mathematics

MDPI

Article

Neural Rendering-Based 3D Scene Style Transfer Method via Semantic Understanding Using a Single Style Image

Jisun Park [1] and Kyungeun Cho [2,*]

[1] Department of Multimedia Engineering, Dongguk University-Seoul, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Republic of Korea; jisun@dongguk.edu
[2] Division of AI Software Convergence, Dongguk University-Seoul, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Republic of Korea
* Correspondence: cke@dongguk.edu; Tel.: +82-2-2260-3834

Abstract: In the rapidly emerging era of untact ("contact-free") technologies, the requirement for three-dimensional (3D) virtual environments utilized in virtual reality (VR)/augmented reality (AR) and the metaverse has seen significant growth, owing to their extensive application across various domains. Current research focuses on the automatic transfer of the style of rendering images within a 3D virtual environment using artificial intelligence, which aims to minimize human intervention. However, the prevalent studies on rendering-based 3D environment-style transfers have certain inherent limitations. First, the training of a style transfer network dedicated to 3D virtual environments demands considerable style image data. These data must align with viewpoints that closely resemble those of the virtual environment. Second, there was noticeable inconsistency within the 3D structures. Predominant studies often neglect 3D scene geometry information instead of relying solely on 2D input image features. Finally, style adaptation fails to accommodate the unique characteristics inherent in each object. To address these issues, we propose a novel approach: a neural rendering-based 3D scene-style conversion technique. This methodology employs semantic nearest-neighbor feature matching, thereby facilitating the transfer of style within a 3D scene while considering the distinctive characteristics of each object, even when employing a single style image. The neural radiance field enables the network to comprehend the geometric information of a 3D scene in relation to its viewpoint. Subsequently, it transfers style features by employing the unique features of a single style image via semantic nearest-neighbor feature matching. In an empirical context, our proposed semantic 3D scene style transfer method was applied to 3D scene style transfers for both interior and exterior environments. This application utilizes the replica, 3DFront, and Tanks and Temples datasets for testing. The results illustrate that the proposed methodology surpasses existing style transfer techniques in terms of maintaining 3D viewpoint consistency, style uniformity, and semantic coherence.

Keywords: 3D style transfer; neural rendering; neural radiance fields; semantic feature matching

MSC: 97R40; 68T07

Citation: Park, J.; Cho, K. Neural Rendering-Based 3D Scene Style Transfer Method via Semantic Understanding Using a Single Style Image. *Mathematics* **2023**, *11*, 3243. https://doi.org/10.3390/math11143243

Academic Editors: Hongang Qi, Yan Liu and Jun Miao

Received: 7 June 2023
Revised: 16 July 2023
Accepted: 21 July 2023
Published: 24 July 2023

1. Introduction

The COVID-19 pandemic increased the demand for three-dimensional (3D) virtual environments applicable to virtual reality (VR)/augmented reality (AR) and metaverses across diverse societal sectors, including education, training, and the arts. This has led to a marked rise in remote practical and application classes in addition to non-face-to-face service training. Other areas witnessing increased utilization include digital twin-based on-site monitoring and prediction, alongside the generation of free-viewpoint content. Despite these advances, the creation of various 3D virtual environments remains labor-intensive. Consequently, as the assortment of styles necessitating construction expands,

the associated time and costs escalate. For instance, consider an already-constructed 3D room featuring plastic chairs, white walls, and red floors. To modify the room's style to incorporate wooden chairs, walls, and floors, meticulous manual adjustments to texturing in accordance with the specific object and area are required. To circumvent this challenge, investigations have sought to automate the conversion of 3D virtual environment styles, as depicted in Figure 1. This is achieved by rendering a style-converted output image via artificial intelligence, which utilizes the final rendered two-dimensional (2D) image of the 3D virtual environment as the input to the style transfer network [1–7]. This methodology facilitates style conversion while retaining the structure of the original image by employing a technique to learn the feature vectors of both the original RGB image and style RGB image procured within the virtual environment.

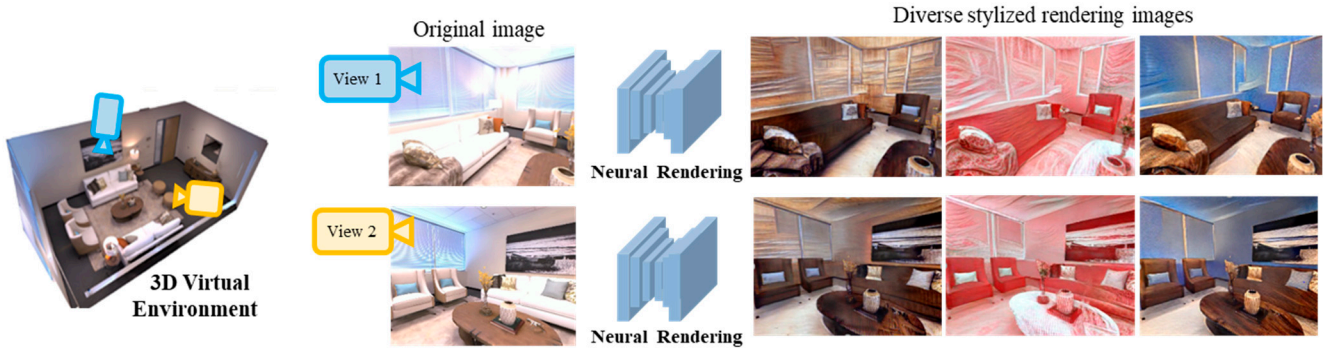

Figure 1. Conceptual schematic representation of 3D style transfer via rendering.

Nevertheless, the existing studies on style transfer have three primary limitations. First, training a style transfer network tailored to 3D virtual environments necessitates a considerable quantity of style image data that mirror the viewpoints of the virtual environment. Existing studies related to style transfer rendering generally feature restricted viewpoints, primarily because frequent viewpoint alterations within a 3D virtual environment render procuring numerous style images with analogous viewpoints challenging. The second issue pertains to a lack of consistency within the 3D structure. In single-frame image unit style transfer studies [8–11], only 2D input image features are considered, thereby excluding 3D scene geometry information. Consequently, when the viewpoint is altered, as shown in the first row of Figure 2, the 3D structure of red and yellow dashed box areas is not preserved. As indicated by the yellow and red regions, despite occupying the same position, the colored area of the wall undergoes minor variations in each frame owing to style transfer. Such inconsistencies are unsuitable for real-time style transformation rendering, such as VR/AR, as the scene perpetually flickers and the 3D structure of the scene modifies when the video is observed continuously.

Second, current studies do not embrace semantic 3D style transfer. Recent investigations have sought to mitigate the 3D inconsistency issues encountered during style transfer by considering the properties of consecutive frames during video-style transfer [12,13]. Moreover, studies have been conducted to perform style transfer without disturbing the 3D structure by employing artificial intelligence to assimilate 3D information, such as neural radiance fields [14,15]. Nevertheless, these studies, while successfully converting the overall scene style, failed to adapt the style to accommodate the unique characteristics of each object, as exemplified in the second row of Figure 2. In the style image, for instance, the walls are blue, and the furniture presents a wooden texture; however, the style-converted outcome does not replicate these attributes as green and blue dashed box areas. To execute style transfer for each object, style images from numerous viewpoints in an environment akin to 3D virtual space are required. However, this escalates the time and cost of constructing learning data, and the desire to apply more styles augments the burden of data compilation.

