International Journal of Geo-Information

MDPI

Article

Diffusion Model-Based Cartoon Style Transfer for Real-World 3D Scenes

Yuhang Chen [1], Haoran Zhou [2,*], Jing Chen [1], Nai Yang [1], Jing Zhao [3] and Yi Chao [1]

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China; cyhang@cug.edu.cn (Y.C.); 20201003695@cug.edu.cn (J.C.); yangnai@cug.edu.cn (N.Y.); chaoyi@cug.edu.cn (Y.C.)
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[3] Provincial Surveying and Mapping Production Archives of Hubei, Wuhan 430074, China; xfhybxk@sina.cn
* Correspondence: zhouhr@aircas.ac.cn

Abstract

Traditional map style transfer methods are mostly based on GAN, which are either overly artistic at the expense of conveying information, or insufficiently aesthetic by simply changing the color scheme of the map image. These methods often struggle to balance style transfer with semantic preservation and lack consistency in their transfer effects. In recent years, diffusion models have made significant progress in the field of image processing and have shown great potential in image-style transfer tasks. Inspired by these advances, this paper presents a method for transferring real-world 3D scenes to a cartoon style without the need for additional input condition guidance. The method combines pre-trained LDM with LoRA models to achieve stable and high-quality style infusion. By integrating DDIM Inversion, ControlNet, and MultiDiffusion strategies, it achieves the cartoon style transfer of real-world 3D scenes through initial noise control, detail redrawing, and global coordination. Qualitative and quantitative analyses, as well as user studies, indicate that our method effectively injects a cartoon style while preserving the semantic content of the real-world 3D scene, maintaining a high degree of consistency in style transfer. This paper offers a new perspective for map style transfer.

Keywords: LDM; geographic scenes; stylization; semantic preservation; image maps

check for updates

Academic Editors: Wolfgang Kainz and Florian Hruby

Received: 9 May 2025
Revised: 14 July 2025
Accepted: 28 July 2025
Published: 4 August 2025

Citation: Chen, Y.; Zhou, H.; Chen, J.; Yang, N.; Zhao, J.; Chao, Y. Diffusion Model-Based Cartoon Style Transfer for Real-World 3D Scenes. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 303. https://doi.org/10.3390/ijgi14080303

Copyright: © 2025 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background and Challenges in Map Style Transfer

With the rapid development of geographic information technology, map style, as a collection of cartographic design features with aesthetic cohesion and uniqueness [1], plays a crucial role in the visualization process. Cartography is undergoing a profound transformation from functional descriptions to an integration of science, art, and emotional expression [2]. In this context, map style transfer, as an innovative means of map visualization aimed at migrating the visual style of a reference image to a target map, has become a research hotspot in the field of cartography.

In the early stages, map style transfer relied on the expert experience of cartographers for interactive style migration [3,4]. With the advancement of machine learning technology, mainstream research has shifted towards techniques based on convolutional neural networks (CNN) and generative adversarial networks (GAN) [4–6]. These methods have achieved the extraction of style elements from artworks for application to standard maps

or remote-sensing imagery, or style transfer between the two [4,5], such as converting remote-sensing imagery into standard map styles [7–9] or reproducing the unique charm of historical maps [10]. However, mainstream GAN technologies suffer from inherent limitations such as difficult training, uncontrollable generation, and instability, which make it challenging to ensure style consistency across different scenes and scales and constitute a major technical bottleneck in this domain.

In recent years, the emergence of diffusion models (DM) [11,12] has offered a new opportunity to address these challenges. These models learn a reverse process of gradual denoising from pure noise. This method not only maintains high-quality generation but also significantly reduces computational costs, outperforming GAN in terms of the quality, diversity, and stability of the generated images. Latent diffusion models (LDM) are a key optimization and concrete implementation of traditional DM and have shown tremendous potential in style transfer tasks. Nevertheless, current LDM-based style transfer methods often rely on complex text prompts or require tedious optimization adjustments for different inputs, which makes it difficult for them to guarantee both the accuracy of style injection and the integrity of content preservation in an automated, prompt-free workflow.

Currently, the field of map style transfer faces a trio of intertwined challenges. The primary challenge lies in the inherent conflict between "artistic expression" and "semantic preservation", where existing technical paths often fall into a dilemma: one class of methods pursues strong artistic effects at the cost of the clarity of key geographic information, while the other preserves content but only achieves superficial color transformations, failing to capture the essence of complex styles. Secondly, there remains a reliability bottleneck at the technical implementation level. GAN, due to the inherent instability of their training process, lead to uncontrollable and difficult-to-reproduce results. While emerging LDM have achieved breakthroughs in quality, they are often overly reliant on complex external text prompts, which makes both approaches unable to meet the stringent demands for robustness in automated application scenarios. The most critical research gap lies in the limitation of scene dimensionality. The existing work is almost entirely confined to 2D planar representations (such as standard maps or remote-sensing imagery), lacking specialized and effective high-fidelity style transfer frameworks for photorealistic 3D scenes, which can provide richer spatial structures and texture details.

Therefore, how to overcome these difficulties, find an ideal balance between preserving geographic semantics and injecting artistic style, and develop an automated, high-fidelity stylization framework for detail-rich photorealistic 3D scenes constitutes a core research gap in the map style transfer domain that urgently needs to be filled.

1.2. Our Work

Inspired by this, this paper proposes a DM-based framework for the cartoon style transfer of photorealistic 3D scenes. This paper selects the cartoon style as its target because cartoonization offers an artistic generalization of the complex real world through simplified geometric shapes, clear outlines, and vibrant color blocks. It transforms complex geographic scenes into easily understandable and visually appealing representations. This characteristic not only effectively reduces visual information redundancy but also preserves, and even emphasizes, the core structures and spatial relationships of geographic elements, which makes it an ideal choice for balancing information communication and artistic aesthetics in cartographic visualization.

We employ a pre-trained LDM as the foundational generation framework and integrate a pre-trained low-rank adaptation (LoRA) module to stably inject a specific cartoon style texture. We fuse this with ControlNet Tile to ensure high preservation of the input scene's structure and details. Furthermore, we innovatively combine the MultiDiffusion strategy to

coordinate the global style and utilize DDIM Inversion to generate structured initial noise. Together, these components ensure a high quality and high consistency in style injection. This achieves, without the need for any text prompts or other additional conditional inputs, the generation of an output image for any given region of a photorealistic 3D scene that both highly preserves the original scene's semantic content and structure and significantly embodies a vivid cartoon style, with highly consistent style effects. This provides new ideas and solutions for map style transfer.

This paper also validates the superiority of this method in terms of content preservation, style effectiveness, and user acceptance through detailed qualitative analysis, quantitative evaluation, and a user study.

2. Related Works

This section aims to review two core areas related to this research: general image style transfer methods and specific map style transfer methods. As mentioned in the Introduction, existing methods face inherent challenges in balancing artistic effects and information preservation. By reviewing the developmental trajectory of existing technologies, this review will not merely list previous works but will also delve into the technical evolution and key methods in these two fields, deeply analyzing the common limitations and specific bottlenecks in current research and thereby demonstrating the necessity and innovation of the DM-based photorealistic 3D-scene stylization framework proposed in this paper.

2.1. Image Style Transfer Methods

Style transfer is a key technique in the fields of image processing and computer vision, with the core task of imparting the visual style of one image onto another. Image style transfer technology has evolved rapidly from early manual texture synthesis techniques [13,14] to modern deep learning methods dominated by GAN and DM [15–17].

GAN, due to their outstanding image generation capabilities, were once widely used in style transfer tasks [18]. However, the training process of GAN is known for its "stability issues" and can be overly unconstrained and uncontrollable [19,20], often leading to mode collapse or semantic distortion, which reduces the reliability of GAN in tasks that require high fidelity and stable output.

In recent years, DM, with their high-quality generation and stable, controllable advantages, are gradually becoming the new paradigm in the field of style transfer, giving rise to a variety of technical paths. These paths can be broadly divided into two categories: those that require pre-training of the style and those that perform on-the-fly inference without pre-training.

