Resolution-Stable Arbitrary Style Transfer via Cross-Scale Low-Frequency Attention Anchoring

Tong Bu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1092216489@qq.com

Hanxi Zhong
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail:bi8bobi8bo0811@qq.com

Yifei Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1310774954@qq.com

Abstract—**Changes in content resolution can cause StyTR-2 to shift its attention correspondences and output texture organization. To address this problem, we propose a training-free cross-scale low-frequency attention-response anchoring method. A frozen low-resolution branch provides a reference whose decoder responses are spatially aligned with those of the high-resolution branch. Only the low-frequency component of the cross-scale response difference is injected, correcting global correspondences while retaining high-resolution texture details. Mutually exclusive diagnostic, development, and test sets are used to evaluate consistency in attention responses, perceptual features, and pixel outputs across three resolution pairs. The proposed method consistently reduces cross-resolution drift. Ablations over frequency range, intervention location, and decoder stage further support the roles of low-frequency constraints, attention-space intervention, and multi-stage anchoring. Overall style statistics remain stable, although the method incurs a small loss of local structural fidelity and additional inference latency. Because no model weights are updated, the method can serve as a resolution-stability extension to StyTR-2 for preview-to-high-resolution export workflows.**

Keywords- Arbitrary Style Transfer, Stytr-2, Cross-Scale Attention, Low-Frequency Anchoring, Resolution Stability

I. INTRODUCTION

A. Research Background and Problem Raising

Image style transfer is designed to preserve the semantic structure of the content image while introducing the color, texture, and brushstroke characteristics of the reference image. Gatys et al. established a neural style transfer paradigm [1] using the deep features of pre-trained convolutional networks and Gram matrices. Subsequently, feedforward networks [2], AdaIN [3], and attention-style methods [4-5] significantly improved generation efficiency and arbitrary style adaptability. In recent years, Transformer has relied on long-range dependency modeling to enter the field of arbitrary style transfer. StyTR-2 uses content and style dual encoders, content-aware positional encodings, and multi-layer Transformer decoders to improve global style coordination [6].

In contrast to earlier methods that relied primarily on local convolutional kernels and channel statistics, Transformer-style transfer expanded content and stylistic features into sequences of tokens, determining from which stylistic locations a content location should derive information through global relationships between queries, keys, and values. The mechanism is able to organize colors and textures across a longer spatial distance and is suitable for handling complex, non-local correspondences between arbitrary styles; but the result also relies on token grids, position coding, and attention normalization. When the input size changes, the number of content tokens, the spatial sampling density, and the original image range covered by a single token will all change. Even if

the content semantics and style reference remain unchanged, the correspondence established by the decoder may not naturally remain consistent. Therefore, resolution is not only a computational parameter, but can also be a behavioral variable in Transformer-style transfer.

Existing research focuses on content-style matching quality, local pattern repetition, semantic structure retention, or computational complexity [5-9, 13-18], while changes in input resolution during deployment are often seen as preprocessing details. In practical applications, the same content image may be input to the model at different scales, such as preview, edit, and high-resolution export. If the model establishes a significantly different content-style correspondence after the scale change, the high-resolution output will not only increase the detail, but may also change the color distribution, texture layout, and local outline. This phenomenon is defined in this paper as cross-resolution response drift and measured as the distance between the low, high-resolution output and the intermediate attention response.

This issue is directly related to the Transformer-based image style transfer project. In interactive applications, users tend to quickly preview the style effect at a lower resolution, and then select a satisfactory content-style combination for high-definition export. Ideally, HD results should add detail to the overall color scheme, main brushstroke orientation, and visual center of gravity at the preservation preview stage; if there is a large change in attention at both scales, the preview cannot reliably represent the final result. Therefore, this paper does not understand multi-scale as the superimposition of more feature layers, but translates it into a measurable Transformer response stability problem.

B. Problem Analysis and Methodological Ideas

An upfront no-training diagnosis showed that the cross-resolution response drift of the six decoder attention stages of StyTR-2 was positively correlated with the output drift. However, training directly through consistency loss systematically lowers the gradient energy and Laplace variance, producing a stable but smoothing result. Based on this negative result, this paper no longer adds training constraints, but starts with the response frequency composition of the inference stage: low-resolution responses provide a more stable global correspondence, while high-resolution responses contain the necessary local details. If only the low-frequency part of the difference between the two is anchored, it is theoretically possible to correct the global correspondence without overwriting the high-frequency residual.

In order to further clarify the source of the drift, in this paper, the intermediate response of the StyTR-2 decoder is observed in stages, and the training consistency constraints, attention response correction, and output spatial fusion are compared, respectively. Preliminary experiments show that although strong consistency constraints can reduce cross-scale differences, they will weaken gradient energy and local textures; similar problems exist with full-frequency response replacement. Based on these observations, this paper limits the scope of intervention to the low-frequency difference of the attention response, and examines the position and number of layers of action on the development set. The overall study flow is shown in Fig. 1.

C. Research Content and Contributions

The main contributions of this paper include: (1) starting from the intermediate response of the decoder, establishing a cross-resolution diagnostic method for StyTR-2, decomposing the scale difference of the same content-style combination into three levels of attention, perceptual features and pixels; (2) proposing a cross-scale low-frequency attention-response anchoring mechanism, which corrects the low-frequency correspondence of the six decoding stages with the low-scale response without increasing the learnable parameters under the condition that the model weights are frozen, while retaining the local details of the high-resolution branch; (3) verifying the stability benefits of the method through independent tests of frequency, position of action and layer number ablation, and three-stage resolution, and quantifying its cost in terms of structural fidelity, inference delay, and GPU memory usage.

