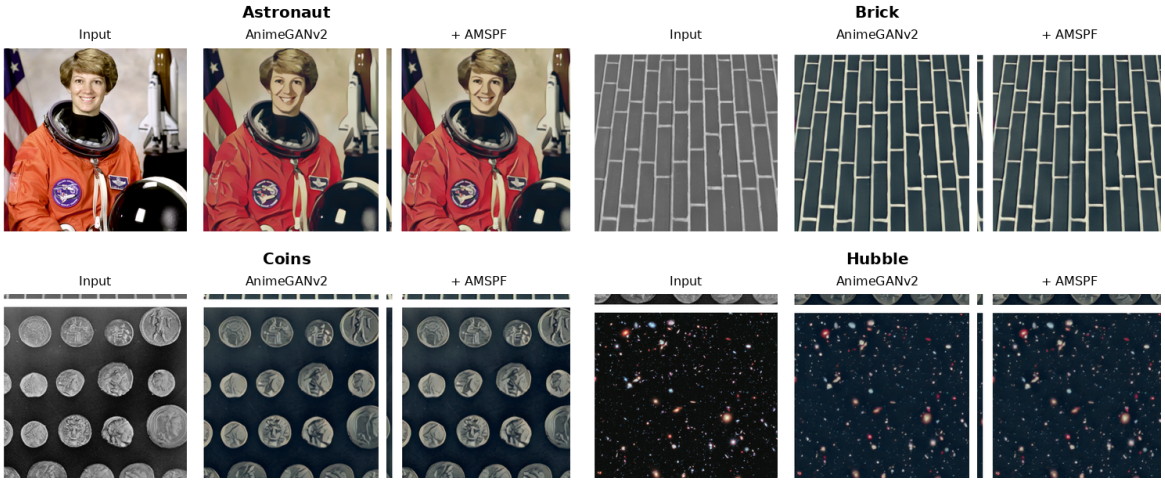

Figure 5. Representative qualitative results. Each row shows the input image, frozen AnimeGANv2 output, and AnimeGANv2 output after AMSPF refinement.

Table 8. Resolution-wise runtime of the generator and AMSPF refinement on the CPU pilot environment.

Resolution	Generator	AMSPF	Total	Overhead
256	1008.36	10.88	1019.25	1.08%
384	2956.85	35.47	2992.32	1.20%
512	5064.67	44.47	5109.14	0.88%

5.5 Adaptive Fusion, Similarity, and Color Stability

The observed α values ranged from 0.10 to 0.24, with a mean of approximately 0.2115. Texture-rich grass samples reached the lower bound because their edge-density estimates were high, whereas several structurally sparse images reached the upper bound. This confirms that Equation 10 behaves adaptively rather than collapsing to an effectively constant coefficient.

The mean SSIM decreased only from 0.6230 to 0.6180, while gradient correlation changed from 0.5336 to 0.5319. Colorfulness moved from 31.771 to 31.651. The limited chromatic change is consistent with applying contrast enhancement only to the luminance channel and leaving the a and b channels unchanged. Concerns about content leakage and representational disturbance in style-related transformations provide additional motivation for conservative, localized modification [32]. Unsupervised image-to-image translation similarly treats preservation of key source structures as a separate objective from appearance transfer [33].

5.6 Runtime and Resolution Analysis

In the 24-image primary experiment, the baseline generator required 5651.74 ms per image on average, whereas the complete pipeline required 5704.58 ms. AMSPF therefore added 52.84 ms, or approximately 0.93% of the baseline inference time. The separate resolution study is summarized in Table 8.

Figure 6 plots the same measurements before the figure appears. A dual-axis representation is used because the generator and total-pipeline runtimes are measured in seconds, whereas the AMSPF refinement requires only tens of milliseconds. The left axis therefore reports AnimeGANv2 and total runtime in seconds, while the right axis reports AMSPF runtime in milliseconds; this preserves the actual measured scale without visually collapsing the refinement curve.

