Colour Style Generation for Digital Images: A Controllable Diffusion Model with Semantic Segmentation Constraints

Yingqiang Chen[1], Hongliang Du[1,*]
[1]School of Media, Weifang University, Shandong, Weifang 261021, China
Corresponding author: Hongliang Du (e-mail: 18265696603@163.com).

ABSTRACT Artificial intelligence generation techniques are increasingly reshaping digital image colour editing by enabling flexible and controllable style reconstruction. A controllable diffusion model with semantic segmentation constraints is proposed to address semantic mismatch, boundary colour bleeding, and insufficient regional control in global colour grading. Built on a latent diffusion framework, the model integrates semantic segmentation, multi-scale regional feature encoding, Lab-space colour statistics, style condition embedding, cross-attention injection, semantic-mask-guided denoising, and multi-objective consistency optimisation. Region-specific colour constraints are dynamically introduced during reverse diffusion to preserve object structure, coordinate local tones, and support continuous adjustment of style intensity, hue, and saturation. Experiments show that the proposed model achieves an FID of 23.7, 19.4% lower than ControlNet, while SSIM reaches 0.891. Compared with ControlNet, LPIPS decreases from 0.171 to 0.142 and colour error ΔE drops from 8.2 to 6.8. Semantic mIoU increases to 82.6%, boundary F1 reaches 87.4%, and style similarity improves to 0.913, confirming superior generation quality, semantic preservation, and controllability. Ablation results verify complementary contributions from segmentation, embedding, consistency, and boundary smoothing.

INDEX TERMS Digital images; Colour style generation; Semantic segmentation; Controllable diffusion models

I. INTRODUCTION

The development of digital image processing (DIP) has progressed from the traditional colour mapping and statistical matching to generative modelling through the advances of artificial intelligence and intelligent algorithm. Colour style generation is widely used in film and TV colour grading, digital art, cultural image restoration and visual content creation, because of its ability to change the colour tone, colour range and visual atmosphere of the image. However, most of the current methods are based on global colour constraints and cannot properly model the colour changes of semantic regions; this results in colour bleeding, semantic region mismatch and semantic region structure change at the figure-sky, vegetation-building, and figure-vegetation boundaries. To streamline the generation of high-resolution images, Rombach et al. [1] introduced a latent-space diffusion model that generates images using compressed representations, giving a basic framework to the complex process of visual image generation. To make diffusion models better able to control structural information like edges, depth and pose, Zhang et al. [2] added spatial constraints to diffusion models; but, they still haven' t been able to capture regional colour relationships well. Couairon et al. [3] used mask-guided editing for semantic image editing, showing that it is effective to control the region to be edited by using local constraints. Kwon and Ye [4] managed to transform images by separating style and content, which also increased the content preservation, but what is still missing is the explicit relationship between the semantic regions. Kawar et al. [5] and Brooks et al. [6] have added the flexibility of real-image editing from the text optimisation and instruction-driven viewpoints, respectively, but the resulting images are still prone to ambiguity of conditional expressions. To limit editing boundaries, Avrahami et al. [7] adopted a region-blending mechanism to improve the editing naturalness in the local area. Tarrés et al. [8] extended this idea to a parametric style control method, which can serve as a foundation for smoothly varying the style intensity; but the problem of different colour reactivities in the various semantic zones has not been completely solved.

In this paper, we therefore consider a controllable diffusion model, but with an additional semantic segmentation constraint and combine regional feature encoding, colour style condition embedding, semantic constraint denoising and colour consistency optimisation in one model. Experimental validation is performed with regard to the generation

quality, semantic preservation, style control and module contributions, and for the purpose of setting up a method for generation of digital image colour styles that is structurally stable, regionally coordinated and continuously adjustable.

II. MODELLING THE DIGITAL IMAGE COLOUR STYLE GENERATION TASK

A. DEFINITION OF THE COLOUR STYLE GENERATION PROBLEM

The generation of colour styles for digital images can be defined as the mapping of an input image to a generated image with a target colour style, whilst maintaining the stability of the original image's spatial structure, subject contours and semantic content [9]. Let the original image be I, the target style conditions be S, and the generated result be $\hat{I}$; the colour style generation process can then be represented as:

$$\hat{I} = G(I; S; \theta) \tag{1}$$

where $G(\cdot)$ denotes the controllable generative model, and θ represents the model parameters. The optimisation objective is to strike a balance between content preservation and style transfer, namely:

$$\min_{\theta} L = \lambda_c L_{content} + \lambda_s L_{style} + \lambda_r L_{region} \tag{2}$$

Here, $L_{content}$ is used to constrain the consistency of the image structure, L_{style} is used to measure the degree of matching with the target colour style, and L_{region} is used to limit colour shifts within different semantic regions [10]. The overall relationship between the input, constraints and output for this task is shown in Figure 1.