The first category, requiring pre-training, can be further divided into fine-tuning-based and inversion-based methods. In the fine-tuning path [21], the framework of StyleDiffusion [22] includes a style transfer module that needs to be fine-tuned, while DiffStyler [23] cleverly uses LoRA, a lightweight fine-tuning technique, to encapsulate the style. In the inversion-based path [24], InST [25] "inverts" the artistic essence of a style image into an optimizable text embedding. Both fine-tuning and inversion methods require a time-consuming "preparation" process before stylization can be performed.

To pursue higher efficiency and flexibility, more recent research has begun to explore completely training-free and optimization-free on-the-fly style transfer routes. These methods require no prior style learning and dynamically complete the stylization at the time of inference, for example, by introducing a cross-image attention mechanism [26,27] or, like FreeStyle [28], relying solely on text-described models. From the time-consuming optimization of InST to the lightweight fine-tuning of DiffStyler and, finally, to the completely on-the-fly inference of FreeStyle, we see a technical gradient that constantly trades off

between efficiency, control precision, and versatility, collectively driving the advancement of DM in the field of style transfer.

2.2. Map Style Transfer Methods

Map style transfer is a technique for applying the visual style of a reference image to a target map. Early work relied on the interactive operations of cartographers, such as extracting color schemes from artworks and manually applying them to maps [3,29]. While these methods can produce high-quality maps, they are highly subjective and difficult to scale.

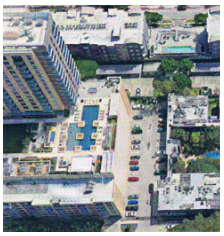
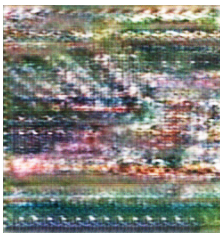
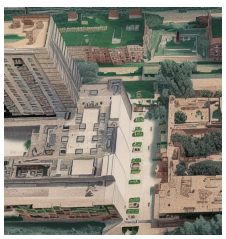
With the development of machine learning, automated map style transfer has become possible. Neural style transfer (NST) has been used to extract textures from artworks and apply them to OSM maps [5]. NST transfers artistic textures by matching statistical information across different feature layers. Its artistic effect is significant, but it comes at the cost of severely compromising the geometric accuracy of the content, causing key structures to become blurred, which makes it more suitable for purely artistic creation rather than functional cartography.

To better preserve the geographic content, research quickly shifted towards GAN. Among them, paired data methods, represented by Pix2Pix [30], use strictly paired training data (e.g., satellite imagery and map tiles) and strong supervisory signals to ensure a high degree of consistency in the content structure. These methods are particularly prevalent in well-defined tasks such as remote sensing-to-map conversion [8,9]. In contrast, unpaired data methods, represented by CycleGAN [18], introduce a cycle-consistency loss, revolutionizing the field by eliminating the dependence on paired data. This has greatly expanded the possible application scenarios; for instance, their use was pioneered to convert modern maps into historical map styles with high quality [10]. CycleGAN enables style transfer between domains that inherently cannot be paired, such as maps and paintings or modern and historical maps. However, the price of this freedom is a lower reliability in content preservation compared to Pix2Pix, which makes CycleGAN more prone to generating "hallucinations" or artifacts that do not conform to geographic reality. Its training process and results are also relatively less stable.

Despite the significant achievements of GAN-based methods, they consistently face challenges related to their inherent instability, mode collapse, and the difficulty of perfectly disentangling content from style. This often leads to the blurring or loss of key information such as map symbols and text and thereby undermines the map's ability to convey information. Although subsequent research has attempted to remedy these issues by introducing more sophisticated loss functions or optimizations for specific domains [10], the inherent flaws of the GAN framework have not been fundamentally solved. Consequently, some research has begun to explore highly customized non-GAN pathways. For example, high-fidelity color transfer for specific types of maps (such as vector maps and terrain maps) has been achieved by quantitatively modeling semantic relationships or designing novel color-organization methods [31,32].

The advent of DM has provided a breakthrough solution to the content preservation dilemma. The ControlNet technique [33], by introducing an additional, trainable network branch, achieves strong conditional control over pre-trained DM, allowing for the use of information such as edges and line art to precisely guide the structure of the generated image. This means that, by inputting the structural information of a map (such as road networks and building outlines), it is possible to freely inject any artistic style without compromising any geographic accuracy. In order to compare the advantages and disadvantages of the above methods more intuitively, we summarize them in Table 1.

Table 1. Analysis of the pros and cons of various methods.

Methodology	Artistic Effect	Content Preservation	Stability	Example Image of 3D Scene
Neural style transfer (NST)	High (Excellent at transferring textures, brushstrokes, and colors; strong artistic feel)	Low (Severely damages geometric structures; roads, text, etc., are difficult to recognize)	High (For a given input, the output is deterministic and consistent)	
Paired GAN (pix2pix)	Medium (More focused on mapping between two style systems rather than free artistic creation)	High (Strong supervisory signals ensure high consistency of content structure)	High (Stable training and inference processes; predictable results)	
Unpaired GAN (CycleGAN)	High (Enables transfer across different domains, e.g., maps and paintings; large creative space)	Medium (Cycle-consistency helps, but can still distort content and introduce artifacts)	Low (Less stable to train than pix2pix; can suffer from mode collapse; results have some randomness)	
Diffusion models (DM)	High (Controllable and high-quality injection of cartoon style; consistent style)	High (Strategies like ControlNet ensure the preservation of 3D geographic structure and semantic information)	Very High (Stable generation process, avoiding issues like mode collapse found in GAN)	

Although the aforementioned studies have achieved high-quality image style transfer using DM, they generally have limitations: they either require complex text prompts or need time-consuming optimization and adjustments for different inputs. Furthermore, these methods are inherently designed for 2D images. When applied to photorealistic 3D scenes, their intrinsic limitations become evident. Lacking an understanding of the scene's 3D geometry, their stylization process is more akin to "texture mapping" on a 2D plane, which makes it difficult to deeply integrate styles with complex building contours and spatial relationships. Consequently, when these methods are directly applied to 3D scene images rendered from different viewpoints, they often lead to severe cross-view style inconsistencies, the loss of key details, and unpleasant "flickering" artifacts during dynamic browsing. These shortcomings indicate that existing general-purpose style transfer techniques cannot meet the stringent requirements of 3D geographic scene visualization for structural fidelity and cross-view stability, which is precisely the core challenge this research aims to address.

To tackle the multiple challenges in map style transfer regarding artistic expression, semantic content preservation, and application scenario extension, this paper selects the powerful DM framework. For the first time, it systematically extends high-quality style transfer research to photorealistic 3D scenes, aiming to fill the current gap in 3D geographic spatial representation in this field.

3. Method

This chapter details a comprehensive methodology for cartoon-style conversion in real-world 3D environments that is structured into three components. The framework development section focuses on the technical implementation of our stylized 3D environment architecture. The intermediate output validation phase ensures quality control of the framework's generation, while the empirical evaluation component combines quantitative analysis with user research to substantiate our findings.

This framework is designed to process high-resolution geospatial imagery. To conduct comprehensive testing and validation, we developed a dataset comprising 60 images from diverse scenarios. The imagery originates from Google Earth, covering various landforms including landscapes, buildings, wastelands, and forests. The remote-sensing images span from 16 to 18 resolution levels to ensure rich detail and texture. For oblique photography data, we performed standardized preprocessing by uniformly cropping them into 1024 × 1024 pixel blocks to obtain standard input for the model.

3.1. Framework Development

This chapter will detail a comprehensive method that integrates initial noise control, detail repainting, and global coordination strategies to achieve the cartoon style transformation of remote-sensing images and real scene 3D scenes. The overall technical route is shown in Figure 1.