II. Related Work And Problem Formulation

A. Neurostyle transfer

Optimized neural style transfer updates the pixels through iteration, so that the generated image matches the content characteristics and style statistics at the same time [1], which has strong flexibility but high inference time. Johnson et al. used perceptual loss for feedforward generation networks, which greatly increased speed, but a single model usually served a limited style [2]. AdaIN transitions any style in real time by aligning the mean and variance of content and style characteristics [3]. WCT further matches covariance statistics in the deep feature space [13], while Avatar-Net achieves multi-scale zero sample migration through feature decoration and multi-layer reconstruction [14]. Such statistical transformation methods have good efficiency and generalization ability, but the description of local semantic correspondence is still limited.

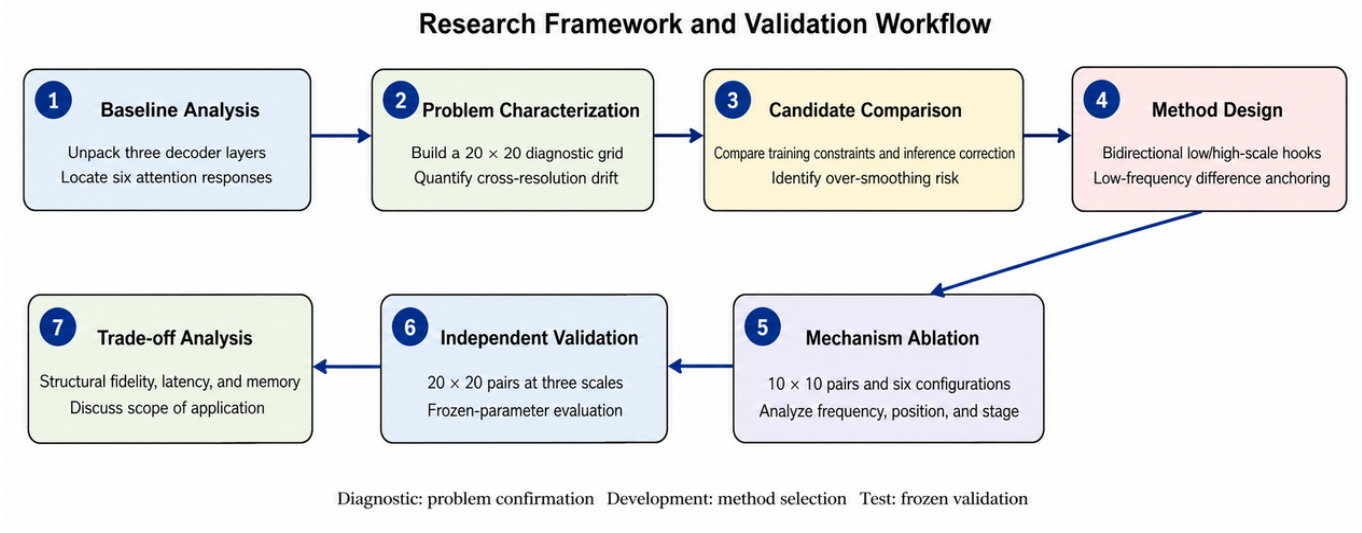

Figure 1. Research Framework and Validation Process

B. Attention & Transformer Style Transfer

SANet reorganizes local style patterns based on the similarity between content and style characteristics, and combines multi-layered features and identity loss to maintain content structure [4]. AdaAttN further uses shallow and deep features to calculate point-by-point weighted statistics to achieve local adaptive normalization [5]. In the direction of linear transformation and reversible network, LST uses a data-driven transformation matrix to balance efficiency and content affinity [15], ArtFlow mitigates content leakage in repeated migration through reversible flow [16], and CAP-VSTNet further discusses the maintenance of feature-pixel affinity [17]. StyTR-2 uses two-way Transformer encoders and multilayer decoders to model long-range dependencies, and proposes content-aware positional encoding for style transfer [6]. Subsequent studies have improved the matching process from the perspectives of stabilizing key style positions [7], aesthetic pattern repeatability [8], fine-grained visual token transforms [9], and hierarchical window attention [18]. The above work mainly optimizes the quality and efficiency of a single migration. This paper focuses on the internal response and output consistency of the same content-style combination at different content resolutions.

In StyTR-2, the convolutional features are first projected and expanded into a content sequence and a style sequence. The two coding branches model their own long-range relationships respectively. The decoder then completes the style matching with the content representation as the query and the style representation as the key and value. The three-layer decoder updates the content layer by layer, indicating that each layer contains twice the attention calculated by the content query style key

value, so that a total of six observable response stages are generated in one forward. The shallow response is closer to the initial content structure, while the deep response is reorganized based on preamble matching. Only observing the final image cannot distinguish whether the drift comes from internal matching or output post-processing, so this paper saves the six-stage response and the final output at the same time, and places them in the same cross-resolution evaluation framework.

C. Multi-scale processing and positioning in the proposed method

Multi-stroke style transfer generates artistic textures of different sizes through scale or attention control [10], and perceptual factors such as space, color, and scale can also be explicitly controlled during the optimization process [20]. STROTSS describes content-style compromises from optimal transmission and self-similarity, and evaluates different settings through user research [19]. These efforts actively adjust the style scale, spatial correspondence, or optimization goals, which is different from the setting in this paper that keeps the style reference unchanged and only changes the content and output resolution. This paper does not propose a new training network, nor does it claim universal multi-scale feature fusion; the research object is limited to the cross-scale attention response of the StyTR-2 inference stage, and the goal is to reduce the corresponding drift caused by resolution changes while retaining high-resolution details.

From the perspective of research variables, traditional multi-scale enhancement often changes the feature hierarchy, convolutional receptive field, or style texture scale at the same time, and performance changes may be caused by multiple factors. In this paper, the style map preprocessing, model weight, loss network and all inference parameters are fixed, only the content and output resolution are changed, and the low- and high-scale behavior of the same content-style combination is compared. Although this control narrowed the scope of the study, it could clearly attribute the experimental conclusions to cross-scale attention response correction, rather than general attribution to "the use of multi-scale information".