B. MODELLING THE ASSOCIATION BETWEEN SEMANTIC REGIONS AND COLOUR FEATURES

To mitigate the colour bias caused by global colour shift in local regions, the input image is segmented into semantic regions such as sky, vegetation, buildings, people and ground, and a correspondence is established between region categories and colour statistical features [11]. Let the semantic segmentation result be denoted as $M = \{M_k\}_{k=1}^{K}$, where M_k represents the region mask for class k, and K denotes the number of semantic categories. The colour features of each region are described by the mean, standard deviation and colour histogram:

$$F_k = [\mu_k, \sigma_k, H_k] \tag{3}$$

where μ_k and σ_k denote the mean and standard deviation of the region in the Lab colour space, respectively, and H_k is the normalised colour histogram [12]. A semantic–colour mapping relationship is further established as follows:

$$C_k = \Phi(E_k, F_k) \tag{4}$$

where E_k represents the semantic embedding, $\Phi(\cdot)$ is the feature fusion function, and C_k is the region-specific condition vector [13]. This modelling approach enables different semantic regions to be subject to distinct colour constraints, whilst minimising cross-region colour contamination; the specific association process is illustrated in Figure 2.

C. DESCRIPTION OF CONSTRAINTS FOR CONTROLLABLE GENERATION TASKS

To generate controllable colour style, stable constraints between the content structure, semantic regions and target style need to be established. In the generation process, the contours, textural hierarchy, spatial layout of the original image should be basically consistent, so as to prevent the change of the original image structure caused by style transfer. The Semantic Segmentation is used to frame color variations in the different areas such as the sky, vegetation, buildings and people to reduce colour bleedages near the boundaries and mismatches among the regions [14]. The generation direction is determined by the style intensity, hue bias and saturation while joint constraints are imposed according to semantic consistency, regional colour harmony and boundary continuity. This way, the produced image would have the style features of the target image and local regions would have natural transitions and visual plausibility.

III. CONSTRUCTION OF A CONTROLLABLE DIFFUSION GENERATION MODEL WITH SEMANTIC SEGMENTATION CONSTRAINTS

A. OVERALL MODEL ARCHITECTURE

The built model consists of five parts: semantic segmentation, regional feature encoding, colour condition embedding, diffusion-based denoising generation and consistency optimisation. Semantic information is extracted from the input image using a semantic segmentation to get a regional mask and multi-scale semantic information is extracted using a feature encoder; the target style is conditionally expressed by colour statistics and style vectors, which are fed into the denoising network together with the information of each time step. The diffusion module is then constrained by the semantic region constraints in the gradual restoration of the image, so the cross-region colour bleeding is suppressed. The output results are then processed together for the content preservation, style matching and regional consistency, and the stable structure, harmonious colours and adjustable intensity are achieved in the style generation. The whole system architecture is illustrated in Figure 3.

B. SEMANTIC SEGMENTATION AND REGIONAL FEATURE ENCODING OF DIGITAL IMAGES

The input image first passes through a pre-trained semantic segmentation network to obtain pixel-level class probabilities, and region labels are determined via the maximum response:

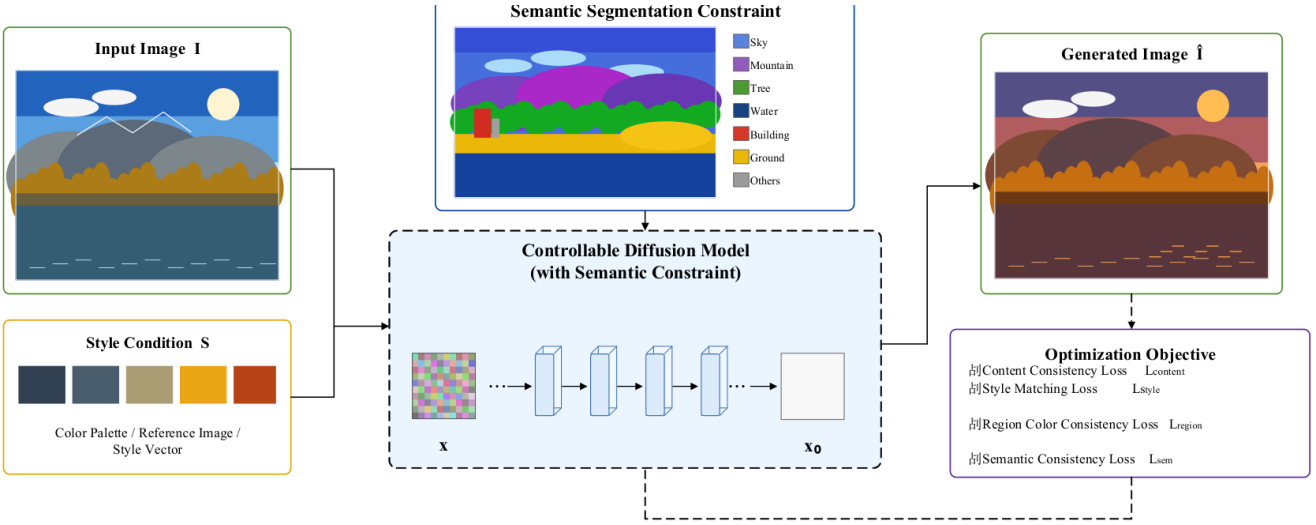

Fig. 1. Schematic Diagram of the Colour Generation Task.

