CoCoDiff: Correspondence-Consistent Diffusion Model for Fine-grained Style Transfer

Wenbo Nie[1,2,*], Zixiang Li[1,2,*], Renshuai Tao[1,2,†], Bin Wu[1,2], Yunchao Wei[1,2], Yao Zhao[1,2]

[1] Institute of Information Science, Beijing Jiaotong University,

[2] Visual Intelligence + X International Joint Laboratory of the Ministry of Education

Motivation

Existing style transfer methods primarily rely on global style alignment and overlook fine-grained semantic correspondences, leading to style mismatches, texture distortions, and structural degradation in complex scenes. Meanwhile, the rich intermediate features and implicit correspondences in pretrained diffusion models remain underexplored, with limited constraints for maintaining semantic consistency.

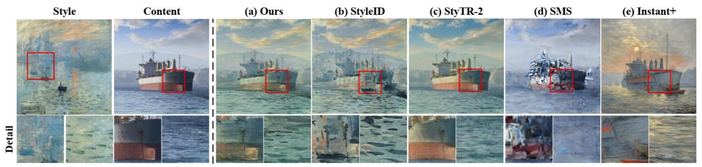

Figure 1: **Comparison of different style transfer methods.** We compare our proposed CoCoDiff with three other representative methods with zoomed-in details.

Motivated by these limitations, CoCoDiff leverages diffusion features to model fine-grained content–style correspondences and enforce consistency, enabling more precise and natural style transfer.

Methodlogy

Algorithm 1: Correspondence-Guided Style Transfer

Input: I_c: content image; I_s: style image; I_{sty_c}: style image with content features; f_θ: pretrained diffusion U-Net; $\mathcal{T}$: candidate timesteps; $\mathcal{L}$: candidate layers; w: injection weight; τ_c, τ_s: stopping thresholds; z: max iterations

Output: $I_{\text{gen}}^{(z^*)}$

// **Stage A: correspondence**

$I_{sty_c} \leftarrow \text{Attn}(Q_s, K_c, V_c)$ // equation 3

Extract $\{F_c^{t,l}\}, \{F_{sty_c}^{t,l}\}$ for $t \in \mathcal{T}$, $l \in \mathcal{L}$; $(t^*, l^*) = \arg\max \mathcal{M}(t,l)$ // equation 5

for each p_c: $p^*_{sty_c} = \arg\max_{p_{sty_c}} \cos\left(F_c^{t^*,l^*}(p_c), F_{sty_c}^{t^*,l^*}(p_{sty_c})\right)$ // equation 4, equation 6

// **Stage B: fitting & control**

for $z = 1$ **to** Z **do**

  $y_{cs} = \sigma(y_s)\frac{y_c - \mu(y_c)}{\sigma(y_c)} + \mu(y_s)$ // equation 11

  $feat[k] \leftarrow w \cdot attn[k][p^*_{\text{sty}_c}] + feat[k][p_c]$ // equation 8

  $\mathcal{L}_{content} = \|\text{Sobel}(I_{\text{gen}}^{(z)}) - \text{Sobel}(I_c)\|$, $\mathcal{L}_{style} = \sum_l \|G(I_{\text{gen}}^{(z),l}) - G(I_s)\|_F^2$ // equation 10

  if $\mathcal{L}_{content} > \tau_c$ ***and*** $\mathcal{L}_{style} < \tau_s$ **then**

    break

return $I_{\text{gen}}^{(z)}$

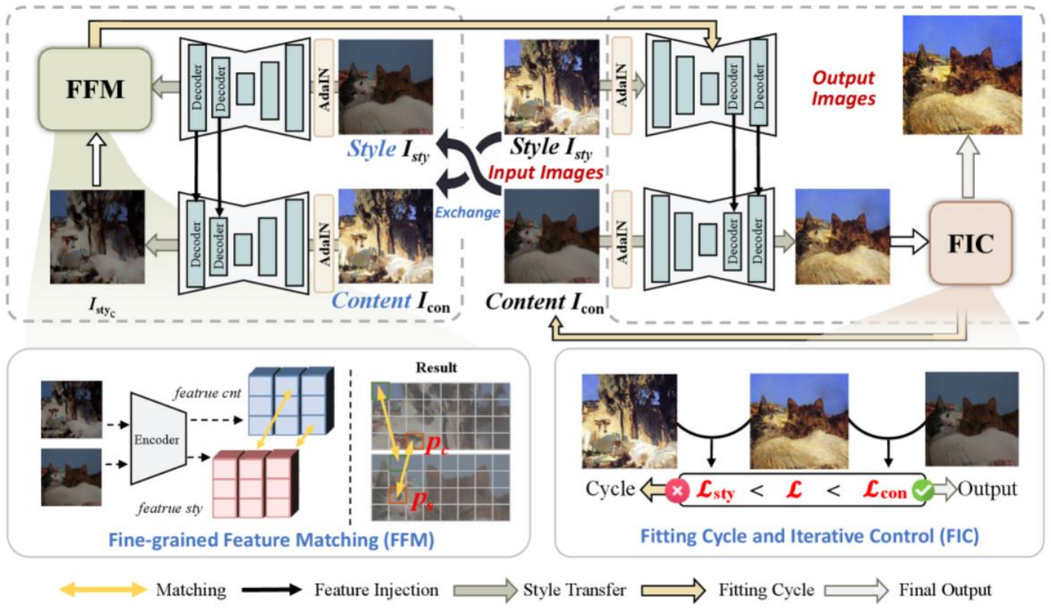

Figure 2: Framework Overview of our proposed Correspondence-Consistent Diffusion (CoCoDiff).