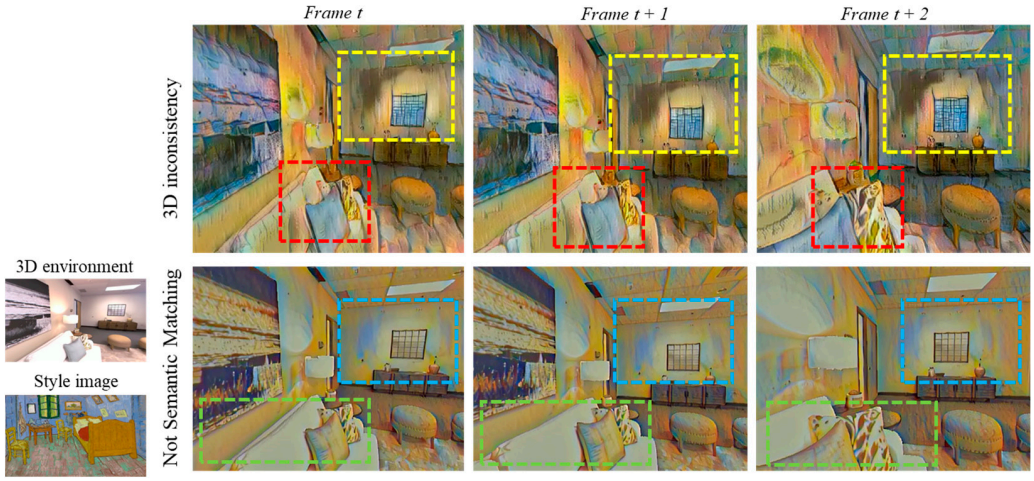

Figure 2. Limitations of current 2D video-style transfer techniques.

Therefore, this study proposes a methodology capable of transforming the style of a 3D scene by considering the style of the semantic domain while preserving the 3D structure using a single style image. The principal contributions of this study are summarized as follows:

(1) The proposed method facilitates a 3D scene style transfer that considers semantic style using only a single style image. Initially, this method learns 3D geometry using neural radiance fields, followed by learning the characteristics of a single style image. This strategy ensures structural alignment within a 3D scene while considering the semantic style.

(2) The proposed method incorporates a semantic nearest-neighbor feature-matching technique, thereby enabling semantic style transfer. By formulating a semantic color guidance image through a k-means-based semantic color guidance image generation module and simultaneously learning the characteristics of the corresponding image and style images, the technique can be adapted to match the style of each object, thus enabling semantic style transfer.

The remainder of this paper is organized as follows. Section 2 discusses recent advancements in the stylistic transformation of 2D images, video stylistic transformation, and 3D stylistic transformation. The proposed methodology is described in Section 3. The experimental results and subsequent evaluations are described in Section 4. Finally, Section 5 presents the conclusions of this study.

2. Related Work

2.1. Style Transfer of 2D Images

The style transfer of 2D images is a seminal field in computer vision studies, inaugurated by the arbitrary style transfer technique postulated by Gatys [16] and subsequently propagated by a myriad of 2D style transfer investigations, encompassing high-quality 2D style transfer [17–21], color matching [22–24], and texture synthesis [25–27]. These studies predominantly employed pretrained deep neural networks, such as the VGG pretrained model [28], to extract style image features and subsequently train a style transformation network to correspond to these features. Nevertheless, because these studies solely converted the style of a single image, a flickering effect manifests when applied to continuous images, such as videos, owing to the collapse of the 3D structure between each frame.

2.2. Style Transfer of Videos

To rectify the flickering phenomenon encountered when style transfer is implemented using a single image, video style transfer studies [29–33] aim to maintain temporal consistency, thereby enabling continuous style transfer. Investigations such as [12,29,30,32]

employed temporal coherency loss to learn temporal consistency between frames, whereas studies such as [13,33] combined the style and characteristics of input images to preserve temporal consistency. Although these techniques address temporal coherence in video images, they do not guarantee coherence in the 3D structure, and the style of the 3D space is not consistently upheld. For instance, limitations emerge in the viewpoint area when applied to a 3D environment as it becomes untenable to convert the style of images from unlearned harmful viewpoints.

2.3. 3D Style Transfer

In recent years, a variety of style transfer studies have been conducted on 3D scenes, following numerous investigations pertaining to 2D images. For instance, the studies of [34,35] converted the style of texture and geometry based on the mesh, whereas studies such as [36,37] learned a 3D scene to apply a style image premised on a point cloud. Nevertheless, in these investigations, the models can prove to be excessively complex, and there are challenges in applying them to 3D environments replete with various objects and backgrounds, akin to real environments. Conversely, in style transfer studies rooted in neural radiance fields [38–42], the 3D structure of a real environment can be effectively learned. However, in previous investigations, such as [43–47], which converted the style predicated on the learned 3D structure, the overall style of the entire 3D environment was converted rather than the characteristics of each individual object. This resulted in a problem, wherein the unique details of each object dissipated, and the distinction between the background and the object became indistinct. Although ARF [43] enhances style quality by transforming styles using local features rather than the entire image, a persistent issue remains. Even if an object exhibits a distinct color, it is converted into a disparate style. Also, StyleRF [44] transforms the grid features based on the reference style, resulting in high-quality zero-shot style transfer. However, it does not consider semantically matching the colors of each object. Therefore, in this study, we propose a semantic 3D style transfer network that can transform styles in accordance with the unique characteristics and details of each object using semantic style feature matching.

3. Proposed Method

We propose a methodology capable of transferring the style of a 3D virtual environment by considering the style of the semantic domain, while maintaining the 3D structure using only a single style image. The proposed method learns 3D information predicated on neural radiance fields [38] and compares and optimizes the characteristics of the rendered and style images, as illustrated in Figure 3.

The proposed methodology for 3D scene style transfer via semantic nearest-neighbor feature matching encompasses two stages. The initial stage entails training the geometric neural radiance fields, and the succeeding stage involves training the style neural radiance field. In the context of geometric neural radiance fields, 360-degree stable-viewpoint images and their corresponding poses were first derived from a 3D virtual environment. Subsequently, a neural radiance field [38] was employed to learn 3D geometric scene information intrinsic to the virtual environment. In the style neural radiance field training stage, style neural radiance fields are honed by optimizing the learned geometric neural radiance fields via semantic nearest-neighbor feature matching. This process incorporates not only learning the characteristics of the style image but also the attributes of the semantic color guidance image, which serves as a guide for style in the semantic domain. Optimized-style neural radiance fields are utilized for differentiable volume rendering, thereby generating 360-degree free-viewpoint images with a converted style while preserving the structure of the 3D virtual environment. In the following sections, a detailed description of each stage is provided.