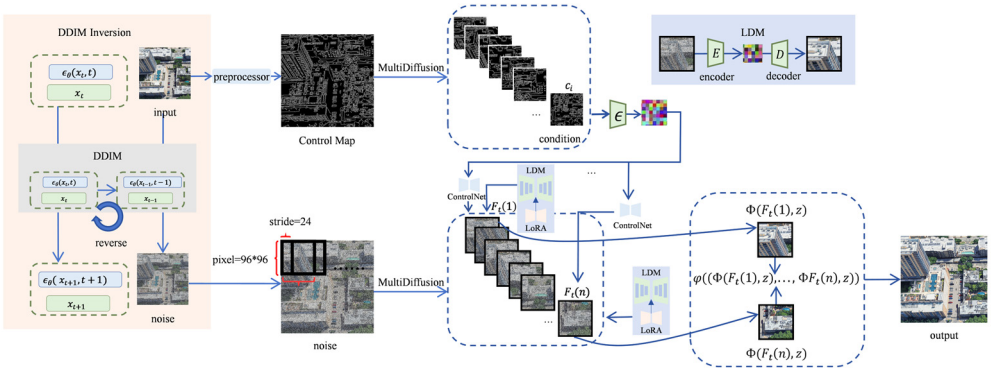


Figure 1. The overall technical approach.

The process begins with applying a noise-reversal strategy to the input image. The generated noisy image is divided into blocks using sliding windows within a defined overlap range, with each block being processed through the denoising model. Concurrently, the ControlNet Tile strategy is employed to implement additional control during denoising. The processed blocks are then reassembled based on their original pixel positions in the image, where overlapping pixels are weighted by their frequency of occurrence during the reassembly process. The reconstructed image is reused as input to repeat the denoising cycle until completion. The ControlNet Tile strategy corresponds to Section 3.1.3, while the noise-reversal strategy aligns with Section 3.1.5.

3.1.1. Latent Diffusion Models

Diffusion models represent a class of probabilistic generative models [11,34] that are trained to reverse the process of probability diffusion. Recently, diffusion models have garnered significant attention for their exceptional ability to generate high-quality images. Diffusion models consist of forward and reverse diffusion processes. In image generation, the forward diffusion process incrementally introduces random noise into an initial image

x_0, resulting in a noisy image x_t at a given time step t, which is a mixture of the original image x_0 and noise ϵ [11]:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon \tag{1}$$

where $\epsilon \sim N(0, I)$ and $\{\alpha_t\}$ are a sequence of noise levels that decreases as the time step increases. The backward process gradually removes noise from the initial noise image. During this process, a neural network $\epsilon_\theta(x_t, t)$ is typically trained to predict the added noise. To address this, latent diffusion models (LDM) [12] shift the diffusion process from the high-dimensional pixel space to a low-dimensional latent space. A a pre-trained variational autoencoder (VAE) [35] is also introduced to encode the image into a latent representation.

The VAE consists of an encoder E and a decoder D. The encoder E compresses the input image $x \in R^{H \times W \times 3}$ into a low-dimensional latent representation z:

$$z = E(x) \in R^{h \times w \times c} \tag{2}$$

where $h \ll H$, $w \ll W$. The decoder D reconstructs the image $\widetilde{x} = D(z)$ back into pixel space. By performing diffusion and denoising in this low-dimensional latent space, the computational complexity and memory requirements of LDM are significantly reduced, which makes it possible to generate high-resolution images with limited computing resources.

3.1.2. LoRA

Low-rank adaptation (LoRA) [36] is an efficient pre-trained model fine-tuning technique commonly used for large language models and diffusion models. It reduces the number of trainable parameters by learning a pair of rank-decomposition matrices while freezing the original weights. Its core idea is that the change in weights (update) during model fine-tuning has a low intrinsic rank. Therefore, we do not need to update the entire large weight matrix but only need to learn a low-rank update.

In standard model fine-tuning, a pre-trained weight matrix $W_0 \in R^{d \times k}$ is updated to $W = W_0 + \Delta W$, where ΔW is the weight change learned during training. The key insight of LoRA is that it does not directly optimize ΔW, which has $d \times k$ parameters, but approximates it via low-rank decomposition. Specifically, ΔW is represented by the product of two smaller matrices, A and B:

$$\Delta W = BA \tag{3}$$

where $B \in R^{d \times r}$, $A \in R^{r \times k}$, and the rank r represent a hyperparameter much smaller than d and k ($r \ll min(d, k)$).

During training, the original weights W_0 remain frozen, and only matrices A and B are trainable. This drastically reduces the number of trainable parameters from $d \times k$ to $r(d + k)$. In the model's forward pass, the modified layer's computation is as follows:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{4}$$

where x is the input and h is the output. At initialization, A is typically initialized from a Gaussian distribution, while B is initialized to zero. This ensures that $\Delta W = 00$ at the start of training, making the fine-tuned model identical to the pre-trained model, which guarantees training stability.

In our framework, we utilize a pre-trained LoRA module to encapsulate a specific "cartoon" art style. This module contains low-rank matrices A and B for key layers in the LDM (such as the cross-attention layers). By loading this LoRA module, we can stably and efficiently inject the specific style during the generation process without altering the weights of the base LDM and thereby fix the artistic style of the generated image.

3.1.3. ControlNet Tile

To highly preserve the structure and content of the input image, we introduce the ControlNet technique [33]. This technique can enhance the control over a pre-trained LDM by imposing additional control conditions. We use the original image as an additional control condition, in conjunction with the ControlNet Tile pre-trained model, to control the image generation.

Under the same hyperparameters, we decoded the generation results at the 10th time step for both the denoising process with ControlNet Tile and that without ControlNet Tile, as shown in Figure 2. Compared to group b, group a showed a significant improvement in preserving the structure and detailed content of the input image during the denoising process. Additionally, Stable Diffusion relies on the classifier-free guidance (CFG) technique to generate high-quality images. The formula for the CFG is as follows [37]:

$$\epsilon_{prd} = \epsilon_{uc} + \beta_{cfg}(\epsilon_c - \epsilon_{uc}) \tag{5}$$

where ϵ_{prd}, ϵ_{uc}, ϵ_c, and β_{cfg} refer to the model's final output, unconditional output, conditional output, and user-specified weight, respectively. When using ControlNet to add additional conditional control, it can be added to either ϵ_{uc} or ϵ_c. To maintain high-quality control capabilities during the denoising process without text prompts, we add ControlNet to both ϵ_{uc} and ϵ_c. As indicated by the CFG formula, when there are no text prompts, the conditional output is the same as the unconditional output, which completely eliminates the CFG's control guidance, leaving only the control of ControlNet. The comparison results are shown in Figure 3, where it can be seen that adding ControlNet to both ϵ_{uc} and ϵ_c highlights the detail content of the original image and results in generated images that are closer to the given style.

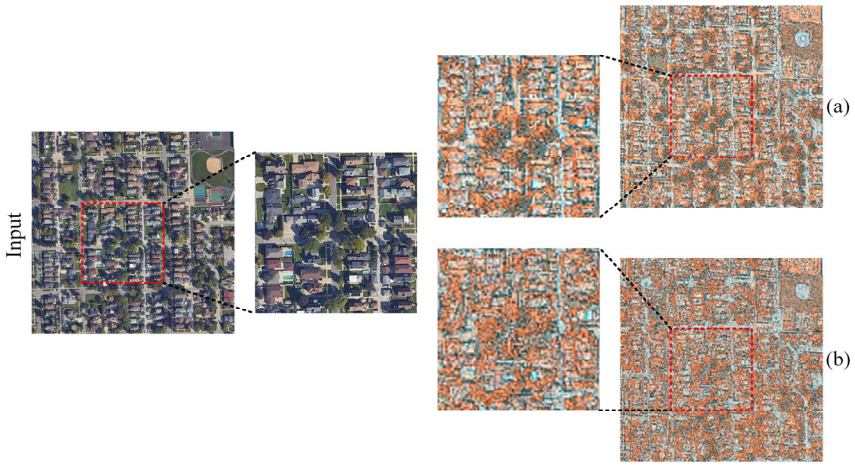


Figure 2. Comparison of 10-step denoising results between group a with the ControlNet Tile strategy introduced and group b without the ControlNet Tile strategy, with local effect images all magnified at 200%.