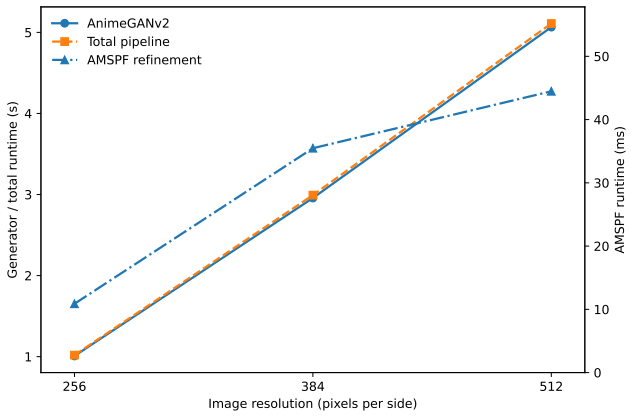

Figure 6. Resolution-wise mean runtime of the frozen AnimeGANv2 generator, AMSPF refinement, and complete pipeline on the CPU pilot environment. The left axis reports generator and total runtime in seconds; the right axis reports AMSPF runtime in milliseconds.

The increase in processing time with resolution is expected because both branches operate on larger spatial representations. The 512-pixel generator value in Table 8 differs from Table 5 because the resolution study uses eight fixed cases, whereas the primary result averages all 24 samples. Reviews of generative adversarial networks for computer vision likewise identify computational efficiency and evaluation methodology as important alongside generation quality [34].

5.7 Contextual Interpretation

The experiment does not support a claim that AnimeGANv2+AMSPF is universally superior to modern style-transfer architectures. The study does not reproduce their training data, style collections, perceptual protocols, or benchmark settings. It demonstrates a narrower result: when a fixed pretrained cartoonization model already provides an acceptable style, deterministic multi-scale refinement can improve source-boundary retention with little additional runtime. This distinction aligns with broader work on controllable image synthesis, where preserving selected characteristics while manipulating others remains an important design challenge [35].

6. Discussion

The findings indicate that AMSPF operates primarily as a targeted structural correction mechanism rather than as a second stylization stage. The measurable change is concentrated in edge preservation, while SSIM, gradient correlation, and colorfulness remain comparatively stable. This behavior is consistent with the intended separation of responsibilities: the generator determines artistic appearance, and AMSPF selectively restores boundary information that remains recoverable from the original photograph.

6.1 Why Multi-Scale Reinforcement Helps

The mean Edge-Preservation F1 gain of 0.0189 occurred in 18 of 24 pilot samples and was accompanied by a positive 95% confidence interval. Multi-scale agreement helps because thin texture responses often disappear as smoothing increases, whereas dominant structural boundaries remain detectable over more than one scale. Averaging the three edge maps therefore acts as a simple confidence mechanism before fusion.

This does not imply that every source contour should survive cartoonization. Texture simplification is part of the target appearance. The adaptive density term is therefore critical: a fixed edge coefficient would treat a smooth portrait and a grass image identically, whereas the proposed rule reduces reinforcement in high-density scenes.

6.2 Stylization–Structure Trade-off

The small SSIM decrease illustrates the expected trade-off between exact photometric similarity and selective contour reinforcement. If the objective were only to maximize SSIM, a trivial solution would be to reduce the degree of stylization. Such a result would be contrary to the purpose of cartoonization. The relevant observation is that the edge-preservation improvement is obtained without a large shift in global similarity or colorfulness.

The gradient-correlation change is also small. AMSPF does not attempt to reconstruct the complete gradient field of the input; it selectively darkens high-confidence boundaries. This explains why the primary task-specific metric improves while the global gradient statistic remains essentially unchanged.

6.3 Luminance Anchoring and Visual Stability

Equation 11 can reduce local intensity around strong edges. Without correction, repeated or dense boundaries could become unnaturally dark. The bounded luminance ratio and small blending coefficient limit this effect. The final $L^*a^*b^*$-domain operation follows the same conservative design: only the luminance channel is mildly equalized, and the original post-fusion luminance receives 75% of the blend. The 0.38% change in colorfulness indicates that this stage does not substantially alter the chromatic character of the generator output.

6.4 Computational Practicality

The proposed refinement adds no trainable parameters. Its operations consist of filtering, adaptive edge detection, simple statistics, pixel-wise arithmetic, color conversion, and local contrast processing. In the recorded CPU environment, the 52.84 ms average cost is small relative to the 5.65 s generator time. The absolute times should not be generalized to graphics-processing-unit, mobile, or embedded hardware, but the measured ratio confirms that the refinement did not introduce another costly generative network.