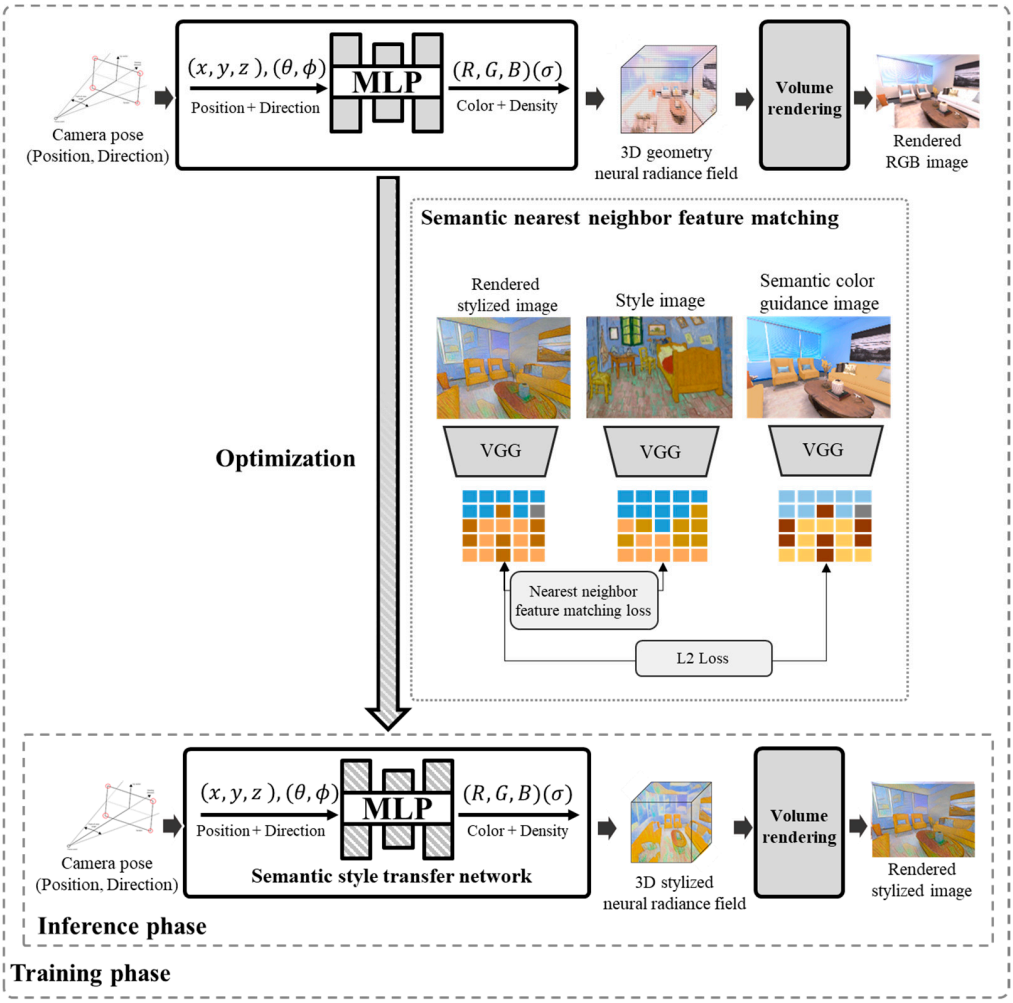

Figure 3. Schematic representation of proposed method for semantic 3D scene style transfer.

3.1. Geometry Neural Radiance Fields Training Stage

First, the proposed approach learns the geometric intricacies of a 3D virtual environment. We employ the architectural design of neural radiance fields, as delineated in [38], to train these fields within a 3D environment. Neural radiance fields, which are differentiable rendering models, can comprehend 3D information. We utilized a multilayer perception (MLP) network, formed of fully connected layers, represented as $\mathcal{F}_\Theta(x, d) \rightarrow (c, \sigma)$. Here, $x = (x, y, z)$ defines the 3D position of a sampled point, while $d = (\theta, \phi)$ denotes the viewpoint's direction. Upon inputting these two data points, the network is trained to embed the 3D geometric information, outputting the c (RGB color value) and σ (density value) at the given spatial point.

$$\mathcal{F}_{\text{geometry}}(x, d) \rightarrow (c, \sigma) \tag{1}$$

In this study, we obtained the geometric data of the 3D virtual environment by training a geometric neural radiance field network using RGB images and pose values sourced from the 3D environment, as shown in Equation (1). The trained $\mathcal{F}_{\text{geometry}}$ then served as the foundational model for the training of Equation (2), $\mathcal{F}_{style}$.

$$\mathcal{F}_{style}(x, d) \rightarrow (c', \sigma') \tag{2}$$

3.2. Style Neural Radiance Fields Training Stage

During this phase, $\mathcal{F}_{\text{geometry}}$, having undergone training, is optimized into $\mathcal{F}_{style}$ to learn the style. The learning structure of the network is shown in Figure 4. We begin by initializing the parameters of $\mathcal{F}_{\text{geometry}}$ network, which were trained for the 3D structure. When we input the position value and direction at a specific spatial point, the network outputs color and density. Subsequently, volume rendering was used to depict the image at that spatial point. Initially, the model generates a conventional RGB image; however, as the learning progresses, it begins to convert an image to the chosen style. To learn the style, we extracted features of the style and rendered images using a pretrained VGG model [28] and gradually reduced the loss by comparing these features. Furthermore, the model was trained to decrease the loss between the original image and the rendered image by creating a semantic color guidance image, aiding in the semantic conversion of the style. We employed a pretrained segmentation network [48] to derive a segmentation image from both RGB and style images to create a semantic color guidance image. Upon inputting the extracted segmentation and RGB images into the semantic color guidance image generation module, semantically relevant colors were generated. Section 3.2.1 provides a comprehensive explanation of the semantic color guidance image generation module.

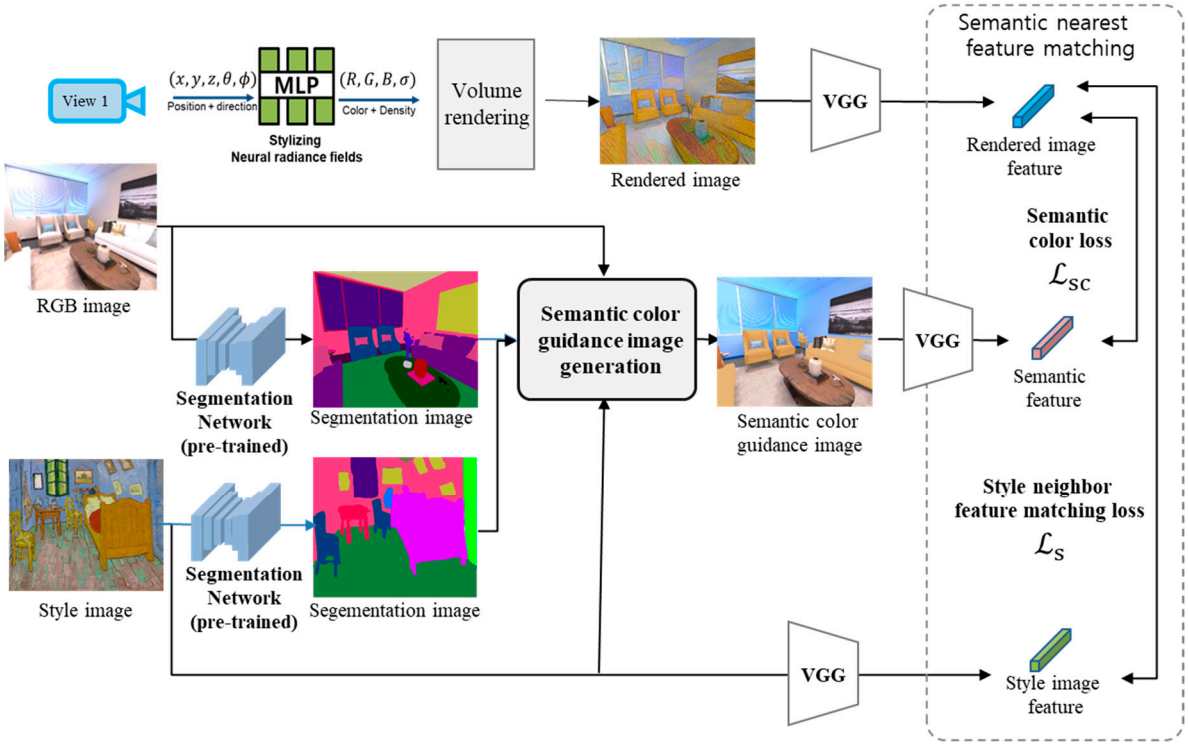

Figure 4. Training progression within the proposed semantic style transfer network.