ControlNet adjusts the guidance strength by multiplying each connection in Stable Diffusion with a weight value w. The model's performance with different weight values is illustrated in Figure 4. As the weight value increases, the effect of detail redrawing becomes more pronounced. Lower weight values result in the loss of more detail information in the image, with lower color saturation and contrast, leading to less accurate color reproduction and a less distinct cartoon style presentation. In contrast, higher weight values significantly enhance the saturation and contrast of the image content, with details becoming too prominent, making the overall image appear too busy. After careful comparison and testing, we ultimately set the weight value to 0.85. At this value, the generated images

are able to effectively restore the overall semantic information of the input image while achieving a desirable style effect.



Figure 3. Comparison of control injection methods. Group a adds ControlNet control only to ϵ_c. Group b adds ControlNet control to both ϵ_{uc} and ϵ_c.

3.1.4. MultiDiffusion

Images generated with ControlNet Tile control maintain a high degree of similarity to the original images and effectively exhibit a distinct cartoon style. However, the model still alters numerous details in the original images. To further refine the details of the generated images and more effectively showcase the style transformation, we employ the MultiDiffusion strategy [38]. We divide the noisy image x_t into small segments using a sliding window approach, with each window sized at 96 × 96 pixels and a stride of 24 for the window movement. We record the pixel ranges (h_{start}, h_{end}, w_{start}, w_{end}) for each segment on x_t. If $x_t \in \mathbb{R}^{m \times n}$, the number of segments that x_t is divided into by the 96 × 96 window can be calculated using the following formula:

$$num = \left(\left\lfloor \frac{m-96}{stride} \right\rfloor + 1\right) \times \left(\left\lfloor \frac{n-96}{stride} \right\rfloor + 1\right) \quad (6)$$

The divided image blocks $\{F_t(1), \ldots, F_t(n)\}$ are each input into the pre-trained model for reverse denoising $\Phi(F_t(I), z)$. After denoising, the resulting blocks are reassembled according to their pixel ranges in x_t, with overlapping pixel values weighted by their

overlap count to obtain the final pixel values. This process yields x_{t-1}, and the above steps are repeated until the denoising is complete.

$$x_T,\ x_{T-1},\ \ldots,\ x_0\ s.t.\ x_{T-1} = \varphi(\ x_T) \tag{7}$$

The introduction of MultiDiffusion allows for the gradual adjustment and optimization of detail levels during the generation process, resulting in a more refined and closely matched representation of details to the original image. In Figure 5, we compare the output effects of images with and without the MultiDiffusion strategy. Specifically, in image ③, the intersection appears chaotic, with various traffic marking lines on the road differing significantly from the input image in terms of distribution and distorted forms. Additionally, the transitions between different ground features are not smooth, which leads to an overall abstract effect. In contrast, image ⑤ accurately represents the road intersection, with the marking lines on the road closely matching the input image. The transitions between elements are also more harmonious. Further examination of image ④ reveals that the structure of the buildings has been significantly distorted, which makes it impossible to identify the details of the building's sides. Moreover, the distribution of roads on both sides of the buildings is severely distorted, resulting in a significant difference in detail from the input image. In image ⑤, however, the outline structure and details of the buildings are well-preserved, with the transitions between the buildings and the surrounding roads being more natural and fluid, with a higher similarity to the input image.



Figure 4. Comparison of control weights. Group a applies a low weight of 0.4. Group b uses a standard weight of 1.0. Group c uses a high weight of 1.6.

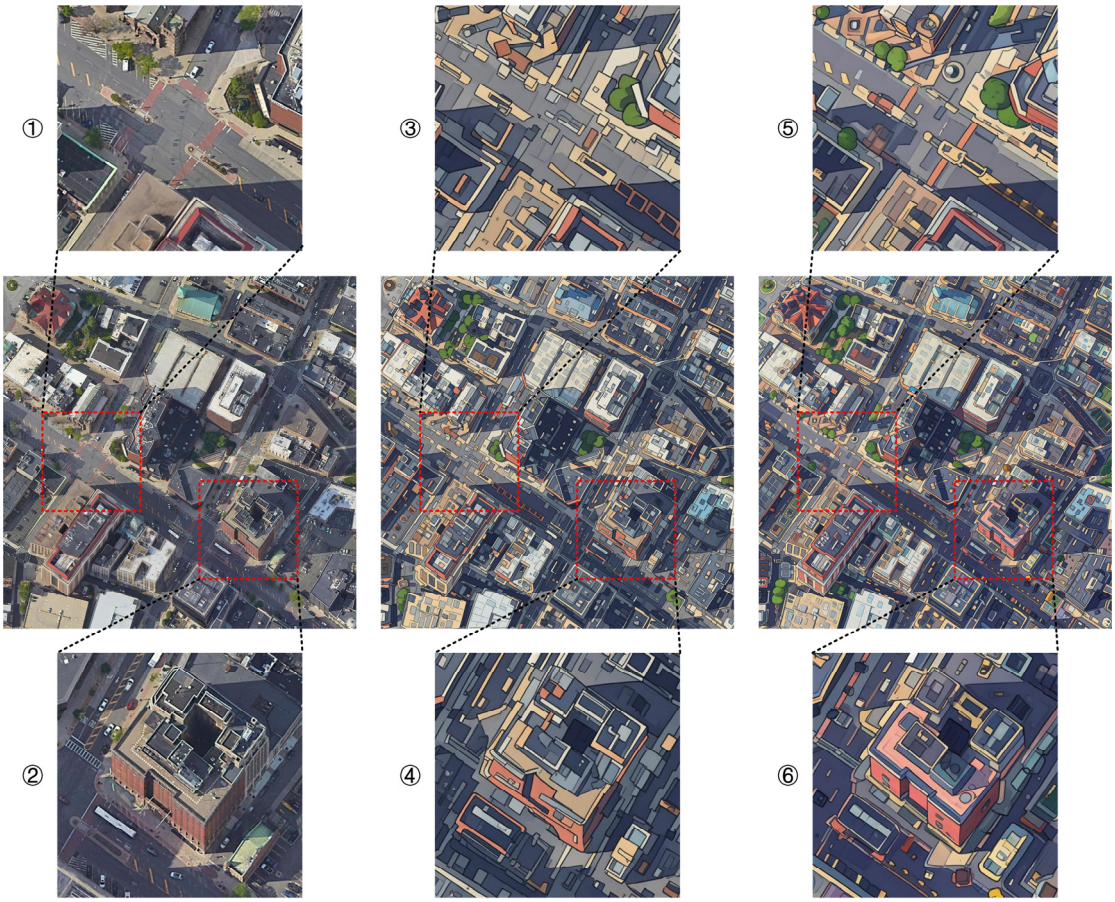


Figure 5. Effect comparison: group a without MultiDiffusion strategy, group b with MultiDiffusion strategy. The local effect images are all magnified at 250%.

In conclusion, the images with the introduced MultiDiffusion strategy better balance the style transfer effect and detail preservation quality of the generated images, bringing them closer to the input image and improving the overall quality and visual effect of the image generation.

3.1.5. Initial Noise Control

The images generated using the aforementioned method achieve decent results in preserving semantic content and style transfer, but they often exacerbate unnecessary details, which makes the generated images appear cluttered. To more stably control the generation outcomes of the model, we introduce the DDIM Inversion strategy. DDIM specifies how to obtain x_0 from x_t as outlined in the following formula, where $\sigma_t \epsilon_t$ represents the added random noise.