III. CROSS-SCALE LOW-FREQUENCY ATTENTION-RESPONSE ANCHORING

A. General framework of the methodology

The proposed method preserves the content encoder, style encoder, transformer parameters, and convolution decoder of StyTR-2, and only extends the response processing flow at the inference stage. The low-resolution reference forward is used to capture the response of the six attention stages and its two-dimensional token grid; after the high-resolution forward reaches the corresponding stage, spatial alignment, response differential, low-pass filtering, and residual reinjection are performed in turn. TABLE 1 summarizes the main differences between this method and the official inference pipeline.

TABLE I. COMPARISON OF THE OFFICIAL STYTR-2 INFERENCE PROCESS WITH THE EXTENSIONS IN THE PROPOSED METHOD

Processing session	Official StyTR-2	The proposed method extends
enter organization	Single Content Scale & Style Chart	The same content is constructed in low and high scales; style input remains fixed
Attention Observation	Directly execute the decoder without saving the phase response	Register a response hook to capture 6 responses from a 3-layer decoder
token space handling	No cross-scale processing	Record 2D grid, bilinear alignment low scale response
Response Updates	The original response continues to pass	Calculate the difference and only inject the 5 × 5 low-pass component back
Model parameters	Use official pre-training weights	Weight completely frozen, no new learnable parameters added
The Evaluation Process	Mainly based on single generation quality	Joint evaluation of response, output, fidelity, latency and memory

Note: This method is an extension of the attention response in the inference stage and does not modify the official pretraining parameters and subject network structure.

B. Problem Definition and Intervention Locations

For a given content image C and a style image S, low-resolution and high-resolution content C are constructed using the same aspect ratio, and the style image S maintains a fixed preprocessing scale. Note that the frozen StyTR-2 is F, and its kth decoder attention phase output response is R(k). Low-resolution and high-resolution forward are obtained by R(k) and R(k) respectively. In this paper, k traverses two content queries in a three-layer decoder: a style key-value attention operation, with a total of six stages.

The core of cross-resolution stability is not to require low- and high-resolution output to be exactly the same, but to require that the two remain similar after alignment — the style corresponds to the overall visual organization, while allowing the high-resolution branches to express additional details. Therefore, this paper simultaneously measures the intermediate response drift, VGG feature drift, pixel drift, edge map drift, and color statistical drift.

Even if the two inputs are only scaled by the same artwork, the local edge's landing point in the mesh, content-aware positional encoding, and the competitive relationship between adjacent tokens will change. In the diagnostic phase of this paper, the six-stage response drift was observed to change in the same direction as the final VGG/pixel drift, so the intermediate response rather than the final pixel was used as the main intervention location.

C. Low-Frequency Response Differential and Anchoring

First, the low-resolution response is restored to the spatial feature map according to its two-dimensional token grid, and aligned to the high-resolution token grid by the bilinear interpolation operator U (·). The cross-scale response difference for stage k is defined as:

$$\Delta \perp (k) = U\left(R_{l\perp(k)}\right) - R_{h\perp(k)} \tag{1}$$

Differential Δ aves angers include both global corresponding changes and local high-frequency changes. Direct use of full differential results in a high-resolution response approaching a low-resolution response, possibly erasing local textures. In this paper, the low-frequency difference is extracted using a 5 × 5 average pooling P (·) with a step size of 1 and boundary zeroing, and the high-resolution response is injected with the coefficient α:

$$R_{hat}, h \perp (k) = R_{h\perp(k)} + \alpha \cdot P_{5(\Delta\perp(k))} \tag{2}$$

All experiments were fixed α = 1 and K = 5 and were not adjusted on the official test set. Since only the differential low-frequency portion is added, the high-frequency residuals not covered by the low-pass operator in the high-resolution response are still provided by the original branch. After the revision of R, the subsequent decoding module is passed in, and the anchoring output I is finally obtained.

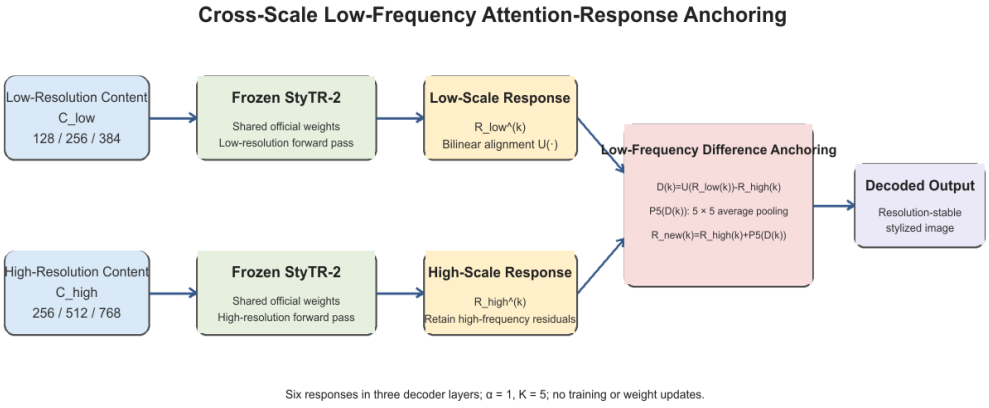

Figure 2. Cross-scale low-frequency attention response anchoring method

Averaging pooling is not interpreted here as an ideal frequency domain filter, but rather as a parameterless, localized and repeatable low-pass approximation. Its role is to preserve the cross-scale differences that are consistent in the direction within the neighborhood, so that the fine-grained differences of the symbols that change rapidly cancel each other out. The former typically corresponds to color areas, primary texture trends, and wide-range content-style matches, while the latter is more likely to include edges and fine strokes specific to high-resolution branches. Since $\alpha = 1$ acts only on the difference after the low pass, it is not equivalent to replacing the high-resolution response with a low-resolution response; full-frequency replacement ablation is a direct test of this difference. Unlike the light reality style transfer method WCT ² [23], which uses wavelet transform to separate high and low-frequency information, this paper does not reconstruct the image frequency band, but only corrects the cross-scale low-frequency difference of the intermediate response of the Transformer.