3.2.1. Semantic Color Guidance Image Generation

This section outlines the procedure for employing the semantic color guidance image generation module to achieve semantic colors, as illustrated in Figure 5. Initially, decomposition into object-level images was performed using style and segmentation images. Subsequently, from the object-layer image partitioned into distinct object classes, K color codes were obtained through k-means clustering, with an emphasis on extracting the color exhibiting the highest occurrence ratio.

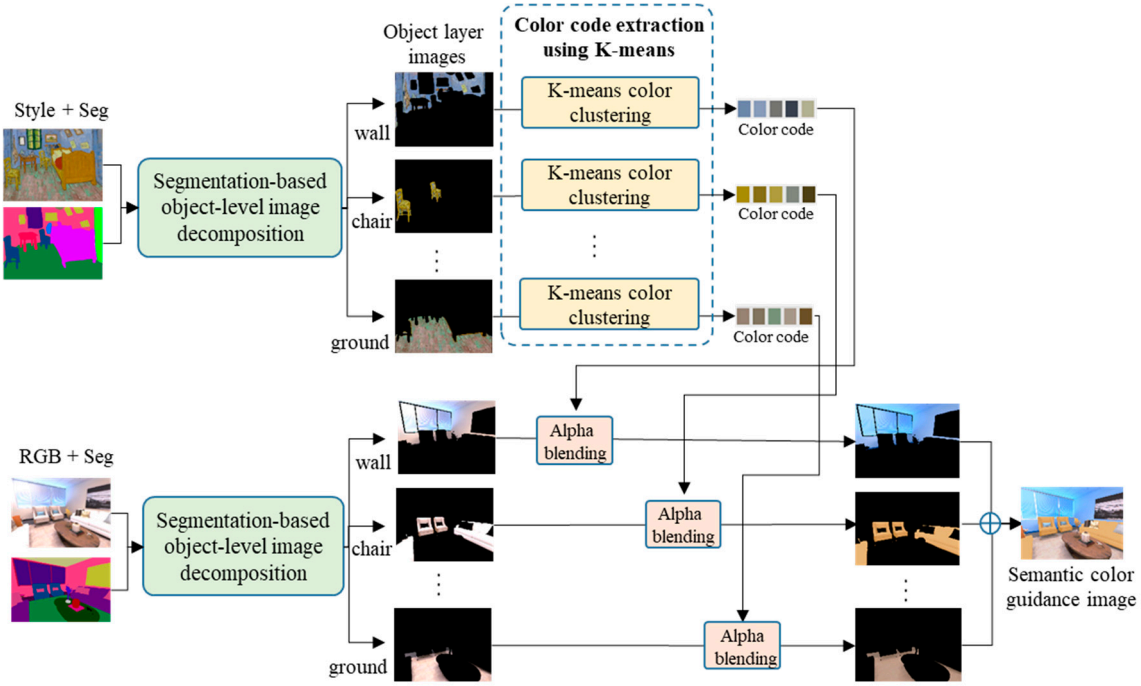

Figure 5. Semantic color guidance image generation process.

In this study, we extracted five representative color codes using five Ks to perform the analysis. Subsequently, alpha blending was applied to the object-layer image of the decomposed RGB image using the first color from the representative color code for each layer. Alpha blending is a technique [49] utilized to merge two images into a single composition. By employing Equation (3), the pixel color *Ca* with opacity αa is superimposed onto the pixel color *Cb* with opacity αb, resulting in the creation of a new color *Co* with its respective opacity.

$$C_o = \frac{C_a \alpha_a \; C_b \alpha_b (1 - \alpha_a)}{\alpha_o} \tag{3}$$

Once the blending of the style image color with each object layer was accomplished, a semantic color guidance image was generated by combining all layers. The resultant image possesses semantic interpretability and aids in inducing color learning for each object during the subsequent step of semantic nearest-neighbor feature matching.

3.2.2. Semantic Nearest-Neighbor Feature Matching

As depicted in Figure 6, to facilitate the acquisition of semantic style transfer, the loss in the rendered image was determined by employing the semantic color guidance and style images. During the loss calculation, the pretrained VGG model [28] was employed to extract the feature maps $\mathcal{F}_{semantic}$, $\mathcal{F}_{style}$, and $\mathcal{F}_{render}$ from the semantic color guidance images, style images, and rendered images, respectively. To achieve a localized style conversion, the discrepancy between $\mathcal{F}_{style}$ and $\mathcal{F}_{render}$ was calculated using the style neighbor feature-matching loss ($\mathcal{L}_s$). For semantic style transfer, the distinction between $\mathcal{F}_{semantic}$ and $\mathcal{F}_{render}$ was measured as the semantic color loss ($\mathcal{L}_{sc}$). A comprehensive explanation of each loss is provided below.

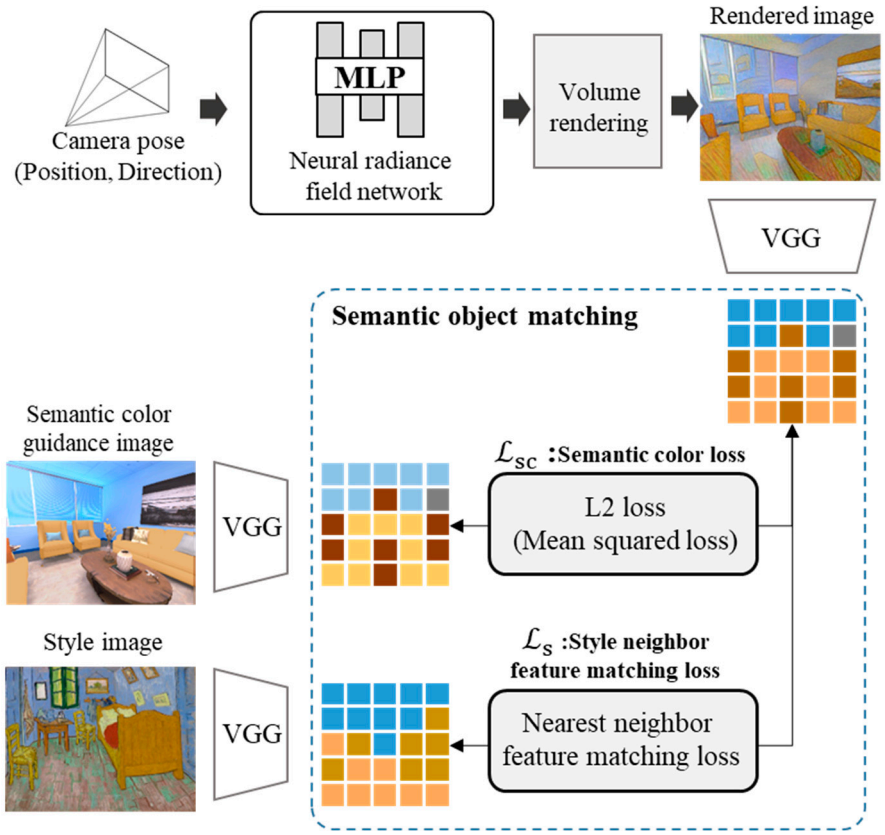

Figure 6. Semantic nearest-neighbor feature-matching process.