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta^{(t)}(x_t) + \sigma_t \epsilon_t \tag{8}$$

Inversion is commonly used in text-guided image editing [39]. Unlike the forward diffusion process that introduces random noise into the initial image, the inversion process involves finding an initial noise vector. When the prompt and this initial noise vector are input into the diffusion process, it generates the input image while preserving the model's editing capabilities. In DDIM Inversion, the goal is to reverse the sampling process. We aim to obtain an image filled with noise, which, when used as the starting point for the sampling process, generates the input image. This process can be modeled as how to obtain

x_t from x_0. To invert, we simply rewrite Formula (8) to solve for x_t's expression, setting σ to 0 to eliminate the random noise:

$$x_t = \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_t - 1}}\left(x_{t-1} - \sqrt{1-\alpha_{t-1}}\epsilon_\theta^{(t)}(x_t)\right) + \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x_t) \tag{9}$$

The initial noise image obtained through DDIM Inversion is more stable in preserving the semantic information of the input image compared to random initial noise images. As shown in Figure 6, the noise image obtained after DDIM Inversion serves as a better starting point for the sampling process that is more conducive to restoring the structure and details of input image. In the comparison, we can observe that image ⑤ is more precise in the reconstruction of open spaces and ground features. In image ③, the open spaces contain a large number of details that are not from the input image, and the re-rendering of containers and vehicles on the ground is excessive, resulting in an overall chaotic appearance. In contrast, image ⑤ is able to restore the various ground features in the input image with high quality, providing better visual quality. Similarly, in image ④, the elements on the ground have undergone significant changes, with there being many details that do not exist in the input image, which makes the image overall appear disordered. In image ⑥, the distinction between ground features is more pronounced, and the semantic content of the input image is more accurately restored, which results in a more fluid and harmonious overall image.

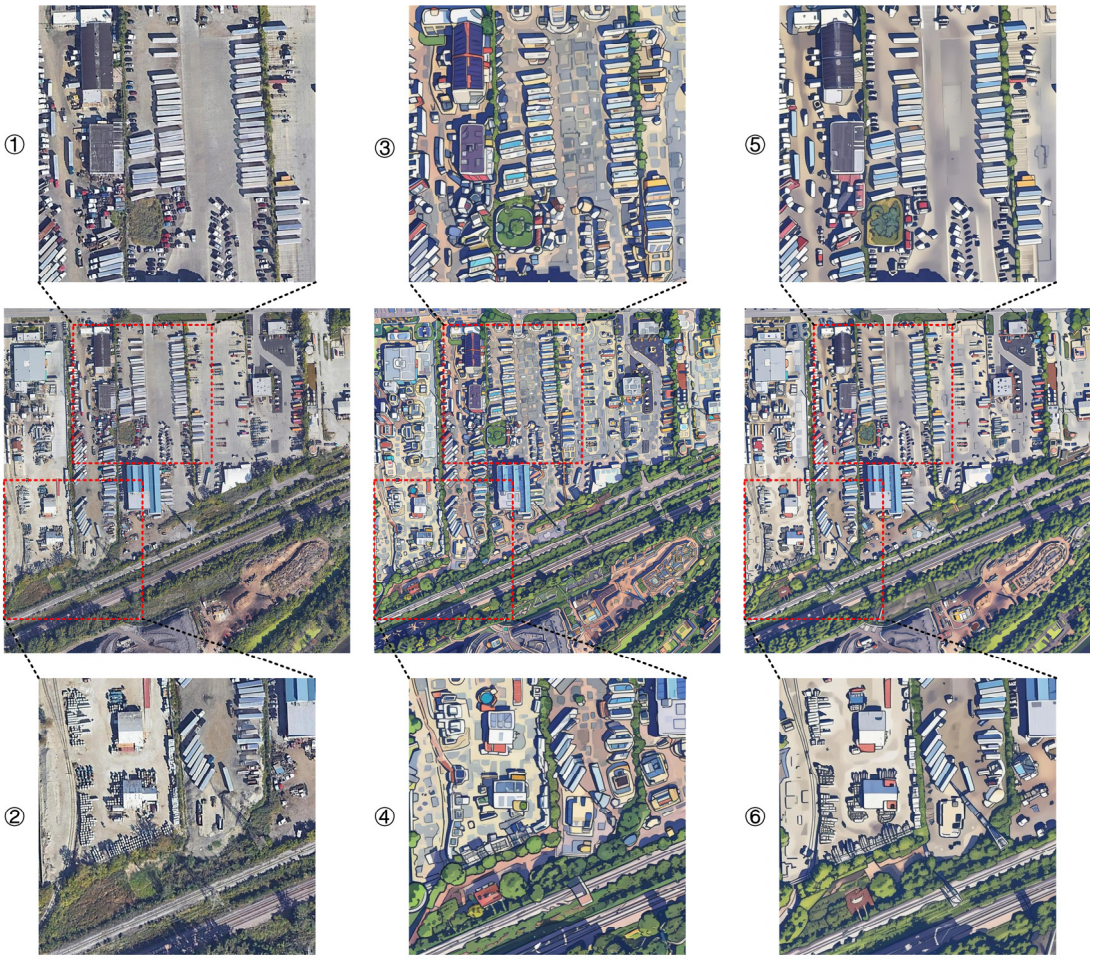


Figure 6. Effect comparison: group b uses the 10-step noise inversion result as the initial noise, group a does not incorporate the DDIM Inversion strategy, with both groups' local effect images magnified at 200%.

In summary, DDIM Inversion enhances the process of maintaining details and structure during image generation by offering an initial noise image that closely aligns with the semantic content of the input. This results in the production of higher-quality images that are more faithful to the original semantics.

3.2. Intermediate Output Validation

Our method is training-free, thus requiring no additional training data. The experiments were conducted on an NVIDIA RTX 4080 GPU(NVIDIA Corporation, Santa Clara, CA, USA). For a single 1024 × 1024 image, the average sampling time is approximately 41 s. We used fixed hyperparameters and the DDIM sampler, performing a total of 20 sampling steps for each image. The model is based on SD1.5, utilizing publicly available pre-trained models for inference. Based on the above methods, we developed a system for the cartoon style transfer of photorealistic 3D scenes, as shown in ① and ④ of Figure 7. Users can arbitrarily select a region on the screen, and the photorealistic 3D scene of that region will be sent as input to the model for cartoon style transfer. Once the conversion is complete, the corresponding cartoon-style effect for that region can be seen on the screen, as shown in ② and ⑤ of Figure 7. The model can also display a curtain effect, where dragging a curtain divider left and right on the screen, as shown in ③ and ⑥ of Figure 7, reveals the cartoon-stylized 3D scene on the left side of the curtain line and the original photorealistic 3D scene on the right.



Figure 7. Cartoon style transfer system is applied to two real world 3D scenes a and b respectively.

As shown in the stylized images in Figure 7, our method demonstrates remarkable consistency in style transfer. Whether processing different types of images or facing diverse scene contents, it can effectively achieve cartoon style transfer while maintaining semantic content. The experimental results show the following:

(1) Our proposed method, while preserving the semantic content of the image, can reasonably adjust elements such as shape, brushstrokes, lines, and colors to achieve an outstanding cartoon effect;
(2) When processing various complex and diverse scenes, our method can accurately maintain the integrity of the semantic content;
(3) Furthermore, when generating stylized images, our method can avoid any obvious adverse effects, such as halo effects, ensuring the clarity and visual quality of the image.

These advantages enable our method to demonstrate high-quality performance in the task of cartoon style transfer for photorealistic 3D scenes. It not only accurately expresses stylistic features but also effectively maintains the recognizability of the image content. In summary, our method possesses strong robustness, a high degree of content preservation, and accurate style expression. These characteristics make our method highly practical and broadly applicable in real-world scenarios.

3.3. Empirical Evaluation Protocol

To conduct a rigorous empirical evaluation of this framework, we designed and implemented a detailed questionnaire survey. A total of 33 participants were recruited for this survey, all of whom were university students or graduate students, with an age range of 19–28 years (mean age = 23.4 years, standard deviation = 2.1 years). This survey used a convenience sampling method to ensure diversity in the sample's background. Participant backgrounds covered geographic information science, computer science, art and design, and other humanities and social sciences. About half of the participants had knowledge related to geographic information or cartography, while the other half were users from non-specialized backgrounds. All participants volunteered to participate after giving informed consent.