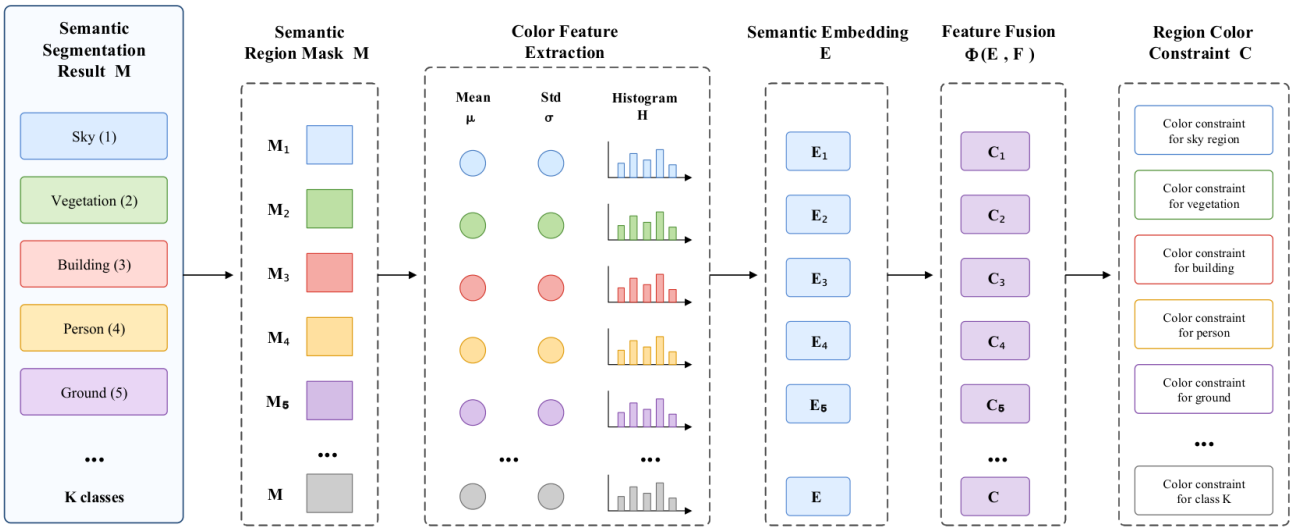

Fig. 2. Semantic Colour Association.

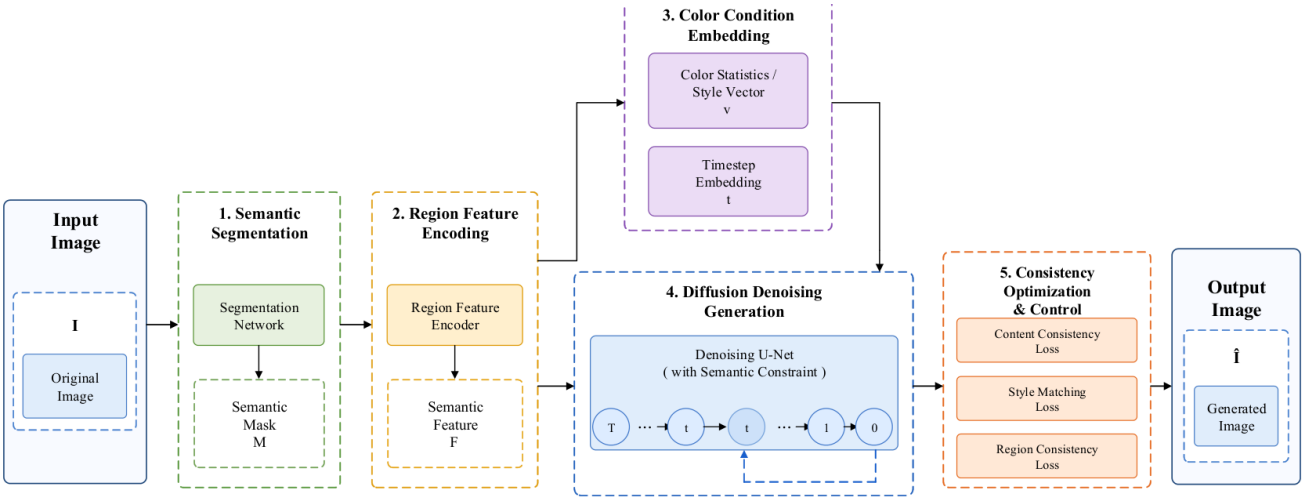

Fig. 3. Overall Architecture of the Model.

$$M(x,y) = \arg\max_k P_k(x,y \mid I) \tag{5}$$

where P_k is the probability that a pixel belongs to the semantic region of class k, and M is the segmentation mask [16].