Style neighbor feature-matching loss ($\mathcal{L}_s$): To capture intricate high-frequency visual details in the style, the style feature map and the rendered feature map are utilized to compute the style nearest-neighbor feature-matching loss, as illustrated in Equation (4). By employing nearest-neighbor feature matching [43], the focus was on learning local-level features rather than the overall image style.

$$\mathcal{L}_{sc}\left(\mathcal{F}_{render}, \mathcal{F}_{style}\right) : \mathcal{L}_{nnfm}\left(\mathcal{F}_{render}, \mathcal{F}_{style}\right) = \frac{1}{N}\min_{i',\, j'} D\left(\mathcal{F}_{render}(i,j),\ \mathcal{F}_{style}(i',j')\right) \quad (4)$$

$\mathcal{F}_{render}(i,j)$ and $\mathcal{F}_{style}(i',j')$ denote the feature vectors at positions i, j and i', j', respectively. The cosine distance between these vectors is calculated and subsequently minimized using Equation (5).

$$\mathrm{D}(v_1, v_2) = 1 - v_1^T v_2 / \sqrt{v_1^T v_1 v_2^T v_2} \quad (5)$$

Semantic color loss ($\mathcal{L}_{sc}$): To learn semantic characteristics, the mean square loss is employed to minimize the disparity between the two feature vectors, as demonstrated in Equation (6).

$$\mathcal{L}_s(\mathcal{F}_{render},\ \mathcal{F}_{semantic}) : \mathcal{L}2(\mathcal{F}_{render},\ \mathcal{F}_{semantic}) = \sum_{i=1}^{n} (\mathcal{F}_{render},\ \mathcal{F}_{semantic})^2 \quad (6)$$

Total loss: The final total loss is given by Equation (7). In the equation, λ is weights that control the importance of the semantic terms.

$$\mathcal{L}_{total} = \mathcal{L}_{sc}\left(\mathcal{F}_{render}, \mathcal{F}_{style}\right) + \lambda \cdot \mathcal{L}_s(\mathcal{F}_{render},\ \mathcal{F}_{semantic}) \quad (7)$$

4. Experiments

This section presents an overview of the experiments conducted to validate the performance and to analyze the results of the proposed semantic 3D style transfer framework.

4.1. Experimental Setup

The experiments were performed using a computer equipped with an Intel(R) Xeon(R) Gold 5218R 2.10 GHz processor and 32 GB of RAM. In addition, an NVIDIA graphics card RTX 3090 was utilized to train the proposed network. The training process was conducted in a development environment running on Ubuntu 20.04. To evaluate the effectiveness of the proposed style transformation network within 3D space, we assessed three aspects: 3D consistency, semantic consistency, and style consistency. These evaluations aimed to measure the quality and coherence of style transfer results. To conduct these assessments, we employed the following datasets and evaluation metrics.

4.1.1. Datasets

The experiment used three diverse datasets: Replica [50], 3DFront [51], and Tanks and Temples [52]. These datasets were rich in complex structures, intricate details, and a wide array of objects. Replica [50], a high-quality 3D virtual environment dataset, comprises meshes, high-quality textures, and semantic data. In this study, harnessed RGB images and camera pose data were collected from the Room0, Room1, and Office2 environments for both training (600 sets) and testing (300 sets) for 3D geometry and semantic style transfer training. Subsequent tests assessed the semantic consistency via image segmentation. Additionally, to validate the effectiveness of our proposed method in both indoor and outdoor scenarios, we employed the 3DFront [51] and Tanks and Temples [52] datasets. The 3DFront dataset [51] incorporates 360-degree RGB images and their associated pose data, captured across a range of virtual indoor environments. In particular, we used data from Rooms 0044 and 1013 in our experiments. Conversely, the Tanks and Temples dataset [52] consists of 360-degree RGB images and camera pose information derived from actual environments, with an experiment deploying playground data.

To evaluate the ability of our proposed method to transfer a variety of styles, we conducted additional experiments utilizing an assortment of style images. These included Van Gogh's room, wooden room, outdoor winter scenes, and sketches.

4.1.2. Evaluation Methods

The performance evaluation in this experiment encompassed three key aspects: 3D, stylistic, and semantic consistencies. The evaluation methods employed for each aspect are as follows:

(1) 3D consistency evaluation:

To assess the consistency of different viewpoints within a 3D environment, we adopted the evaluation method proposed in [53]. This method calculates the temporal warping error between two frames in a video with viewpoint change. The evaluation formula, as shown in Equation (9), involves a warped frame denoted by $\hat{V}_{t+1}$ and a non-occlusion mask $M_t \in \{0, 1\}$, indicating the non-occluded regions. We evaluated consistency from two perspectives:

(a) Short-range consistency: This aspect focused on evaluating the consistency between nearby novel views, calculated at 1-frame intervals.
(b) Long-range consistency: The consistency between faraway novel views was assessed and calculated at 5-frame intervals.

$$E_{warp}(V_t, V_{t+1}) = \frac{1}{\sum_{i=1}^{N} M_t^{(i)}} \sum_{i=1}^{T-1} M_t^{(i)} \left\| V_t^{(i)} - \hat{V}_{t+1}^{(i)} \right\|_2^2 \quad (8)$$

(2) Style consistency evaluation:

Style consistency was assessed using a perceptual metric (LPIPS) [54]. LPIPS [54] is an effective tool for gauging the perceptual similarity between two images, as defined in Equation (7). To compute the LPIPS, we input the two images under comparison into a pretrained VGG network, which enabled the extraction of the feature values from the middle layer. Subsequently, we measured and evaluated the similarities between the two features. In this study, we compared the style similarity between style images and rendered images.

$$LPIPS = \sum_{l} \frac{1}{H_{l,} W_{l}} \sum_{h,w} \left\| w^{l} \bigodot \left(\hat{y}^{l}_{hw} - \hat{y}^{l}_{0hw} \right) \right\|^{2}_{2} \tag{9}$$

(3) Semantic consistency evaluation:

For semantic consistency evaluation, we utilized a pretrained segmentation model to extract segmentation results. These results were then compared with the segmentation of the original RGB images to assess semantic matching [55]. If the unique characteristics of each object were not preserved and appeared to be generally blurred, this was indicative of low semantic consistency. The segmentation evaluation formula, as shown in Equation (6), calculates pixel accuracy.

$$\text{Pixel Accuracy(PA)} = \frac{TP + TN}{FP + FN + TP + TN} \tag{10}$$

By employing these evaluation methods, we aimed to comprehensively assess the 3D, stylistic, and semantic consistencies within the proposed framework, thereby gaining insight into the performance and quality of the style transfer process.