The questionnaire consisted of basic participant information and specific evaluation tasks. The former included age, gender, and experience with remote-sensing imagery; the experimental materials used in this user study were all generated from a dataset we constructed ourselves. This dataset was sourced from Google Earth and included three different scales (large, medium, small) of photorealistic 3D scenes, corresponding to Google Earth's level 18, 17, and 16 imagery, respectively. All original oblique photogrammetry images were preprocessed and cropped into standardized 1024 $\times$ 1024 pixel tiles before being input into our model. To eliminate the potential impact of scale changes on the generated image results, we set up three experimental groups in the specific evaluation, corresponding to large-, medium-, and small-scale images. The large scale was defined as containing a small part of a community, such as a street and its surrounding scene; the medium scale was defined as being able to contain an entire community; and the small scale was defined as being able to contain a large area of a community. The comparison included three types of photorealistic 3D scenes: orthophoto view, 30° oblique view, and 60° oblique view. The generated cartoon-style photorealistic 3D scenes are shown in the stylized images in Table 2. The experimental tasks consisted of four sub-tasks that focused on content preservation, style transfer, and user preference, as listed in the fourth column of Table 2. Each sub-task was a multiple-choice question structured as a five-point Likert scale. The tasks for the three parts were identical. The specific evaluation tasks are shown in Table 2.

Table 2. Design of experiments.

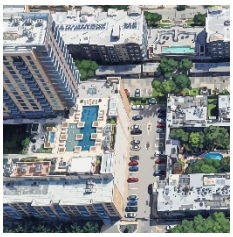
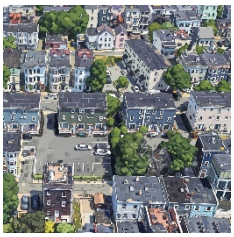
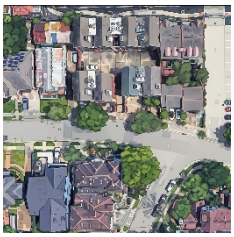
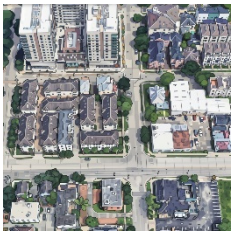
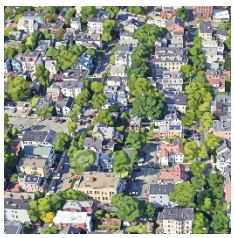
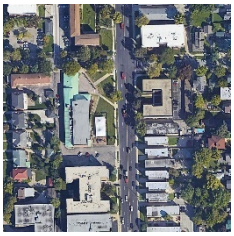
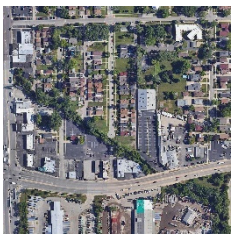
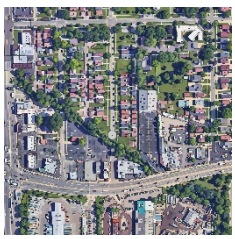
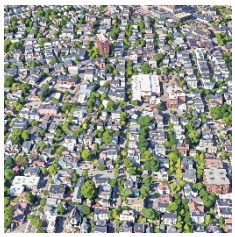
Scale		Original Scenes	Stylized Scenes	Tasks
Large	Group One			T1: The stylized scene presents a distinct cartoon style? T2: The stylized scene fully preserves the content of the original scene? T3: Compared to the original scene, is the level of information loss in the stylized scene acceptable to me? T4: I like this stylized scene?
	Group Two			
	Group Three			
Medium	Group One			
	Group Two			
	Group Three			

Table 2. *Cont.*

Scale		Original Scenes	Stylized Scenes	Tasks
Small	Group One			T1: The stylized scene presents a distinct cartoon style? T2: The stylized scene fully preserves the content of the original scene? T3: Compared to the original scene, is the level of information loss in the stylized scene acceptable to me? T4: I like this stylized scene?
	Group Two			
	Group Three			
T5: How does the above stylized scenes affect your perception?				

During the questionnaire survey, participants were first asked to fill in their basic information. They were then asked to browse three pairs of photorealistic 3D scenes at a certain scale, with unlimited browsing time. After finishing, they rated the 4 sub-tasks. After completing the ratings, they proceeded to the next part of the browsing. This process was repeated until all three browsing tasks were completed. After finishing all parts, participants were also asked to write down their impressions of the stylized images. In the end, we collected 33 questionnaires. We calculated the average score for each task and conducted statistical analysis using scale size as the independent variable and user scores as the dependent variable. We used a multi-paired sample Friedman test to explore the significance of differences in the scores for the four tasks under different scales, with a set significance level of 0.05. The statistical results are shown in Table 3.

As shown in Table 3, for tasks 3 and 4, there is a statistically significant difference in the scores given by participants across different scales. Further multiple comparison analysis for tasks 1 and 2 was conducted using the Nemenyi test, and the results are presented in Table 4. It is evident from the table that there is no statistically significant difference in the scores given by participants for the cartoon style presentation and semantic content preservation of the stylized scenes across different scales.

The above results yield the following conclusions:

(1) For task 1 (The stylized scene presents a distinct cartoon style?), the average score is 4.606, with no significant difference across different scales, which indicates that the stylized scene effectively presents the cartoon style under various scales;

(2) For task 2 (The stylized scene fully preserves the content of the original scene?), the average score is 4.394, with no significant difference across different scales, which indicates that the stylized scene can effectively preserve the original content;

(3) For task 3 (Compared to the original scene, is the level of information loss in the stylized scene acceptable to me?), the average score is 4.424, with no significant difference across different scales, which indicates that users generally find the level of information loss in the stylized scene acceptable;
(4) For task 4 (I like this stylized scene?), the average score is 4.475, which suggests that users generally favor this stylized scene, and that their preference is not influenced by the scale;
(5) We extracted keywords from all the impressions left by users, and the results show that most participants gave positive feedback such as "interesting", "realistic", "with a cartoon style", "attractive", and "beautiful and vivid" after completing the questionnaire. The feedback from users about the stylized scene is positive and affirmative.

Table 3. Statistical results of Friedman test for multiple matched samples of four tasks.

Task	Scale	Average Score	Overall Average	F	P
Task 1	Large	4.455	4.606	0.188	0.020
	Medium	4.697			
	Small	4.667			
Task 2	Large	4.515	4.394	0.163	0.018
	Medium	4.273			
	Small	4.394			
Task 3	Large	4.394	4.424	0.073	0.368
	Medium	4.485			
	Small	4.394			
Task 4	Large	4.424	4.475	0.056	0.558
	Medium	4.515			
	Small	4.485			

Table 4. Post-hoc multiple comparison analysis results for tasks 1 and 2.

Task	Scale Level		Mean Difference (I–J)	P
	Level I	Level J		
Task 1	Large	Medium	−0.242	0.401
	Large	Small	−0.212	0.510
	Small	Small	0.030	0.900
Task 2	Large	Medium	0.242	0.302
	Large	Small	0.121	0.721
	Small	Small	−0.121	0.721

3.4. Quantitative Evaluation

To better assess the content preservation of the cartoon-style photorealistic 3D scenes, we used several widely recognized evaluation metrics for quantitative analysis. The LPIPS (Learned Perceptual Image Patch Similarity) measures the perceptual similarity between images; a lower value indicates greater similarity. The SSIM (Structural Similarity Index) evaluates structural similarity; a higher value is better. The FID (Fréchet Inception Distance) measures the distance between the distribution of generated images and real images; a lower value indicates better generation quality and diversity. The CLIP Score [40] quantifies the accuracy of content preservation by calculating the cosine similarity between the input and output images in the CLIP image embedding space; a value closer to 1 indicates better content preservation.

As shown in Table 5, we compared our method with several mainstream baseline methods on 60 images that cover various scenes such as landscapes, buildings, wastelands, and forests. All image generation processes used a unified set of hyperparameters. The experimental results clearly demonstrate the superiority of our method.

Table 5. Style transfer method score sheet. ↓ indicates that lower scores are better, while ↑ indicates that higher scores are better.

Metric	Our Method	NET	pix2pix	CycleGAN	ControlNet
LPIPS ↓	0.138	0.484	0.605	0.617	0.516
SSIM ↑	0.865	0.499	0.047	0.409	0.171
FID Score ↓	177.08	502.82	541.137	401.513	247.0703
CLIP Similarity ↑	0.911	0.689	0.561	0.6153	0.705

In terms of content preservation, our method achieved a CLIP Score of 0.911, significantly higher than all competing methods, which proves the exceptional capability of our framework in preserving the core semantic information of the original scene. At the same time, an SSIM value of 0.865 indicates its excellent performance in maintaining the structural integrity of the scene. In terms of generation quality and perceptual similarity, our method also leads, with its LPIPS score (0.138) and FID score (177.08) being the lowest. This means that our generated results have the smallest perceptual difference from the original images, and that their image distribution is closer to the real world.