Subsequently, a multi-scale convolutional encoder is utilised to extract texture, edge and spatial structural features, and the mask is downsampled to the corresponding scale to obtain region-constrained features:

$$F_k^{\ell} = \mathrm{Down}(M_k, \ell) \cdot E_{\ell}(I) \tag{6}$$

In particular, $E_{\ell}(I)$ represents the image features at the ℓ layer [17]. To enhance the stability of the region representation, mask-weighted pooling is performed on regions of the same type, and the results are fused with the class embeddings:

$$Z_k = \alpha \frac{\sum_{x,y} M_k(x,y) F_k(x,y)}{\sum_{x,y} M_k(x,y) + \varepsilon} + (1-\alpha) E_k \tag{7}$$

where α is the fusion coefficient between visual features and semantic embeddings, with a range of [0,1]. A grid search was performed within the range 0.1 – 0.9 with a step size of 0.1 and segmentation retention rate, SSIM and LPIPS were comprehensively compared. The value of α = 0.6 was chosen for the subsequent experiments as the regional structure is retained to the maximum level while the colour is controlled to the optimal level [18].

C. MECHANISM FOR EMBEDDING COLOUR STYLE CONDITIONS

The target colour style is characterised by a combination of reference images, colour palette parameters, or discrete style labels. First, the mean, standard deviation and histogram features in the Lab colour space are extracted, and a style vector is obtained via fully connected mapping:

$$v_s = \mathrm{MLP}([\mu_s, \sigma_s, H_s]) \tag{8}$$

The style vector, temporal step encoding and regional semantic features are jointly embedded as follows:

$$C_t = \mathrm{LN}(W_s v_s + W_t e_t + W_z Z_k) \tag{9}$$

where e_t denotes the time-step embedding, Z_k denotes the region features for the i-th class (k), and W_s, W_t and W_z are learnable projection matrices [19]. The fusion condition C_t is injected into the denoising network via cross-attention, enabling different semantic regions to undergo differentiated updates according to the target hue, brightness and saturation, thereby enhancing the accuracy of style response and local control capabilities. The conditional embedding structure is shown in Figure 4.

D. DESIGN OF THE SEMANTIC CONSTRAINED DIFFUSION GENERATION PROCESS

The diffusion generation process reconstructs images through forward noise addition and backward denoising. Let the original latent variable be x_0, and the noise-added result at step t be x_t; the forward process is expressed as:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon \tag{10}$$

where $\bar{\alpha}_t$ is the cumulative noise scheduling coefficient, and ε is Gaussian noise [20]. In the backward denoising stage, the semantic region features Z_k and the colour conditions C_t are fed into the noise prediction network:

$$\hat{\varepsilon}_t = \varepsilon_{\theta}(x_t, t, C_t, Z_k) \tag{11}$$

The semantic mask is employed to define the scope of the feature updates per region, controlling the sky, vegetation, buildings and people regions with their respective colour constraints and consistency constraints at the boundaries to reduce colour bleeding and semantic mismatches. Both the noise prediction error and the semantic region deviations slowly decrease with more denoising steps; and the model is converging faster and with smaller fluctuation if semantic constraints are used. The changes in error are displayed in Figure 5.

E. COLOUR CONSISTENCY OPTIMISATION AND GENERATION CONTROL STRATEGY

The training phase uses the content preservation, style matching, regional consistency and regional boundary smoothing constraints, which are used to guarantee the colour harmony between different semantic regions, with the weight values set at 0.30, 0.25, 0.30 and 0.15 respectively. On the side of generation control, the style intensity coefficient was 0.2~1.0 (step 0.2); the hue was adjusted by $\pm 15°$, and the saturation adjustment range was 0.8~1.2. The number of diffusion steps was defined to be 50, while the conditional guidance scale was set to 7.5 for a balance between style responsiveness and the structural stability. The values of the relevant parameters, and the reasons for them, are provided in Table 1.