6.5 Practical Implications

The modular design is useful when a deployed cartoonization backbone already has an acceptable visual style. AMSPF can be enabled, disabled, or modified without retraining the generator. Its parameters are interpretable, and the same generated image can be compared before and after refinement. Such behavior is advantageous in lightweight digital-art tools, educational systems, offline utilities, and other workflows where predictable post-processing is preferred to additional learned complexity.

The present experiment validates this behavior only for the selected AnimeGANv2 Paprika backbone. Although the equations are not mathematically dependent on that generator, compatibility with other cartoonization, transformer, or diffusion backbones remains an empirical question.

6.6 Failure Modes and Limitations

Very fine repetitive texture remains a difficult case because edge detectors can respond strongly to patterns that are not semantically important. Low-contrast boundaries may also be underrepresented if they remain weak at all three smoothing scales. In addition, the generator may intentionally simplify a region for artistic reasons; reintroducing some source contours can then reduce stylistic abstraction even when the boundary metric improves.

The largest limitation is the compact pilot dataset. Twenty-four variants are derived from only 12 underlying source images, so they are not 24 independent photographic subjects. The Wilcoxon result therefore provides evidence for the pilot set rather than population-level generalization. Only one generator and one style were evaluated, and a formal human perceptual study was not conducted. The objective metrics describe structural behavior but cannot establish universal artistic preference.

The method also uses fixed bounds for α, β, luminance correction, and CLAHE strength. These settings were kept constant for reproducibility and were not tuned per image. Broader evaluation may identify content categories that benefit from different parameter ranges.

6.7 Threats to Validity

Internal validity is affected by parameter selection and implementation details. The structural metric is aligned with the method's objective because both derive evidence from source-image edges. Complementary SSIM, gradient, color, runtime, and qualitative analyses reduce but do not remove this concern. External validity is limited by the small public sample set and single cartoonization backbone. Construct validity is limited by the absence of a universally accepted objective measure of cartoonization quality. Statistical validity is limited by the small effective number of independent source images. Accordingly, the reported significance result should be interpreted as pilot-level evidence requiring broader replication.

7. Conclusion

This study introduced a low-overhead adaptive multi-scale structure-preserving refinement framework for pretrained photo cartoonization. The method separates artistic transformation from structural correction: a frozen AnimeGANv2 generator produces the initial cartoon, while a parallel multi-scale branch extracts reliable source-image boundaries and

controls their reintegration through an edge-density-dependent fusion coefficient. Local luminance anchoring and controlled $L^*a^*b^*$ refinement limit excessive darkening and preserve the visual character of the generated image.

On the 24-sample pilot set, mean Edge-Preservation F1 increased from 0.6240 to 0.6429, a relative improvement of 3.02 percent, with positive changes in 18 samples. The paired edge improvement was statistically significant within the pilot sample, whereas the changes in structural similarity and gradient correlation were not. The refinement added approximately 52.84 ms to the average CPU inference time, corresponding to 0.93 percent end-to-end overhead.

The results support the use of training-free structural refinement when the pretrained style is already satisfactory but local boundaries require additional preservation. The study does not establish universal superiority over other cartoonization systems because validation was limited to one backbone, one style, and a compact source collection. Future work should evaluate larger independently acquired datasets, additional cartoonization backbones, semantic edge discrimination, adaptive parameter prediction, and formal perceptual studies with human observers.

Declarations

Acknowledgements. The authors have no acknowledgements to declare.

Funding. The authors received no specific funding from any public, commercial, or not-for-profit funding agency for the research, authorship, or publication of this article.

Author Contributions. Kavali Manisha contributed to the conceptualization, methodology, software implementation, experimentation, data analysis, visualization, and preparation of the original manuscript. A. Soujanya contributed to supervision, validation, interpretation of the results, and critical review and editing of the manuscript. Both authors reviewed and approved the final version of the manuscript and agree to be accountable for the integrity of the work.