In summary, these quantitative data strongly prove that our method achieves the best performance currently available in balancing style injection and content preservation. It is capable of achieving high-quality artistic stylization while maximally preserving the original information of the input image.

4. Discussion

This research proposed and validated a DM-based framework for the cartoon style transfer of photorealistic 3D scenes. This section aims to provide an in-depth analysis of the research findings, rather than a simple repetition. Our core goal is to solve the difficult balance between "artistic expression" and "geographic semantic preservation" in traditional map style transfer. Both quantitative analysis and user studies jointly confirm that our framework successfully preserves the complex structure of the original 3D scenes while injecting a distinct and consistent cartoon style.

In terms of the precision of content preservation, our framework introduces the revolutionary ControlNet technology [36], using the original scene structure as a strong conditional input. This fundamentally solves the core problem of content and style entanglement in traditional GAN methods [9,18], which leads to the loss of geographic information.

Secondly, regarding the quality and stability of style transfer, this method discards the unstable training of GAN [20,24] and instead adopts a highly stable LDM as its framework. We innovatively combine the LDM with the MultiDiffusion strategy to coordinate global style and use DDIM inversion to generate structured initial noise, jointly ensuring the high quality and high consistency of the style injection and effectively addressing the research goal of "style consistency".

Compared to other DM methods that rely on text prompts [10], the innovation of this framework lies in its automated workflow that requires no additional prompts. This greatly enhances the practicality and reproducibility of the method when dealing with large-scale geographic data, solving a key bottleneck in the practical application of existing advanced models.

This study also, for the first time, extends style transfer from traditional 2D map images [9] to photorealistic 3D scenes, solving the key challenge of cross-view style consistency that previous methods have never faced.

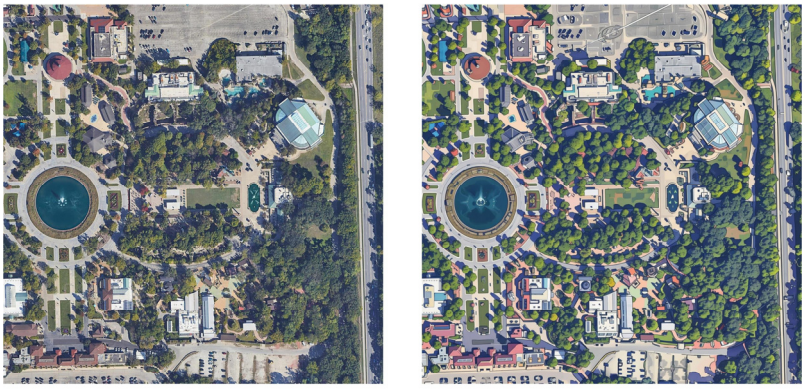
As shown in Figure 8, we used remote sensing images under different scenes as input, and the generated stylized images still maintain the semantic content of the original images while effectively reflecting the cartoon style. Although the cartoon style transfer method proposed in this paper can achieve high-quality results on both photorealistic 3D scenes and remote-sensing imagery, there are still the following shortcomings:

(1) The generation process requires a certain amount of time. The experiments in this paper were conducted on an NVIDIA RTX 4080 GPU, and for a single 1024 × 1024 image, the average sampling time is about 41 s, which limits the possibility of real-time generation. In terms of efficiency, it is necessary to start from aspects such as model structure optimization, knowledge distillation, and hardware acceleration to compress the generation time from the minute level to the second or even millisecond level, which is a prerequisite for achieving interactive applications and large-scale deployment.

(2) Although this method performs excellently when processing high-resolution remote-sensing imagery from Google Earth and preprocessed 1024 × 1024 oblique photogrammetry tiles, its application on standard maps is not ideal. As shown in Figure 9, the core advantage of our framework lies in the cartoon-style simplification of texture-rich real-world scenes. When the input is standard map data, which are already highly abstract and symbolic, the model has difficulty performing effective secondary simplification. The generated results tend to remain at a superficial level of color enhancement and edge tracing, rather than producing creative style transfer. This indicates that future work requires the model to be able to intelligently identify different data types and adopt differentiated stylization strategies.



Figure 8. *Cont.*



Input Output

Figure 8. Cartoon style transfer effect of remote-sensing images.

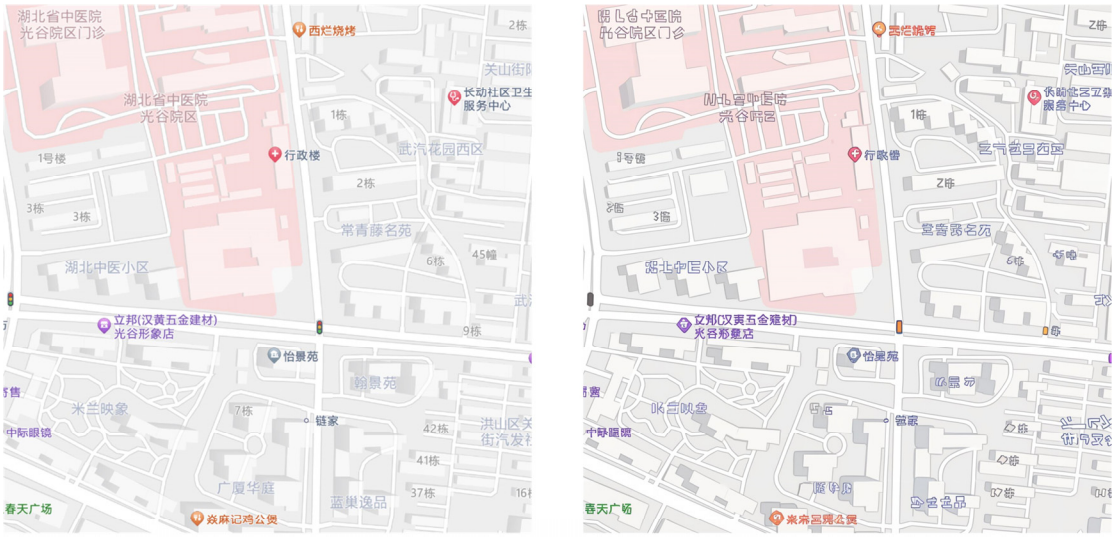


Figure 9. Effect diagram of standard map style transfer (All Chinese characters appearing in the figure indicate place labels).

5. Conclusions

In this paper, we introduced a method for geographic image cartoon style transfer based on DM. Our method is different from existing methods, as we use a pre-trained LDM and a pre-trained LoRA model as the generation backbone, which is combined with ControlNet strategies, MultiDiffusion strategies, and noise inversion to control the content and style transformation of the generated image. This method can achieve style transfer without additional optimization or text prompts. For remote-sensing or photorealistic 3D images of various scenes and contents, our method has shown superior performance in terms of visual quality, cartoon style consistency, and content information preservation. These results have greatly advanced the field of geographic image style transfer. The contributions of this paper include the following:

(1) Proposing an automated framework based on DM that requires no prompt guidance and can stably transfer complex cartoon styles to photorealistic 3D scenes. We have demonstrated the possibility of achieving high-quality, high-consistency artistic visualization without sacrificing the fidelity of geographic information, providing technical support for the evolution of cartography from a traditional functional expression to a narrative medium that integrates science and art;

(2) We conducted quantitative, qualitative, and user evaluations of this method. The results show that the method performs with high quality in both image content preservation and style transfer effects.