D. Inference Steps and Complexity

The official StyTR-2 baseline requires only one high-resolution forward. This method first carries out a low-resolution forward to cache six responses, and then carries out a high-resolution forward with hook correction. The model parameters, decoder weights, and VGG evaluation network remain frozen, so there are no training sets, optimizers, or new learnable parameters. The extra graphics memory comes mainly from the low-resolution response cache; the extra computation comes from the low-resolution forward. As the target resolution increases, the computational ratio of low-resolution branches to high-resolution branches decreases, so the delay ratio overhead is expected to decrease with the increase of the target scale.

The specific reasoning process is divided into four steps. The first step is to construct a low-resolution content input according to half of the target output scale, execute the frozen StyTR-2 forward and cache six responses by layer; the second step is to use the target resolution content to start the high-resolution forward, restore the two-dimensional mesh of the cache response at the corresponding stage, and complete the bilinear alignment; the third step is to calculate the cross-scale difference and apply a 5 × 5 low-pass, and only add the low-frequency correction amount to the current high-resolution response; the fourth step is to continue to hand over the corrected sequence to the original decoder and the convolution decoding network to obtain the final stylized image. The whole process does not change the model file, close the anchor hook to restore the official baseline.

If the time-consumption of a low-resolution and high-resolution forward is recorded as TXX and TXX respectively, the baseline complexity corresponds to TXX. The method in this paper approximates TXX + TXX as well as a small amount of interpolation and pooling overhead. The size of the response cache is proportional to the number of six-stage low-resolution tokens and channels, not to all high-resolution intermediate features. Therefore, the method is expected to exhibit a significant increase in computational time and a smaller increase in peak memory. Section 5.4 Independent synchronous timing and memory measurements are used to verify this engineering judgment.

IV. Experimental Protocol

A. Data and Partitioning

Experiments use content images and style images from the public style-transfer-dataset [11]. This resource is released under the CC BY license and contains 50 content images, 50 style images, and existing stylized results at different style scales. The proposed method only uses the original content/style images, and does not use the ArtFlow generated results or user ratings as the performance data for this method. In order to avoid confusing problem discovery, method screening, and formal inspection, the 20 × 20 combination was first used for untrained diagnosis; in the remaining 30 mutually exclusive contents and 30 styles, the first 10 × 10 was used as the development/ablation set, and the remaining 20 × 20 was used as the frozen test set.

Development set content numbers are 5, 6, 8, 10, 15, 18, 19, 22, 23, 24, style numbers are 4, 6, 8, 12, 17, 20, 22, 23, 26, 27. Test set content numbers are 26, 28, 29, 30, 32, 33, 34, 35, 36, 39, 40, 41, 43, 44,

45, 46, 47, 48, 49, 50; style numbers are 28, 29, 30, 32, 33, 34, 35, 36, 38, 39, 40, 41, 42, 44, 45, 46, 47, 48, 49, 50. The two collection filenames do not overlap.

The reason for using Cartesian products instead of randomly sampling a small number of pairs is that the result of any style transfer is influenced by both the content structure and the style texture: the same method may behave differently for buildings, characters, or natural scenes, and may also behave differently for paintings, watercolors, or high-contrast textures. The full content x style grid is able to check if each content layer and each style layer is in the same direction separately, while avoiding showing only a few favorable combinations. Since 400 pairs are composed of 20 content and 20 style repeats and cannot be considered as 400 completely independent image samples, this paper uses content × style bidirectional clustering statistics instead of ordinary independent sample confidence intervals.

B. I.Implementing the Setup

The experiment uses the official pre-training weight of StyTR-2 [6], and the operating environment is PyTorch 2.1.0a0, CUDA single card 24 GB. The image maintains its original aspect ratio and is scaled to an integer multiple of 8. The long edge of the style image is fixed at 256; the content image uses 128→ 256, 256→ 512, and 384→ 768 triple low-high long edge combinations. The main method fixes $\alpha = 1$, low-pass kernel $K = 5$, and acts on all six decoder responses. All experiments were gradientless inference.

C. Evaluation indicators and statistical methods

The stability indicators include: (1) the mean value of the six-layer attention response cosine drift; (2) the pre-trained VGG-19 normalized feature distance between the low-resolution output and the aligned high-resolution output [12]; (3) the average absolute pixel drift; (4) the Sobel edge map drift; and (5) the RGB mean and standard deviation drift. Quality protection indicators include VGG content loss, layer-by-layer mean/standard deviation style loss, edge correlation between generated graph and content graph, gradient energy, Laplace variance, and output standard deviation. The lower the drift and loss, the higher the edge correlation; clarity statistics are only used as a deterioration warning and are not interpreted as a monotonous aesthetic quality score.

Content × style combinations are not 400 images independent of each other. In order to avoid mistaking the same content or style for a stand-alone sample, this paper uses content and style to perform 20,000 bidirectional clustering bootstrap for two clustering dimensions: each time the sample content index and style index are played back, and then the pairing average difference of the complete submatrix is calculated, reporting 2.5% and 97.5% quartiles. Development and test sets use the same statistical methods.