Conflict of Interest / Competing Interests. The authors declare that they have no known financial or non-financial competing interests that could have influenced the work reported in this article.

Data Availability. The study utilized publicly available sample images distributed with the scikit-image library. The data used for the reported experiments can therefore be accessed through the corresponding publicly available scikit-image resources. Processed data and experimental results generated during the study may be made available by the corresponding author upon reasonable request.

Code Availability. The experimental implementation and associated computational procedures used to generate the reported results were developed by the authors. Relevant code may be made available by the corresponding author upon reasonable request, subject to applicable institutional and licensing requirements.

Ethics Approval. Not applicable. This study was entirely computational and did not involve the recruitment of human participants, collection of personally identifiable information, clinical intervention, or experimentation involving animals.

Informed Consent. Not applicable, as the study did not involve human participants or the collection of personal or identifiable participant information.

Consent for Publication. Not applicable. The manuscript does not contain identifiable personal information, images, or clinical data requiring individual consent for publication.

Generative AI and AI-Assisted Technologies. Generative artificial-intelligence and AI-assisted tools were used solely to support language refinement, LaTeX preparation, coding assistance, and presentation of the manuscript. The authors independently reviewed and verified all scientific content, experimental procedures, results, interpretations, references, and conclusions. The authors take full responsibility for the accuracy, integrity, and originality of the published work.

Additional Information / Supplementary Material. No additional supplementary material is associated with this article unless explicitly provided with the final published version.

References

[1] X. Zhou, Y. Zheng, and J. Yang, "Bridging the metrics gap in image style transfer: A comprehensive survey of models and criteria," *Neurocomputing*, vol. 624, p. 129430, 2025.

[2] Y. Zhang, Y. Tian, and J. Hou, "Csast: Content self-supervised and style contrastive learning for arbitrary style transfer," *Neural Networks*, vol. 164, pp. 146–155, 2023.

[3] J. Li, L. Wu, D. Xu, and S. Yao, "Soft multimodal style transfer via optimal transport," *Knowledge-Based Systems*, vol. 271, p. 110542, 2023.

[4] M. Liu, S. Lin, H. Zhang, Z. Zha, and B. Wen, "Intrinsic-style distribution matching for arbitrary style transfer," *Knowledge-Based Systems*, vol. 296, p. 111898, 2024.

[5] S. Kim, Y. Min, Y. Jung, and S. Kim, "Controllable style transfer via test-time training of implicit neural representation," *Pattern Recognition*, vol. 146, p. 109988, 2024.

[6] Y. Zheng, J. Jiao, F. Ye, Y. Zhou, and W. Li, "Fast style transfer for ethnic patterns innovation," *Expert Systems with Applications*, vol. 249, p. 123627, 2024.

[7] S. A. Khowaja, L. Nkenyereye, G. Mujtaba, I. H. Lee, G. Fortino, and K. Dev, "Fistnet: Fusion of style-path generative networks for facial style transfer," *Information Fusion*, vol. 112, p. 102572, 2024.

[8] Z. Zhang, Q. Zhang, J. Luan, M. Yang, Y. Wang, and L. Zhao, "Spast: Arbitrary style transfer with style priors via pre-trained large-scale model," *Neural Networks*, vol. 189, p. 107556, 2025.

[9] Y. Dong, S. Liu, Y. Li, and L. Zheng, "Aesthetic-aware adversarial learning network for artistic style transfer," *Neurocomputing*, vol. 646, p. 130431, 2025.

[10] J. Li, Y. Xiang, H. Wu, S. Yao, and D. Xu, "Optimal transport-based patch matching for image style transfer," *IEEE Transactions on Multimedia*, vol. 25, pp. 5927–5940, 2023.

[11] M. Elad and P. Milanfar, "Style transfer via texture synthesis," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2338–2351, 2017.

[12] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure-preserving neural style transfer," *IEEE Transactions on Image Processing*, vol. 29, pp. 909–920, 2020.

[13] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 11, pp. 3365–3385, 2020.

[14] H. Ding, G. Fu, Q. Yan, C. Jiang, T. Cao, W. Li, S. Hu, and C. Xiao, "Deep attentive style transfer for images with wavelet decomposition," *Information Sciences*, vol. 587, pp. 63–81, 2022.