IV. MODEL TRAINING AND OPTIMISATION METHODS

The model training data is divided into training, validation and test sets in a 7:2:1 ratio, and samples are augmented using random cropping, horizontal flipping, luminance perturbation and colour dithering. During the training phase, the backbone parameters of the semantic segmentation network are frozen, with the focus on optimising the regional feature encoder, conditional embedding module and diffusion denoising network [21].For a randomly sampled time step t, Gaussian noise is added to the image latent variable; the diffusion network predicts the noise based on semantic features and target colour conditions, with the basic loss expressed as:

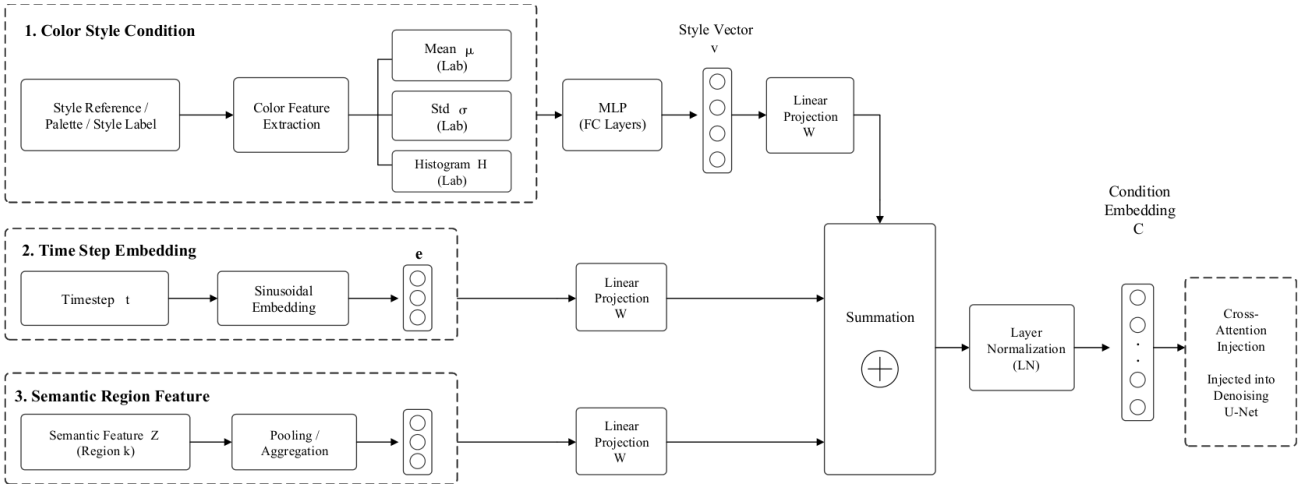

Fig. 4. Conditional Embedding Structure.

TABLE 1. Model Parameter Settings

Parameter Category	Parameter	Value	Basis for Setting
Input Configuration	Input image size	256 × 256 pixels	Balances semantic detail preservation and training efficiency
Semantic Segmentation	Number of semantic classes	19	Defined according to common scene categories in digital images
Feature Encoding	Semantic feature dimension	256	Ensures sufficient representation of regional features
Style Embedding	Style vector dimension	256	Maintains consistency with the semantic feature dimension
Diffusion Process	Number of diffusion steps	50	Balances generation quality and inference cost
Conditional Control	Classifier-free guidance scale	7.5	Determined through validation-set parameter search
Loss Configuration	Content preservation weight	0.30	Strengthens preservation of object structures and textures
	Style matching weight	0.25	Ensures effective response to the target color style
	Regional consistency weight	0.30	Suppresses color deviations within semantic regions
	Boundary smoothing weight	0.15	Reduces color bleeding across semantic boundaries
Generation Control	Style intensity range	0.2–1.0	Supports continuous adjustment from weak to strong styles
	Style adjustment interval	0.2	Facilitates graded control experiments
Color Control	Hue shift range	-15° to 15°	Avoids excessive color shifts and visual distortion
	Saturation coefficient	0.8–1.2	Preserves natural color appearance and tonal hierarchy

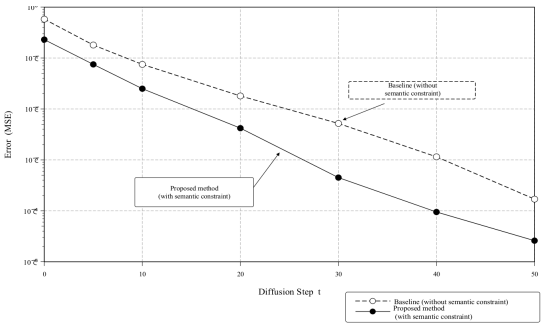

Fig. 5. Changes in Diffusion Error.

$$L_{diff} = \mathbb{E}_{x_0,t,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(x_t, t, C_t, Z)\|_2^2\right] \tag{12}$$