This technology is expected to unleash its potential in multiple fields. In urban planning, it can instantly transform abstract design schemes into visualized blueprints with a futuristic or ecological feel, greatly enhancing the communicative power and public appeal of planning drafts. In the field of cultural heritage protection, this technology can render precise 3D models into hand-drawn or print-style visuals that are full of historical sense, not only providing innovative narrative methods for digital displays but also offering powerful visual tools for virtual restoration and educational dissemination. In addition, in key decision-making scenarios such as emergency response and simulation exercises, highlighting core information and weakening secondary backgrounds through stylization can improve the efficiency of information acquisition and decision-making. We hope that future geographic information expression will no longer be limited to a single, rigid paradigm. Maps will evolve into a dynamic canvas that is full of creativity and capable of carrying emotions and narratives.

Author Contributions: Writing—original draft, Yuhang Chen; formal analysis, Yuhang Chen; methodology, Yuhang Chen; writing—review & editing, Yuhang Chen and Jing Chen; resources, Haoran Zhou and Yi Chao; project administration, Haoran Zhou; supervision, Haoran Zhou, Nai Yang and Jing Zhao; funding acquisition, Nai Yang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the National Natural Science Foundation of China [No. 42171438], with funding provided by Nai Yang.

Data Availability Statement: The remote sensing images, oblique photography images and real-world 3D scenes mentioned in this article are all sourced from Google Earth.

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their useful comments on the manuscript.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kent, A.J.; Vujakovic, P. Stylistic Diversity in European State 1: 50 000 Topographic Maps. *Cartogr. J.* **2009**, *46*, 179–213. [CrossRef]
2. Roth, R.E. Cartographic Design as Visual Storytelling: Synthesis and Review of Map-Based Narratives, Genres, and Tropes. *Cartogr. J.* **2021**, *58*, 83–114. [CrossRef]
3. Christophe, S.; Hoarau, C. Expressive Map Design Based on Pop Art: Revisit of Semiology of Graphics? *Cartogr. Perspect.* **2012**, *73*, 61–74. [CrossRef]
4. Christophe, S.; Mermet, S.; Laurent, M.; Touya, G. Neural map style transfer exploration with GANs. *Int. J. Cartogr.* **2022**, *8*, 18–36. [CrossRef]
5. Bogucka, E.P.; Meng, L. Projecting emotions from artworks to maps using neural style transfer. *Proc. ICA* **2019**, *2*, 9. [CrossRef]
6. Kang, Y.; Gao, S.; Roth, R.E. Transferring multiscale map styles using generative adversarial networks. *Int. J. Cartogr.* **2019**, *5*, 115–141. [CrossRef]
7. Chen, X.; Chen, S.; Xu, T.; Yin, B.; Peng, J.; Mei, X.; Li, H. SMAPGAN: Generative Adversarial Network-Based Semisupervised Styled Map Tile Generation Method. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4388–4406. [CrossRef]
8. Ganguli, S.; Garzon, P.; Glaser, N. GeoGAN: A Conditional GAN with Reconstruction and Style Loss to Generate Standard Layer of Maps from Satellite Images. *arXiv* **2019**, arXiv:1902.05611.
9. Jin, W.; Zhou, S.; Zheng, L. Map style transfer using pixel-to-pixel model. *J. Phys. Conf. Ser.* **2021**, *1903*, 012041. [CrossRef]
10. Li, Z.; Guan, R.; Yu, Q.; Chiang, Y.-Y.; Knoblock, C.A. Synthetic Map Generation to Provide Unlimited Training Data for Historical Map Text Detection. In Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Beijing, China, 2 November 2021; pp. 17–26. [CrossRef]

11. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33, Proceedings of the Annual Conference on Neural Information Processing Systems, Virtual, 6–12 December 2020*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 6840–6851. Available online: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html (accessed on 2 May 2025).
12. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695. Available online: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html (accessed on 2 May 2025).
13. Wang, B.; Wang, W.; Yang, H.; Sun, J. Efficient example-based painting and synthesis of 2D directional texture. *IEEE Trans. Vis. Comput. Graph.* **2004**, *10*, 266–277. [CrossRef]
14. Zhang, W.; Cao, C.; Chen, S.; Liu, J.; Tang, X. Style Transfer Via Image Component Analysis. *IEEE Trans. Multimed.* **2013**, *15*, 1594–1601. [CrossRef]
15. Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; Song, M. Neural Style Transfer: A Review. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 3365–3385. [CrossRef] [PubMed]
16. Zhang, C.; Zhu, Y.; Zhu, S.-C. MetaStyle: Three-Way Trade-off among Speed, Flexibility, and Quality in Neural Style Transfer. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33. [CrossRef]
17. Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; Xu, C. Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning. In Proceedings of the SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–8. [CrossRef]
18. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2232. Available online: https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html (accessed on 2 May 2025).
19. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv* **2019**, arXiv:1809.11096.
20. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral Normalization for Generative Adversarial Networks. *arXiv* **2018**, arXiv:1802.05957.
21. Huang, N.; Tang, F.; Dong, W.; Xu, C. Draw Your Art Dream: Diverse Digital Art Synthesis with Multimodal Guided Diffusion. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 1085–1094. [CrossRef]
22. Wang, Z.; Zhao, L.; Xing, W. StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 7677–7689. Available online: https://openaccess.thecvf.com/content/ICCV2023/html/Wang_StyleDiffusion_Controllable_Disentangled_Style_Transfer_via_Diffusion_Models_ICCV_2023_paper.html (accessed on 2 May 2025).
23. Li, S. DiffStyler: Diffusion-based Localized Image Style Transfer. *arXiv* **2024**, arXiv:2403.18461.
24. Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; Cohen-Or, D. NULL-Text Inversion for Editing Real Images Using Guided Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6038–6047. Available online: https://openaccess.thecvf.com/content/CVPR2023/html/Mokady_NULL-Text_Inversion_for_Editing_Real_Images_Using_Guided_Diffusion_Models_CVPR_2023_paper.html (accessed on 2 May 2025).
25. Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; Xu, C. Inversion-Based Style Transfer with Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 10146–10156. Available online: https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Inversion-Based_Style_Transfer_With_Diffusion_Models_CVPR_2023_paper.html (accessed on 2 May 2025).
26. Alaluf, Y.; Garibi, D.; Patashnik, O.; Averbuch-Elor, H.; Cohen-Or, D. Cross-Image Attention for Zero-Shot Appearance Transfer. *arXiv* **2023**, arXiv:2311.03335.
27. Hertz, A.; Voynov, A.; Fruchter, S.; Cohen-Or, D. Style Aligned Image Generation via Shared Attention. *arXiv* **2024**, arXiv:2312.02133.
28. He, F.; Li, G.; Zhang, M.; Yan, L.; Si, L.; Li, F. FreeStyle: Free Lunch for Text-Guided Style Transfer Using Diffusion Models. *arXiv* **2024**, arXiv:2401.15636.
29. Friedmannová, L. What Can We Learn from the Masters? Color Schemas on Paintings as the Source for Color Ranges Applicable in Cartography. In *Cartography and Art*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–13. [CrossRef]

30. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-To-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134. Available online: https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html (accessed on 12 May 2025).
31. Wu, M.; Sun, Y.; Jiang, S. Adaptive color transfer from images to terrain visualizations. *IEEE Trans. Vis. Comput. Graph.* **2023**, *30*, 5538–5552. [CrossRef]
32. Wu, M.; Sun, Y.; Li, Y. Adaptive transfer of color from images to maps and visualizations. *Cartogr. Geogr. Inf. Sci.* **2022**, *49*, 289–312. [CrossRef]
33. Zhang, L.; Rao, A.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 3836–3847. Available online: https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html (accessed on 2 May 2025).
34. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2256–2265. Available online: https://proceedings.mlr.press/v37/sohl-dickstein15.html (accessed on 2 May 2025).
35. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2022**, arXiv:1312.6114.
36. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
37. Ho, J.; Salimans, T. Classifier-Free Diffusion Guidance. *arXiv* **2022**, arXiv:2207.12598.
38. Bar-Tal, O.; Yariv, L.; Lipman, Y.; Dekel, T. MultiDiffusion: Fusing diffusion paths for controlled image generation. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; Volume 202, pp. 1737–1752.
39. Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv* **2022**, arXiv:2208.01626.
40. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763. Available online: https://proceedings.mlr.press/v139/radford21a.html (accessed on 2 May 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.