[15] S. Liu and T. Zhu, "Structure-guided arbitrary style transfer for artistic image and video," *IEEE Transactions on Multimedia*, vol. 24, pp. 1299–1312, 2022.

[16] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand, "Style transfer for headshot portraits," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 148:1–148:14, 2014.

[17] Y. Men, Y. Yao, M. Cui, Z. Lian, and X. Xie, "Dct-net: Domain-calibrated translation for portrait stylization," *ACM Transactions on Graphics*, vol. 41, no. 4, pp. 140:1–140:9, 2022.

[18] X. Gao and Y. Zhang, "Sragan: Saliency regularized and attended generative adversarial network for chinese ink-wash painting style transfer," *Pattern Recognition*, vol. 162, p. 111344, 2025.

[19] L. M. Gladence, Y.-W. Lai, F.-T. Lee, M.-Y. Chen, and H.-T. Wu, "Generative adversarial network based on cnn classifier predicted scores for image style transfer," *Applied Soft Computing*, vol. 178, p. 113303, 2025.

[20] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos and spherical images," *International Journal of Computer Vision*, vol. 126, no. 11, pp. 1199–1219, 2018.

[21] Z. Zhou, Y. Wu, and Y. Zhou, "Consistent arbitrary style transfer using consistency training and self-attention module," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 16845–16856, 2024.

[22] X. Kong, Y. Deng, F. Tang, W. Dong, C. Ma, Y. Chen, Z. He, and C. Xu, "Exploring the temporal consistency of arbitrary style transfer: A channelwise perspective," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 8482–8496, 2024.

[23] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Explicit filterbank learning for neural image style transfer and image processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2373–2387, 2021.

[24] J. Gao, Y. Sun, Y. Liu, Y. Tang, Y. Zeng, D. Qi, K. Chen, and C. Zhao, "Styleshot: A snapshot on any style," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 2, pp. 1215–1228, 2026.

[25] K. Tang and C. Wang, "Stylerf-volvis: Style transfer of neural radiance fields for expressive volume visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 613–623, 2025.

[26] Z. Hu, B. Ge, and C. Xia, "Multiangle feature fusion network for style transfer," *Image and Vision Computing*, vol. 154, p. 105386, 2025.

[27] Y. Yu, J. Wang, and N. Li, "Foreground and background separated image style transfer with a single text condition," *Image and Vision Computing*, vol. 143, p. 104956, 2024.

[28] Z. Zhou, F. Zhou, and G. Qiu, "Blind image quality assessment based on separate representations and adaptive interaction of content and distortion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2484–2497, 2024.

[29] Y. Liu, Z. Ni, S. Wang, H. Wang, and S. Kwong, "High dynamic range image quality assessment based on frequency disparity," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4435–4440, 2023.

[30] I. Stępień and M. Oszust, "Three-branch neural network for no-reference quality assessment of pan-sharpened images," *Engineering Applications of Artificial Intelligence*, vol. 139, p. 109594, 2025.

[31] S. Pang, X. Chen, Y. Xie, H. Zhan, B. Yin, and Y. Lu, "Diff-tst: Diffusion model for one-shot text-image style transfer," *Expert Systems with Applications*, vol. 263, p. 125747, 2025.

[32] F. Zhang, B. Feng, Z. Xia, J. Weng, W. Lu, and B. Chen, "Conditional image hiding network based on style transfer," *Information Sciences*, vol. 662, p. 120225, 2024.

[33] H. Lee, J. Seol, S. goo Lee, J. Park, and J. Shim, "Contrastive learning for unsupervised image-to-image translation," *Applied Soft Computing*, vol. 151, p. 111170, 2024.

[34] G. Iglesias, E. Talavera, and A. Díaz-Álvarez, "A survey on gans for computer vision: Recent research, analysis and taxonomy," *Computer Science Review*, vol. 48, p. 100553, 2023.

[35] S. Huang, Q. Li, J. Liao, S. Wang, L. Liu, and L. Li, "Controllable image synthesis methods, applications and challenges: A comprehensive survey," *Artificial Intelligence Review*, vol. 57, p. 336, 2024.