To balance structure preservation, style matchin g and regional colour stability, a multi-objective joint loss is employed to optimise the model:

$$L_{total} = \lambda_d L_{diff} + \lambda_c L_{content} + \lambda_s L_{style} + \lambda_r L_{region} + \lambda_b L_{boundary} \tag{13}$$

where the five weights are set to 1.00, 0.30, 0.25, 0.30 and 0.15 respectively [22]. This combination was determined via grid search on the validation set, ensuring that diffusion noise prediction takes centre stage in the optimisation whilst controlling local colour drift and semantic boundary infiltration. The model utilises the AdamW optimiser to update parameters:

$$\theta_{n+1} = \theta_n - \eta_n \frac{\hat{m}_n}{\sqrt{\hat{v}_n} + \varepsilon} \tag{14}$$

The initial learning rate is set to 1×10^{-4} and is gradually reduced using a cosine-annealing strategy to 1×10^{-6}. The number of training epochs is set to 200, with a batch size of 16; mixed-precision training, gradient clipping and exponential moving averages are employed to enhance convergence stability [23]. Training is stopped early if the validation doesn' t improve for 15 epochs and the model parameters that gave the best performance on the validation set are stored. The evolution of training and validation losses is shown in Figure 6 and the specific software and hardware environment and the training hyperparameters are shown in Table 2.

V. EXPERIMENTAL RESULTS AND ANALYSIS

TABLE 2. Training Configuration

Configuration Item	Setting
Training Framework	PyTorch 2.1
Input Resolution	256 × 256 pixels
Optimizer	AdamW
Initial Learning Rate	1×10^{-4}
Batch Size	16
Number of Epochs	200
Learning Rate Schedule	Cosine Annealing
Dataset Split	7:02:01
Mixed-Precision Training	FP16
Early-Stopping Patience	15 epochs

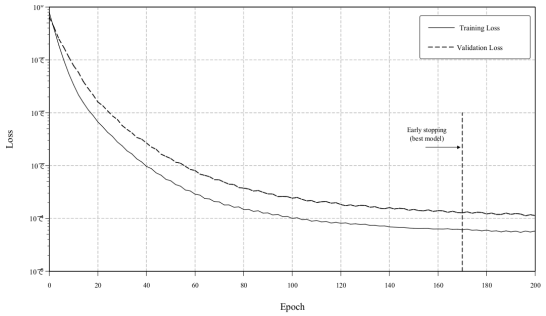

Fig. 6. Training Convergence Curve.

TABLE 3. Environment Configuration Table

Category	Item	Setting
Hardware Environment	Processor	Intel Core i9-13900K
	Memory	64 GB
	Graphics Card	NVIDIA RTX 4090
	GPU Memory	24 GB
Software Environment	Operating System	Ubuntu 22.04
	Deep Learning Framework	PyTorch 2.1
	CUDA Version	CUDA 12.1
Experimental Setting	Input Image Size	256 × 256 pixels
	Dataset Split	7:02:01
	Batch Size	16
	Maximum Number of Epochs	200
	Number of Inference Diffusion Steps	50

A. EXPERIMENTAL ENVIRONMENT CONFIGURATION

The experiments were run on the Ubuntu 22.04 operating system with the following hardware configuration: Intel Core i9-13900K processor, 64 GB of memory, and a video card with 24 GB of memory, NVIDIA RTX 4090. The model was implemented in PyTorch 2.1 and CUDA 12.1, with all images uniformly resized to 256×256 pixels. The data were divided into training, validation and test sets in the ratio 7:2:1 and the same pre-processing procedure was implemented to ensure the experimental fairness. The training batch size was 16, the max training iterations was 200, and the diffusion steps used in the inference phase was 50. The main experimental configurations, software and hardware environment are presented in Table 3.

B. DESIGN OF COMPARATIVE EXPERIMENTS

Three technical approaches were chosen for comparison: AdaIN, WCT, CycleGAN, Stable Diffusion and ControlNet.

TABLE 4. Comparison of Generation Results

Method	FID ↓	SSIM ↑	LPIPS ↓	ΔE ↓
AdaIN	46.8	0.721	0.312	14.8
WCT	43.2	0.746	0.286	13.6
CycleGAN	38.5	0.781	0.243	11.9
Stable Diffusion	32.6	0.833	0.198	9.7
ControlNet	29.4	0.861	0.171	8.2
Proposed Model	23.7	0.891	0.142	6.8

The same training and test data sets, the same input resolution and data pre-processing flows and the same hardware environment were used for training and inference for all models. The measures used in the evaluation were FID, SSIM, LPIPS, ΔE, semantic mIoU and style similarity, corresponding to the metrics of generative realism, structural preservation, perceptual difference, colour deviation, semantic stability and the level of style matching, respectively. This process was repeated five times to average out the fluctuations that might occur due to random initialisation and the result was averaged.

C. ANALYSIS OF COLOUR GENERATION QUALITY

The performance of different methods in terms of colour authenticity, structural preservation and perceptual quality was evaluated using FID, SSIM, LPIPS and ΔE as metrics, with comparisons carried out under a unified test set and inference configuration; the specific results are shown in Table 4.