Attention drift uses the cosine distance between the aligned responses, focusing on comparing the corresponding directions to reduce the effect of absolute amplitude; VGG drift compares the difference in perceived feature space between low-resolution output and zoomed-aligned high-resolution output; pixel drift is more sensitive to color and local position changes. The three types of indicators are expanded step by step from the internal mechanism to the final output. If only the pixel drift decreases and the attention drift remains unchanged, the output may only be smoothed or fused; if the attention, VGG and pixel indicators are improved at the same time, it is more consistent with the interpretation that the internal correspondence is stable. Content, style and edge indicators are used as quality protection conditions to determine whether stability is obtained by destroying the original mission target.

D. Experimental control and reproducibility

Experiments follow the order of first diagnostic, later development, and then freezing tests. The diagnostic set is only used to confirm the research question; the 10 × 10 development set is used to compare the frequency of action and the number of layers, and fix $\alpha = 1$, $K = 5$ and six-stage full-layer anchoring at this stage; the 20 × 20 test set only performs the official baseline and the frozen main method, and does not reselect parameters based on the test results. All pair-by-pair indicators, clustering intervals, completion logs, and efficiency records are saved as recalculable structured files, and qualitative graphs are also generated directly

from the corresponding experimental outputs. This control avoids mixing method discovery, hyperparameter selection, and final validation into the same round of experiments.

The phased design described above separates problem identification, protocol selection, and final validation from each other. The method configuration is only determined on the development set; after entering the independent test, the α, low-pass core and the number of layers of action remain unchanged, thus avoiding the reverse adjustment method using the test results.

V. Results And Analysis

A. Cross-resolution stability on independent test sets

TABLE 3 shows the 256→ 512 frozen test set results. The main method reduced the attention response drift from 0.148458 to 0.103954, a relative decrease of 29.98%; the VGG output drift and pixel drift decreased by 7.51% and 6.68%, respectively. The three confidence intervals were strictly less than zero, and the attention/VGG/pixel drift improved on 400/397/396 pairs, with the mean of 20 content layers and 20 style layers being in the same direction.

TABLE II. EXPERIMENTAL PHASE, DATA SCALE AND USAGE

Experimental phase	Data Composition	Review Size	Experimental Purpose
Problem Found	20 content x 20 style	400 pairs	Analyze six-stage response drift and its relationship to output variance
Mechanism Ablation	10 content x 10 styles, 6 configurations	100 pairs x 6 configurations	Comparison frequency range, action position, number of layers and output fusion
Independent testing	20 content x 20 style, 3 scale, 2 methods	400 pairs × 3 scales × 2 methods	Testing Stability, Scale Consistency and Quality Protection Indicators
Efficiency testing	3 scale × 2 methods	6 sets of configurations	Preheat each group 10 times and synchronize 50 times while recording the memory

Note: Diagnostic, development, and test sets do not overlap; efficiency testing is independent of quality evaluation.

TABLE III. 256→ 512 FROZEN TEST SET MAIN RESULTS (20 X 20 PAIRING)

Specifications	StyTR-2	Method in the proposed method	Relative change	Pairing Difference 95% CI
Attention response drift↓	0.148458	0.103954	−29.978%	[−0.050544, −0.038766]
VGG Output Drift↓	0.210621	0.194794	−7.514%	[−0.018761, −0.012999]
Pixel drift↓	0.099045	0.092427	−6.682%	[−0.008928, −0.004764]
Content loss↓	0.003741	0.003758	+0.452%	[+0.000006, +0.000029]
Style Loss↓	2.793777	2.781211	−0.450%	[−0.042131, +0.015793]
Edge Correlation↑	0.573272	0.565006	−1.442%	[−0.012225, −0.004442]
Gradient Energy	0.048481	0.048186	−0.610%	[−0.000558, −0.000046]
Laplace variance	0.030537	0.030433	−0.343%	[−0.000458, +0.000253]
Output standard deviation	0.251655	0.253954	+0.914%	[+0.000432, +0.004004]

Note: The difference is defined as the method of the proposed method − StyTR-2; the lower the drift/loss indicator, the better. CI uses 20 000 times content × style bi-directional clustering bootstrap.

The quality index showed that the confidence interval of the style loss change was across zero, indicating that the main method did not steadily change the overall style statistics; the Laplace variance also had no significant direction, and the output standard deviation increased by 0.91%. Content loss increased by 0.45%, edge correlation decreased by 1.44%, and gradient energy decreased by 0.61%, indicating that stability improvement is not completely free of cost. The magnitude of the above changes is much smaller than the degradation caused by full frequency replacement, but it should still be explained as a compromise method.

The internal response improves more than the final output because subsequent residual

connections, feedforward layers, and convolutional decoding still retain some of the original high-resolution information. As a result, larger mid-response corrections translate into milder image variations, both reducing overall drift and avoiding direct convergence of high-resolution output to low-resolution results. Consistent improvements in 400 pairs and layering of all content and style suggest that this benefit is not caused by a small number of extreme samples.

B. Stability at Different Resolutions

The attention response drift on the third scale decreased by about 30% to 34%, the VGG drift decreased by 7.4% to 8.1%, and the pixel drift decreased by 6.0% to 7.5%. The style loss and Laplace variance did not change significantly in all three grades; the edge correlation decreased by 1.22%, 1.44% and 0.78%, respectively. 384→ 768 The content loss did not change significantly, indicating that the structural cost did not monotonically increase with the target resolution. The drift is significantly reduced in each of the six attention stages at each scale, ruling out an explanation that the benefit comes from only one level.

The similar magnitude of change on the third dimension indicates that the method does not rely on a single number of tokens or a specific interpolation size. The attentional drift reduction was slightly lower on the intermediate scale and slightly higher on the two-end scale, but did not change monotonically with resolution; the VGG and pixel drift reduction also remained in a narrow range. Existing conclusions are limited to the test interval of 128→ 256 to 384→ 768, and performance beyond 768 long edges still needs to be verified.