4.2. Experimental Results

This section presents a comprehensive analysis of the qualitative and quantitative comparison results for the 3D view consistency, style consistency, and semantic consistency.

4.2.1. 3D-View Consistency Results and Comparison

Figure 7 illustrates the qualitative comparison results of 3D-view consistency. For this evaluation, we compared our proposed method with several existing approaches, including AdaIN [8], WCT [9], AreUST [56], ReReVST [12], MCCNet [13], and ARF [43]. Our analysis revealed notable differences among these methods. AdaIN [8], WCT [9], and AreUST [56], which rely solely on the use of 2D frame information, failed to preserve the shape and style texture of the objects when the viewpoint changed for each frame. ReReVST [12] and MCCNet [13], which utilize 2D video information, maintained the 3D structure and style texture for each frame; however, they did not accurately match the style of individual objects with the target Gogh-style image. Conversely, ARF [43] successfully converted the overall scene style while preserving the 3D structure according to the viewpoint. However, it exhibited errors and mismatches with individual objects. In contrast, our proposed method excelled in transferring the style of each object, matching it semantically with the target Gogh-style image, while preserving the 3D structure in an environment where the viewpoint changes.

The qualitative comparison provided a comprehensive understanding of how our proposed method outperformed the existing approaches in terms of achieving both semantic consistency and maintaining the 3D structure across changing viewpoints.

Figure 7. Qualitative comparisons of 3D-view consistency [8,9,12,13,43,56].

Table 1 presents quantitative comparisons of the short-range consistency calculated at 1-frame intervals using temporal warping error. Our proposed method exhibits a lower temporal warping error than previous methods, ensuring continuous maintenance of 3D-view consistency.

Table 1. Quantitative comparisons for short-range consistency (Bolds indicate the highest consistency; the lower the better).

Data	AdaIN [8]	WCT [9]	AreUST [56]	ReReVST [12]	MCCNet [13]	ARF [43]	Ours
Room 0	1.3610	2.6471	1.5472	1.4704	1.4590	**0.2054**	0.2697
Room 1	1.7866	2.6240	1.8340	1.5739	1.2554	0.3720	**0.2224**
Office 2	1.5438	2.6627	1.6636	1.4346	1.2954	0.2832	**0.2802**
Room 0044	1.7568	2.5653	1.6444	1.5665	1.3563	0.3254	**0.2658**
Room 1013	1.5669	2.3547	1.5335	1.4899	1.3464	0.3865	**0.2654**
Playground	1.8864	2.5182	1.9192	1.8338	1.1646	0.3011	**0.2379**
Average	1.6562	2.5104	1.6396	1.5448	1.3363	0.3341	**0.2613**

Table 2 provides quantitative comparisons of long-range consistency calculated at 5-frame intervals using the temporal warping error. Our proposed method demonstrates a lower temporal warping error than previous methods, even when dealing with long-range viewpoint movements.

Table 2. Quantitative comparisons for long-range consistency (Bolds indicate the highest consistency; the lower the better).

Data	AdaIN [8]	WCT [9]	AreUST [56]	ReReVST [12]	MCCNet [13]	ARF [43]	Ours
Room 0	5.3132	6.8596	4.5472	4.8139	4.9938	2.8125	**1.934**
Room 1	7.9342	6.0772	4.8340	5.6226	2.3598	3.9416	**1.7545**
Office 2	5.4719	6.4132	4.6636	4.3751	2.8393	**2.4181**	2.7884
Room 0044	5.8358	5.6587	4.3654	5.6654	3.6547	3.2587	**2.6987**
Room 1013	6.3548	5.7556	4.3549	6.3254	2.6587	3.2989	**2.3654**
Playground	8.8055	5.8038	4.9192	8.6928	2.2472	3.7675	**1.9527**
Average	6.3572	5.9009	4.6140	5.9556	3.1411	3.2641	**2.3905**

4.2.2. Style Consistency Results and Comparison

Figure 8 demonstrates the capability of our proposed methods to achieve style-consistent 3D scene view style transfer. Our method successfully generates style transfer results in a 3D environment that aligns with the desired style of each image, such as wood-, pink-, or blue-style rooms. For instance, during style transfer with a wood-themed image, the colors of the sofa and chairs changed to brown, matching the furniture color in the style image. Furthermore, the style transfer results maintained 3D-view consistency throughout the scene, preserving the spatial relationships between objects in the environment.

Table 3 presents quantitative comparisons among AdaIN [8], WCT [31], AreUST [56], ReReVST [12], MCCNet [13], ARF [43], and the proposed method. The evaluations were performed using Room 0, Room 1, and Office data from the replica dataset; Room 0044 and Room 1013 data from the 3DFront dataset; and Playground data from the Tanks and Temples dataset. On average, our proposed method achieved an impressive LPIPS score of 0.523, indicating the highest similarity between our style-transferred results and style images.