As shown in Table 4, the proposed model has an FID of 23.7, which is 19.4% lower than that of ControlNet (29.4), indicating that the generated results are more similar to the distribution of real images, while the SSIM of 0.891 is 7.0% higher than that of Stable Diffusion (0.833), which means the generated images can be more stable in terms of subject structure and texture. LPIPS, meanwhile, dropped from 0.171 for ControlNet to 0.142, and ΔE dipped from 8.2 to 6.8, a 17.0% and 17.1% improvement, respectively. They show that semantic region constraints can be used to decrease colour bleed in nearby regions and unnatural colour changes, improving the colour harmony of the images and making them look more realistic.

D. VALIDATION OF SEMANTIC PRESERVATION CAPABILITY

The role of semantic segmentation constraints in preserving the content structure and region boundaries of images was evaluated using semantic mIoU, boundary F1 score and region retention rate. The semantic stability before and after generation by each method was also compared; the results are shown in Table 5.

Table 5 demonstrates the semantic mIoU value of the proposed model is 82.6 per cent, 4.7 per cent higher than ControlNet (77.9 per cent), meaning that class information of each semantic region is well preserved during the generation process. The boundary F1 score improved by 4.2 from 83.2% to 87.4% and the region retention rate improved

TABLE 5. Comparison of Semantic Capabilities

Method	Semantic mIoU (%) ↑	Boundary F1 (%) ↑	Region Retention (%) ↑
AdaIN	65.3	69.8	74.6
WCT	67.9	72.1	76.8
CycleGAN	71.6	76.4	81.3
Stable Diffusion	74.8	79.7	84.5
ControlNet	77.9	83.2	87.6
Proposed Model	82.6	87.4	91.2

TABLE 6. Comparison of Style Control Performance

Method	Style Similarity ↑	Control Linearity ↑	Regional Color Error ↓	Stability (%) ↑
AdaIN	0.741	0.762	13.2	76.8
WCT	0.758	0.781	12.4	78.5
CycleGAN	0.803	0.814	10.3	82.1
Stable Diffusion	0.852	0.867	8.6	86.7
ControlNet	0.879	0.901	7.4	89.3
Proposed Model	0.913	0.941	5.9	92.4

by 3.6 percentage points from 87.6% to 91.2%. Compared to AdaIN, the semantic mIoU of the proposed model was improved by 17.3 percentage points, which demonstrates that region masking and multi-scale feature encoding can greatly mitigate the colour blending and semantic drift at the boundaries of people, buildings and the sky.

E. ANALYSIS OF STYLE CONTROL PERFORMANCE

To evaluate the model's responsiveness under different colour styles and control intensities, we compared style similarity, control linearity, regional colour error and stability in repeated generation; the relevant experimental results are shown in Table 6.

As seen in Table 6, the proposed model achieves a high score of 0.913, which is a 3.9 per cent improvement over the score of ControlNet model (0.879), and shows that the target colour style can be well embedded in the generation process. With this, the control linearity is 0.941 which is higher than the 0.901 result for ControlNet, indicating a more continuous relationship between the control and the response as the style intensity goes up from 0.2 to 1.0. Regional colour error has dropped from 7.4 to 5.9, down 20.3 per cent and repeatability stability has risen to 92.4 per cent. The results show a better local control accuracy, and the avoidance of over-colouring of regions due to global colour adjustment, through embedding the semantic features and the style conditions jointly.

F. ABLATION EXPERIMENT ANALYSIS

To investigate the actual contributions of each component module to model performance, we removed the semantic segmentation constraint, the regional colour consistency loss, the style condition embedding and the boundary smoothing constraint one by one, whilst keeping all other training conditions consistent. The results are shown in Table 7.

As can be seen in Table 7, the removal of the semantic segmentation constraint resulted in an increase in FID from 23.7 to 31.8, whereas the semantic mIoU decreased by 8.5 points from 82.6% to 74.1%, showing that the semantic segmentation constraint is important to maintaining the stability of the regional structure. After the region consistency loss was removed, ΔE went up from 6.8 to 9.1, which is a 33.8% increase of colour error and when the style conditional embedding was removed, the style similarity reduced from 0.913 to 0.842. With the omission of the boundary smoothing constraint, FID increased to 26.7. The entire model showed the best result in all measures and confirmed the synergic optimisation effect of the different modules.