TABLE IV. RELATIVE CHANGE OF FREEZING MAIN METHOD AT THREE RESOLUTION LEVELS

Scales	Attention Drift	VGG Drift	Pixel drift	Content loss	Style Loss	Edge Correlation	Output standard deviation
128→256	−34.024%	−8.143%	−7.481%	+0.825%	+0.048%	−1.216%	+1.717%
256→512	−29.978%	−7.514%	−6.682%	+0.452%	−0.450%	−1.442%	+0.914%
384→768	−31.891%	−7.403%	−5.958%	+0.224%	−0.515%	−0.782%	+1.569%

Note: Each file is the same 20 × 20 frozen test grid; the 95% CI of the three main drift indicators in the bidirectional clustering on the three scale is strictly favorable.

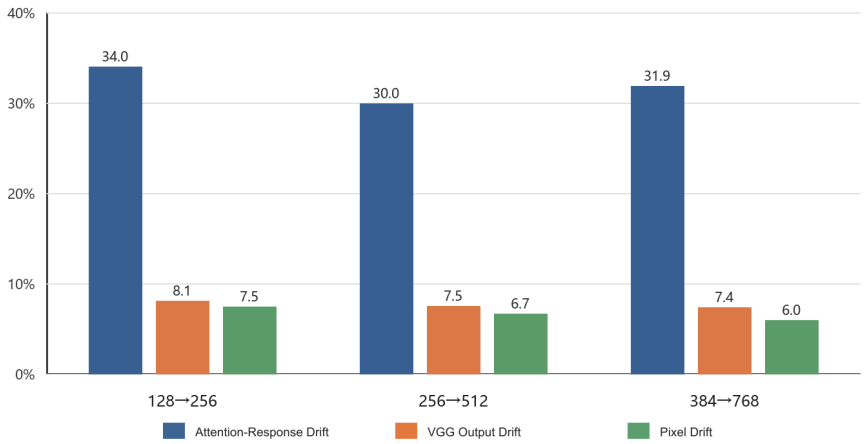

Figure 3. Attention, VGG and Pixel Drift Reduction at Three Gear Resolution

C. Mechanism Ablation Analysis

Full-frequency full-layer replacement reduced attention response drift by 85.81%, but VGG drift improved erratically and increased style loss by 69.65%, gradient energy by 25.75%, and Laplace variance by 34.34%. This means that the closer the response value is to the lower resolution, the more stable the visual result is: full-frequency replacement overrides the local texture borne by the high-resolution branch. In contrast, low-frequency full-layer anchoring retains high-frequency residuals, leaving style loss unchanged while achieving significant stability gains.

Output low-frequency fusion directly pulls high-resolution results to low-resolution results in the image space, so it can mechanically reduce VGG and pixel drift, but the intermediate attention drift is completely unchanged, and is accompanied by a 27.42% increase in style loss and a 5.09% decrease in gradient energy. This negative control shows that the goal of the method in this paper is not simply to smooth the output, but to stabilize the internal correspondence that produces the output. Both L1-only and L3-only have significant but weak drift improvement, and full-layer anchoring achieves greater comprehensive income, supporting cross-layer cumulative effect.

These results support the method design from different aspects. Full-frequency replacement shows that excessive depression of response differences sacrifices high-resolution textures; output fusion can directly reduce image distance without changing intermediate attention; single-layer anchoring cannot prevent subsequent stages from re-accumulating drift. Therefore, low-frequency constraints, spatial intervention of attention, and multistage combination all play a role in the final effect.

Fig. 4 retains the qualitative comparison in the original manuscript, which is used to visually observe the differences in the overall color matching, texture organization, and local details of the negative control such as the official StyTR-2 and full-frequency replacement; Fig. 5 further calculates the error intensity and improvement distribution with the original test output, which is used to locate the specific changes in cross-resolution consistency. The two drawings provide visual quality evidence and error mechanism evidence, respectively, to complement each other.

Fig. 5 uses 6 sets of raw outputs that were saved in advance at the time of the test run, rather than being calculated twice from the compressed graph of the paper. In order to present the benefit range at the same time, two pairs with larger and smaller pixel drift were selected from these 6 groups; this selection does not participate in parameter determination and does not represent the global best or worst results of 400 test pairs. The error map is calculated from the low-resolution output and the high-resolution output after bilinear downsampling, pixel by pixel. StyTR-2 uses the same color scale as the method in this paper. The error change map shows the error decrease after anchoring in green and the error increase in red. The pixel and VGG drift of both samples decreased, but the local error increase region was still visible, indicating that the method improved overall cross-scale consistency, rather than ensuring monotonous optimization of each pixel or local structure.

D. Reasoning Efficiency Analysis

The peak allocated memory increment of this method is always less than 0.8%, indicating that the response cache is not the main capacity bottleneck. The delay increase decreased from 107.0% at 256 output to 39.2% at 768 output, in line with the expectation that the proportion of low-resolution branch relative calculations decreased with the increase of the target scale. This cost may be acceptable for offline high-resolution creations; additional low-resolution fronts remain a major limitation for live previews.