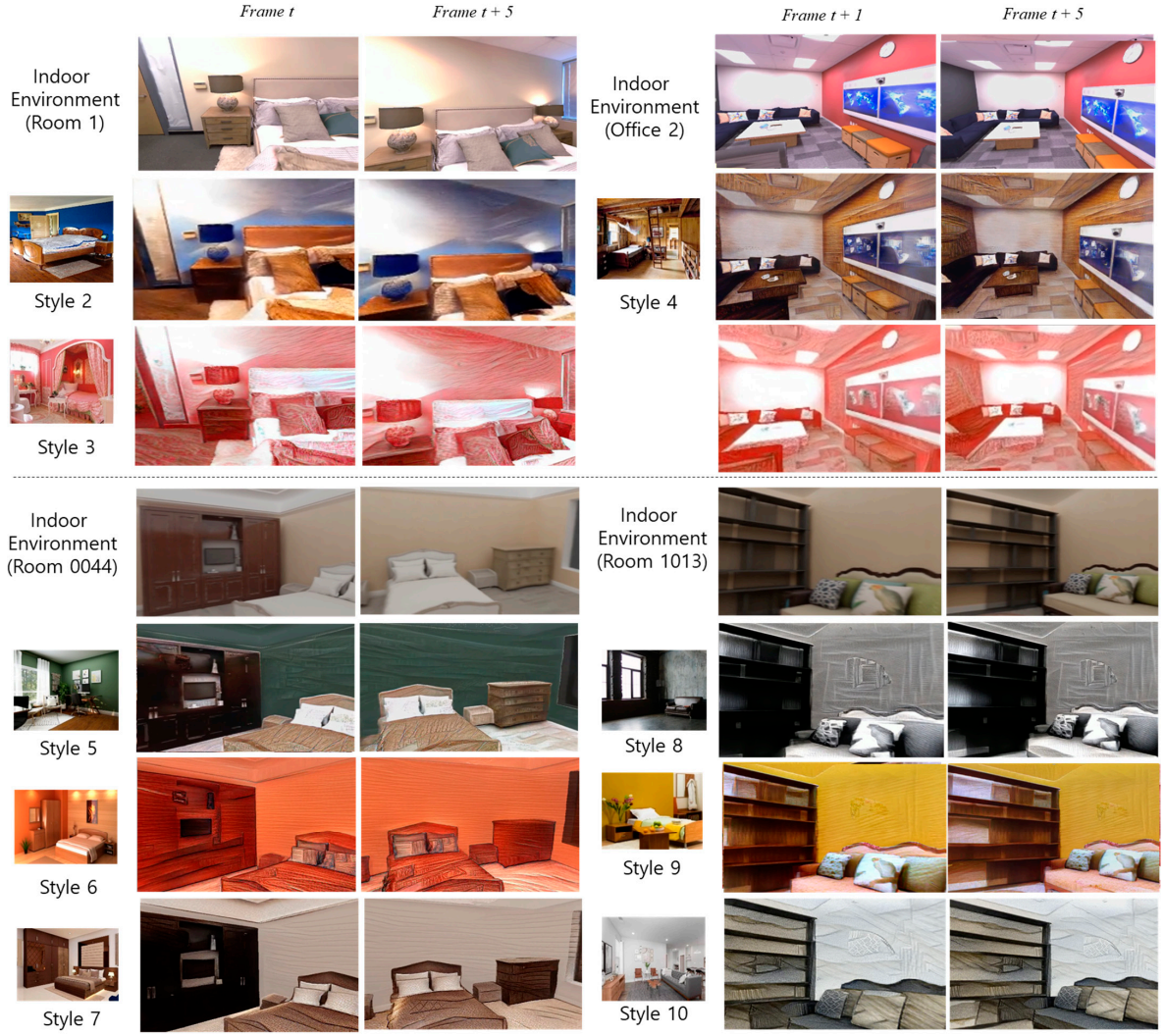

Figure 8. Qualitative comparison of style consistency in indoor environments using the Replica [50] and 3DFront [51] datasets.

Table 3. Quantitative comparisons on style consistency (Bolds indicate the highest consistency; the lower the better).

Style	AdaIN [8]	WCT [31]	AreUST [56]	ReReVST [12]	MCCNet [13]	ARF [43]	Ours
Style 1	0.7650	0.7183	0.7535	0.8686	0.8917	0.5854	**0.5232**
Style 2	0.6953	0.7897	0.7316	0.7929	0.8850	0.6098	**0.5003**
Style 3	0.6651	0.6912	0.7268	0.7775	0.8194	**0.5862**	0.5874
Style 4	0.8017	0.7187	0.7330	0.7484	0.7810	0.5663	**0.5262**
Style 5	0.5988	0.7889	0.7392	0.7747	0.7063	0.5636	**0.5164**
Style 6	0.7692	0.7752	0.7532	0.8173	0.7298	0.5829	**0.5062**
Style 7	0.6495	0.6923	0.7125	0.8469	0.7148	**0.5669**	0.5813
Style 8	0.6339	0.7739	0.7248	0.7648	0.7435	0.5723	**0.5116**
Style 9	0.6741	0.7918	0.7410	0.8371	0.8925	**0.5611**	0.5637
Style 10	0.6811	0.7307	0.8016	0.7065	0.8065	**0.5609**	0.5942
Style 11	0.6834	0.6923	0.6980	0.8040	0.7309	0.5810	**0.5785**
Style 12	0.7828	0.7825	0.7948	0.7593	0.7962	0.6019	**0.5057**
Style 13	0.7494	0.7281	0.7783	0.7331	0.8679	0.5919	**0.5568**
Average	0.7038	0.7441	0.7453	0.7870	0.7973	0.5792	**0.5424**

Figure 9 demonstrates the effectiveness of the proposed method in transferring styles, not only within virtual 3D environments but also in real outdoor 3D environments. The proposed method adeptly transforms the textures of diverse objects and areas, such as trees or floors, to align with each style image, while preserving the overall shape of the scene.

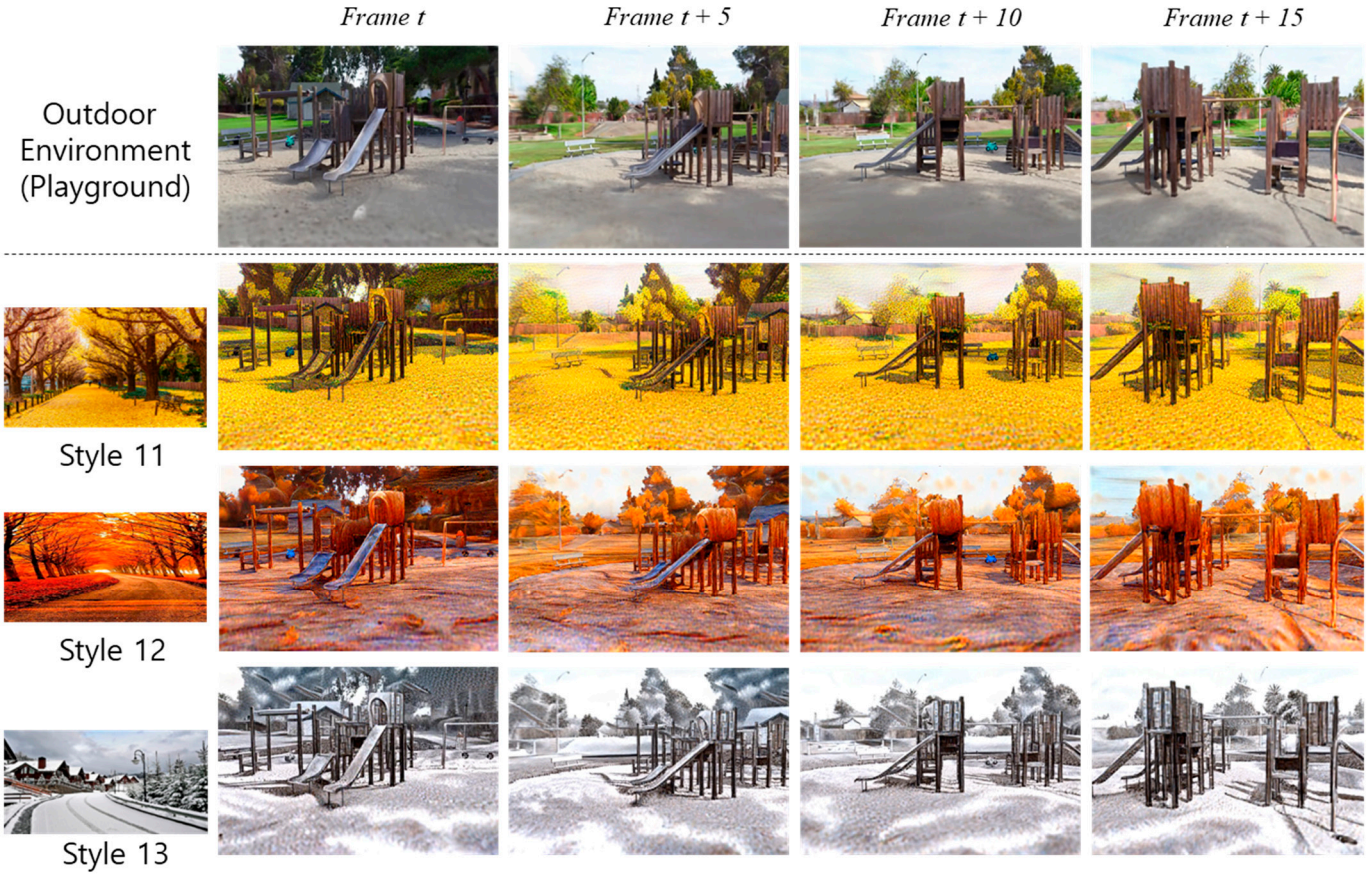

Figure 9. Style consistency results in outdoor environments using the Tanks and Temples [52] dataset.