TABLE 7. Comparison of Ablation Experiments

Model Variant	FID ↓	Semantic mIoU (%) ↑	ΔE ↓	Style Similarity ↑
Without Semantic Constraint	31.8	74.1	9.6	0.874
Without Regional Consistency Loss	28.9	79.5	9.1	0.886
Without Style Condition Embedding	30.4	80.1	8.4	0.842
Without Boundary Smoothing	26.7	80.8	7.7	0.901
Full Model	23.7	82.6	6.8	0.913

VI. CONCLUSIONS

In the digital image colour style generation task, we encounter problems like local colour mismatch, semantic boundary infiltration and insufficient control accuracy, so we build a controllable diffusion generative model with semantic segmentation constraints to solve these problems. We have synergistically controlled the content-structure preservation and target style transfer through regional feature encoding, embedding colour style by semantic constraint, and colour consistency optimisation. Experiment results demonstrate that the model' s FID was lowered to 23.7, SSIM was raised to 0.891, semantic mIoU was increased to 82.6% and style similarity was increased to 0.913, which validates the effectiveness of semantic priors in improving generative realism, regional stability and style responsiveness. The key novelty is on one hand the simultaneous modelling of semantic regions and colour conditions, and on the other hand the dynamic injection of the regionlevel constraints in the diffusion process. Because of the possible inaccuracy of the segmentation network, the number of semantic classes and the amount of calculation required for high-resolution inference, colour deviations may occur in the complex boundaries and small target areas. In the future, the method might be extended with weakly-supervised semantic parsing, multi-scale dynamic control and lightweight sampling mechanism and also be extended to video colour generation to improve the adaptability to complex scenes and the temporal consistency.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, et al., "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF

conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
[2] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 3836–3847.
[3] G. Couairon, J. Verbeek, H. Schwenk, et al., "Diffedit: Diffusion-based semantic image editing with mask guidance," arXiv preprint arXiv:2210.11427, 2022.
[4] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," arXiv preprint arXiv:2209.15264, 2022.
[5] B. Kawar, S. Zada, O. Lang, et al., "Imagic: Text-based real image editing with diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 6007–6017.
[6] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 18392–18402.
[7] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18208–18218.
[8] G. C. Tarrés, D. Ruta, T. Bui, et al., "Parasol: Parametric style control for diffusion image synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2432–2442.
[9] S. Huang, Q. Li, J. Liao, et al., "Controllable image synthesis methods, applications and challenges: a comprehensive survey," Artificial Intelligence Review, vol. 57, no. 12, p. 336, 2024.
[10] T. T. Nguyen-Quynh, S. H. Kim, and N. T. do, "Image colorization using the global scene-context style and pixel-wise semantic segmentation," IEEE Access, vol. 8, pp. 214098–214114, 2020.
[11] O. Tasar, S. L. Happy, Y. Tarabalka, et al., "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp. 7178–7193, 2020.
[12] Y. Yu, D. Li, B. Li, et al., "Multi-style image generation based on semantic image," The Visual Computer, vol. 40, no. 5, pp. 3411–3426, 2024.
[13] L. Qi, L. Yang, W. Guo, et al., "Unigs: Unified representation for image generation and segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6305–6315.
[14] W. Luo, S. Yang, X. Zhang, et al., "Siedob: Semantic image editing by disentangling object and background," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1868–1878.
[15] D. Pavllo, A. Lucchi, and T. Hofmann, "Controlling style and semantics in weakly-supervised image generation," in European conference on computer vision. Cham: Springer International Publishing, 2020, pp. 482–499.
[16] Y. Jia, L. Hoyer, S. Huang, et al., "Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control," in European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024, pp. 91–109.
[17] Z. Lin, Z. Wang, H. Chen, et al., "Image style transfer algorithm based on semantic segmentation," IEEE Access, vol. 9, pp. 54518–54529, 2021.
[18] Y. Zhao, Z. Zhong, N. Zhao, et al., "Style-hallucinated dual consistency learning for domain generalized semantic segmentation," in European conference on computer vision. Cham: Springer Nature Switzerland, 2022, pp. 535–552.
[19] X. Gong, S. Chen, B. Zhang, et al., "Style consistent image generation for nuclei instance segmentation," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 3994–4003.
[20] Y. Yu, C. Wang, Q. Fu, et al., "Techniques and challenges of image segmentation: A review," Electronics, vol. 12, no. 5, p. 1199, 2023.
[21] A. Satchidanandam, R. M. S. Al Ansari, A. L. Sreenivasulu, et al., "Enhancing style transfer with GANs: perceptual loss and semantic segmentation," Int. J. Adv. Comput. Sci. Appl, vol. 11, pp. 321–329, 2023.
[22] T. Chen, Y. Yao, X. Huang, et al., "Spatial structure constraints for weakly supervised semantic segmentation," IEEE Transactions on Image Processing, vol. 33, pp. 1136–1148, 2024.
[23] K. Psychogyios, H. C. Leligou, F. Melissari, et al., "Samstyler: Enhancing visual creativity with neural style transfer and segment anything model (sam)," IEEe Access, vol. 11, pp. 100256–100267, 2023.