Experiment

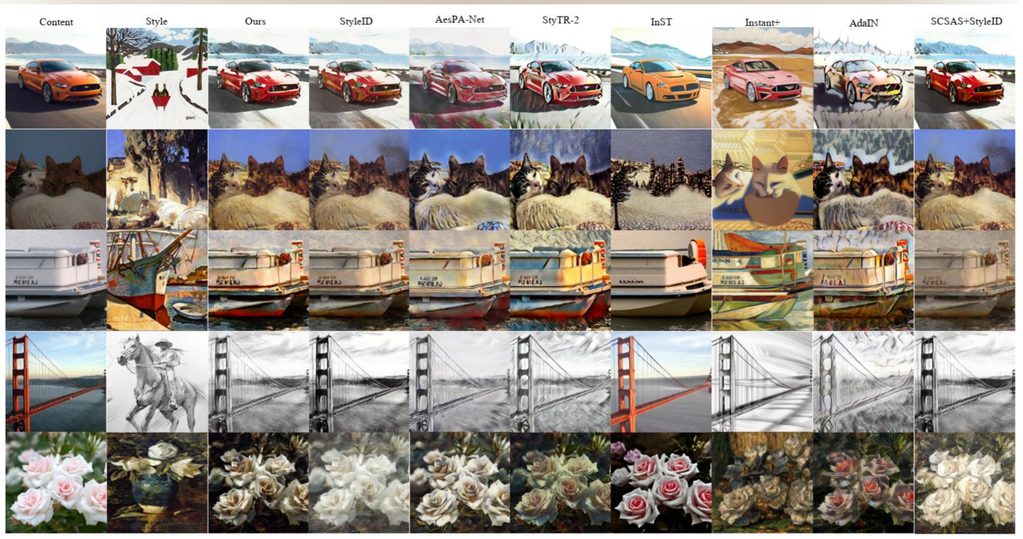

Figure 3: Qualitative comparison. We compare **CoCoDiff** (Ours) with seven representative methods, selected from diffusion-based, patch-based, CNN-based, transformer-based, and other approaches, to provide a comprehensive evaluation.

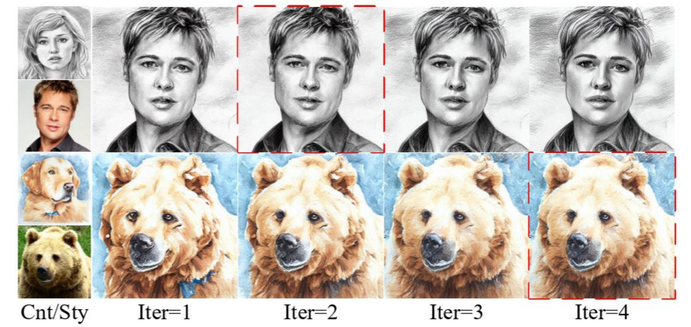

Figure 6: **Visual results of the iteration process.** The group with the best quality is highlighted by a red line.

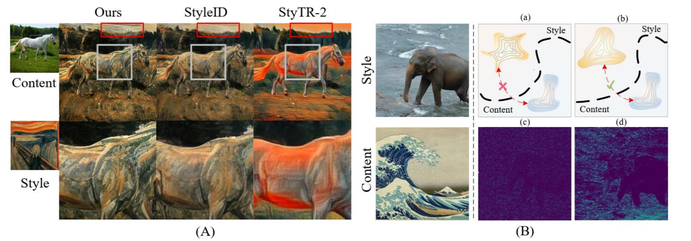

Figure 4: **(A) Qualitative comparison with additional zoomed-in details**. We compare our method with StyleID and StyTR^2 as baseline approaches, highlighting the differences through zoomed-in details. **(B) The illustration of cycle-based image style transfer.** (a) Direct feature matching between the style image and the content image often results in low matching accuracy and feature correspondence failure. (b) By first transforming the style image to adopt the content image's style before performing feature matching, the matching accuracy is significantly improved. (c) Direct correspondence result. (d) Indirect correspondence result.

Metric	**Ours**	StyleID	AesPA	StyTR^2	InST	Instant+	AdaIN	SCSA	FreeStyle	SMS
FID ↓	**18.432**	21.010	19.645	18.886	21.541	20.982	<u>18.672</u>	20.835	23.654	31.266
LPIPS ↓	**0.549**	0.565	<u>0.556</u>	0.587	0.785	0.584	0.612	0.562	0.689	0.821
ArtFID ↓	**30.100**	34.446	32.124	31.559	40.235	34.820	<u>31.711</u>	34.106	41.640	58.756
CFSD ↓	**0.609**	<u>0.619</u>	0.632	0.687	0.881	0.710	0.642	0.612	0.660	0.704

Table 1: **Quantitative evaluation results.** We compare every methods across multiple metrics. Columns 2^{nd}-9^{th} are reference-guided methods while columns 10^{th}-11^{th} are prompt-based methods.

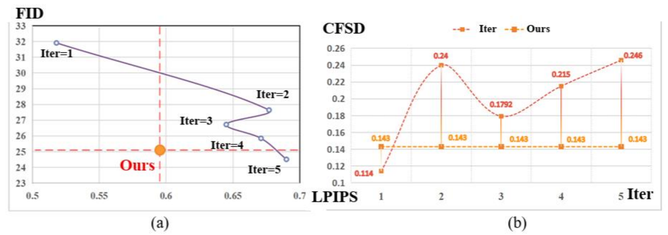

Figure 5: **Quantitative comparison of the cycle module.** (a) Balance between LPIPS and FID metrics across iterations. (b) CFSD variations across iterations.