TABLE V. MECHANISTIC ABLATION OF THE DEVELOPMENT SET (CHANGE RELATIVE TO STYTR-2)

variant	Attention Drift	VGG Drift	Pixel drift	Style Loss	Edge Correlation	Gradient Energy	Laplace variance
Low-Frequency Full Layer (text)	−30.934%	−8.758%	−6.689%	−0.269%	−1.099%	−0.383%	+0.457%
Full Band Full Layer	−85.813%	−3.926%*	−19.782%	+69.651%	−6.982%	−25.748%	−34.342%
Low-Frequency L1-only	−9.961%	−3.118%	−3.298%	+0.452%	+0.478%*	−0.468%	−0.734%*
Low-Frequency L3-only	−9.270%	−6.157%	−5.361%	−0.206%	−0.737%	−0.234%*	+0.218%*
Output Low-Frequency Fusion	0.000%	−57.276%	−55.239%	+27.417%	−1.433%*	−5.087%	+5.206%

Note: * indicates that the bi-directional clustering of this indicator has a 95% CI across zero; the changes listed in the remaining tables are interpreted in the original direction. Ablation uses only the 10x10 development set and is not used to reselect formal test parameters

Figure 4. Qualitative comparison of the main method and the negative control

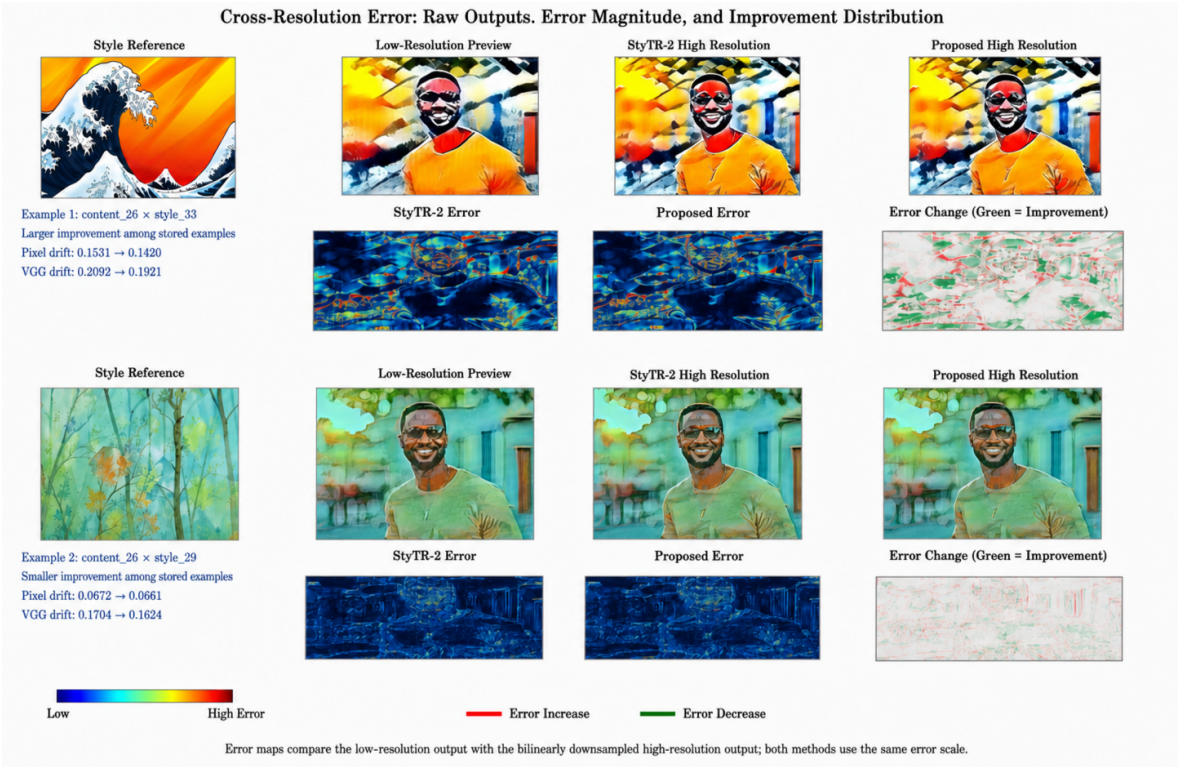

Figure 5. Cross-resolution output, error intensity and improved distribution of pre-stored test samples

TABLE VI. DEPLOYED LATENCY AND PEAK ALLOCATED MEMORY

Scales	StyTR-2 Median Delay/ms	Bit delay/ms in the proposed method	Delayed Amplification	StyTR-2 Video Memory/MB	Memory of the proposed method/MB	Graphics Memory Delta
128→256	16.315	33.775	+107.0%	430.8	433.8	+3.0（+0.70%）
256→512	53.456	86.144	+61.1%	1442.4	1453.6	+11.1（+0.77%）
384→768	157.509	219.251	+39.2%	5686.6	5714.1	+27.5（+0.48%）

Note: Each method warms up 10 times and performs 50 CUDA synchronization timings. StyTR-2 is a high-resolution forward; this method is a low-resolution forward plus a guided high-resolution forward.

The absolute delay also reflects the difference in usage scenarios: at 256 output, the low-resolution auxiliary forward is close to the main forward scale, so the overall time consumption is nearly doubled; after the target scale rises to 768, the high-resolution main branch becomes the main calculation volume, and the relative proportion of the auxiliary branch decreases. As a result, the engineering system can turn off the anchoring during the rapid trial style phase and enable the stable mode when the user confirms the combination and performs the high-definition export to avoid placing all interoperability below the double forward cost.

VI. Discussion

A. Balancing stability and fidelity

Stability improvement is not equivalent to synchronous improvement of various quality indicators. This method reduces the cross-resolution drift steadily, but the edge correlation decreases significantly from 0.78% to 1.44% on the third scale, and the content loss and gradient energy of some scales also change slightly unfavorably. This indicates that after the low-frequency correspondence is corrected, the generated map may shift towards a low-resolution organization on the local contour. At the same time, the style loss and Laplace variance did not deteriorate systematically, the output standard deviation increased, and the qualitative sample did not show obvious smoothness of full-frequency replacement. Therefore, the method in this paper introduces a small but measurable structural fidelity cost while improving resolution stability.

B. Scope of application of the method

This method is suitable for scenarios that require the same content-style combination to maintain visual organization between preview and high-resolution export, such as digital art creation, interactive style editing, and multi-terminal output. The method does not require training, can directly affect the existing StyTR-2 weights, and the graphics memory overhead is small. For tasks that generate only a single fixed resolution, are demanding for real-time, or rely specifically on thin lines and text edges, additional delays and decreases in edge correlation may offset stability gains.

C. Project Implementation and Project Linkage

From the perspective of project implementation, this method can be used as an optional stabilization mode in the Transformer style transfer process without replacing the original StyTR-2 model. The system first receives the content image, style image, and target output size, and the normal mode directly executes the official single forward; the stable mode generates additional half-scale content input, caches the six-stage response, and completes the low-frequency correction in HD forward. The two modes share the same set of official weights and image preprocessing process, which facilitates baseline comparison, quick preview and high-definition export in the same interface, and also keeps the algorithm research consistent with the original project "Transformer-based image style transfer".

By its methodological nature, the scheme is an extension of the attention response in the reasoning phase and does not alter the StyTR-2 subject network. Its engineering significance is to provide an optional stable mode for the preview-HD export process, while giving a cross-resolution evaluation method that is compatible with this mode. Therefore, the project results should be expressed as an improvement in the stability of the Transformer style transfer resolution, rather than an overall improvement in the quality of the single output.

D. Limitations

There are still four limitations to this study. First, the experiment revolves around StyTR-2, and it has not been proven that the same mechanism can be directly migrated to AdaAttN, StyA2K, or other attention architectures. Second, the data comes from a mutually exclusive subset of public 50 × 50 style transfer resources. Although the content/style two-way clustereding statistics are used, the coverage of content and artistic style is still limited. Third, the evaluation is mainly based on perceptual characteristics, statistics and edge agents, and there is still a lack of blind subjective evaluation for cross-resolution consistency. Fourth, α and low-pass nuclei are frozen as a single working point in

the development stage, and dynamic gating is not constructed in this paper to automatically adjust the anchoring strength according to the content structure. Subsequent work should validate generalization on larger Coco/WikiArt subsets, other attention models, and human preference experiments, and explore structurally sensitive adaptive low-frequency weights.

VII. Conclusions

According to the attention and output drift of StyTR-2 under the change of content resolution, this paper proposes a cross-scale low-frequency attention-response anchoring method. The method provides a reference with frozen low-resolution branches, correcting only the low-frequency response difference of the six decoding stages at high-resolution. Independent tests showed that attention, VGG and pixel drift were significantly reduced at the three-stage scale; full-frequency replacement, single-layer anchoring and output fusion experiments further illustrated the role of low-frequency constraints, attention space intervention and multi-stage combination. While the style loss is generally stable, the method introduces small content and edge fidelity costs, as well as 39% -107% delay overhead, and the peak memory increment is less than 0.8%. The above results show that low-frequency response anchoring can improve the cross-resolution consistency of StyTR-2, but its application requires a trade-off between stability, local structure and computational cost.

Acknowledgment

This article is supported by “Xi'an Technology University's School-level College Students' Innovation and Entrepreneurship Training Program in 2025 (Project Number: X202510702184)”.

References

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

[2] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 694-711.

[3] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 1501-1510.

[4] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 5880-5888.

[5] S. Liu, T. Lin, D. He, et al., "AdaAttN: Revisit attention mechanism in arbitrary neural style transfer," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 6649-6658.

[6] Y. Deng, F. Tang, W. Dong, et al., "StyTR2: Image style transfer with transformers," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 11326-11336.

[7] M. Zhu, X. He, N. Wang, et al., "All-to-key attention for arbitrary style transfer," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 23109-23119.

[8] K. Hong, S. Jeon, J. Lee, et al., "AesPA-Net: Aesthetic pattern-aware style transfer networks," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 22758-22767.

[9] J. Wang, H. Yang, J. Fu, et al., "Fine-grained image style transfer with visual transformers," in Proc. Asian Conf. Comput. Vis. (ACCV), 2022, pp. 841-857.

[10] Y. Yao, J. Ren, X. Xie, et al., "Attention-aware multi-stroke style transfer," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 1467-1475.

[11] V. Kitov, "style-transfer-dataset," GitHub repository. [Online]. Available: https://github.com/victorkitov/style-transfer-dataset. Accessed: Aug. 12, 2026.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[13] Y. Li, C. Fang, J. Yang, et al., "Universal style transfer via feature transforms," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[14] L. Sheng, Z. Lin, J. Shao, et al., "Avatar-Net: Multi-scale zero-shot style transfer by feature decoration," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 8242-8250.

[15] X. Li, S. Liu, J. Kautz, et al., "Learning linear transformations for fast image and video style transfer," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 3809-3817.

[16] J. An, S. Huang, Y. Song, et al., "ArtFlow: Unbiased image style transfer via reversible neural flows," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 862-871.

[17] L. Wen, C. Gao, and C. Zou, "CAP-VSTNet: Content affinity preserved versatile style transfer," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 18300-18309.

[18] C. Zhang, X. Xu, L. Wang, et al., "S2WAT: Image style transfer via hierarchical vision transformer using strips window attention," Proc. AAAI Conf. Artif. Intell., vol. 38, no. 7, pp. 7024-7032, 2024.

[19] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 10051-10060.

[20] L. A. Gatys, A. S. Ecker, M. Bethge, et al., "Controlling perceptual factors in neural style transfer," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 3985-3993.

[21] R. Tian, Z. Wu, Q. Dai, H. Hu, Y. Qiao, and Y. Jiang, "ResFormer: Scaling ViTs with multi-resolution training," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 22721-22731.

[22] L. Beyer, P. Izmailov, A. Kolesnikov, M. Caron, S. Kornblith, X. Zhai, M. Minderer, M. Tschannen, I. Alabdulmohsin, and N. Houlsby, "FlexiViT: One model for all patch sizes," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 14496-14506.

[23] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 9036-9045.