4.2.3. Semantic Consistency Results and Comparison

Figure 10 shows the stylized results obtained by applying our proposed method to Room 0, Room 1, and Office 2 data using style images. In the style 1 image, the green box area, which represents the ground, displays a mixture of green and brown colors. The red box area, depicting a chair, exhibits a yellow-brown hue. The yellow box area, indicating a wall, appears blue. As depicted in Figure 10, our proposed method successfully achieves semantically driven style transfer. The ground is accurately colored with a combination of green and brown, while the wall is appropriately rendered in blue. Furthermore, the sofa adopts a yellow-brown shade in accordance with the style 1 image. In contrast, the results obtained from AdaIN [8], WCT [9], ReReVST [12], MCCNet [13], and ARF [43] do not accurately match the colors of the ground, wall, and objects with the style images.

Table 4 presents quantitative comparisons of semantic consistency using a pretrained segmentation model [48]. The proposed method achieved a remarkable semantic segmentation accuracy of 58.8, surpassing all the other tested methods in terms of semantic consistency.

Figure 10. Qualitative comparisons of semantic consistency [8,9,12,13,43].

Table 4. Quantitative comparisons for semantic consistency (Bolds represent the highest consistency).

Data	AdaIN [8]	WCT [9]	ReReVST [12]	MCCNet [13]	ARF [43]	Ours
Room 0	38.2	39.3	**58.7**	57.2	53.4	58.6
Room 1	40.3	41.2	58.3	58.6	52.9	**58.7**
Office 2	35.3	40.3	59.2	58.1	51.3	**59.3**
Average	37.9	40.3	58.7	57.9	52.5	**58.8**

5. Conclusions

This study focused on developing a semantic 3D scene style transfer method that relies solely on a single style image. The proposed method introduced a novel concept of a semantic nearest-neighbor feature-matching method using a neural radiance field. By utilizing a neural radiance field, the method learned 3D information and effectively optimized the rendered image characteristics through semantic nearest-neighbor feature matching with the trained neural radiance field and style image. Notably, the neural radiance field network captured not only the style image characteristics but also those of the semantic color guidance image, which provided style guidance in the semantic domain. The generation of the semantic color guidance image involved alpha blending with a semantic color image using semantic segmentation and k-means clustering.

Following optimization, our method successfully rendered 360-degree free-viewpoint style-transferred images of the 3D virtual environment while preserving its structural integrity. Our experimental results demonstrated the capability of our proposed method to achieve 3D viewpoint, style, and semantic consistency in both indoor and outdoor 3D environments, utilizing the replica dataset and Tanks and Temples datasets. Regarding 3D-viewpoint consistency, our method outperformed previous methods, attaining an average short-range temporal warping error of 0.2613 and long-range error of 2.3905. Furthermore, our method achieved an LPIPS score of 0.5424 for style consistency, indicating a high similarity between our style transfer results and style images. Notably, our proposed method achieved a remarkable semantic segmentation accuracy of 58.8, surpassing all

methods tested for semantic consistencies. These results underscored the effectiveness of our proposed method for semantic 3D scene style transfer, enabling the transfer of styles based on semantic features using only a single style image. We are planning to extend our method to enable semantic 3D style transfer for dynamic 3D scenes. With its broad applications in 3D scene style transfer, this method holds significant potential for various platforms, such as metaverse, VR, and AR.

Author Contributions: Conceptualization, J.P. and K.C.; methodology, software, validation, writing—original draft preparation, J.P.; writing—review and editing, J.P. and K.C.; supervision, project administration, funding acquisition, K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (2022R1A2C2006864).

Data Availability Statement: Replica [50], 3Dfront [51] and Tanks and Temples [52] datasets are used in this study. The datasets can be found here: https://github.com/facebookresearch/Replica-Dataset (accessed on 21 March 2023) and https://www.tanksandtemples.org/ (accessed on 21 March 2023), respectively.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ayush, T.; Thies, J.; Mildenhall, B.; Srinivasan, P.; Tretschk, E.; Yifan, W.; Lassner, C.; Sitzmann, V.; Martin-Brualla, R.; Lombardi, S.; et al. Advances in neural rendering. *Comput. Graph. Forum* **2022**, *41*, 703–735.
2. Xie, Y.; Takikawa, T.; Saito, S.; Litany, O.; Yan, S.; Khan, N.; Tombari, F.; Tompkin, J.; Sitzmann, V.; Sridhar, S. Neural Fields in Visual Computing and Beyond. *Comput. Graph. Forum* **2022**, *41*, 641–676. [CrossRef]
3. Huang, X.; Liu, M.-Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
4. Park, T.; Efros, A.A.; Zhang, R.; Zhu, J.-Y. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020.
5. Fabio, P.; Cerri, P.; de Charette, R. CoMoGAN: Continuous model-guided image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
6. Richter, S.R.; Al Haija, H.A.; Koltun, V. Enhancing photorealism enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1700–1715. [CrossRef] [PubMed]
7. Park, J.; Choi, T.H.; Cho, K. Horizon Targeted Loss-Based Diverse Realistic Marine Image Generation Method Using a Multimodal Style Transfer Network for Training Autonomous Vessels. *Appl. Sci.* **2022**, *12*, 1253. [CrossRef]
8. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
9. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M.-H. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems (NeurIPS)*; The MIT Press: Long Beach, CA, USA, 2017.
10. Li, X.; Liu, S.; Kautz, J.; Yang, M.-H. Learning linear transformations for fast arbitrary style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
11. Sturm, P.; Triggs, B. A factorization based algorithm for multi-image projective structure and motion. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 1996.
12. Wang, W.; Yang, S.; Xu, J.; Liu, J. Consistent video style transfer via relaxation and regularization. *IEEE Trans. Image Process.* **2020**, *29*, 9125–9139. [CrossRef] [PubMed]
13. Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Xu, C. Arbitrary video style transfer via multi-channel correlation. *AAAI* **2021**, *35*, 1210–1217. [CrossRef]
14. Nguyen-Phuoc, T.; Liu, F.; Xiao, L. Snerf: Stylized neural implicit representations for 3d scenes. *arXiv* **2022**, arXiv:2207.02363. [CrossRef]
15. Chiang, P.Z.; Tsai, M.S.; Tseng, H.Y.; Lai, W.S.; Chiu, W.C. Stylizing 3d scene via implicit representation and hypernetwork. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022.
16. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
17. Chen, T.Q.; Schmidt, M.W. Fast patch-based style transfer of arbitrary style. *arXiv* **2016**, arXiv:1612.04337.
18. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherland, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 694–711.

19. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
20. Kolkin, N.; Kucera, M.; Paris, S.; Sykora, D.; Shechtman, E.; Shakhnarovich, G. Neural neighbor style transfer. *arXiv* **2022**, arXiv:2203.13215.
21. Li, Y.; Liu, M.-Y.; Li, X.; Yang, M.-H.; Kautz, J. A closed-form solution to photorealistic image stylization. In *ECCV*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 453–468.
22. Xia, X.; Zhang, M.; Xue, T.; Sun, Z.; Fang, H.; Kulis, B.; Chen, J. Joint bilateral learning for real-time universal photorealistic style transfer. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 327–342.
23. Xia, X.; Xue, T.; Lai, W.S.; Sun, Z.; Chang, A.; Kulis, B.; Chen, J. Real-time localized photorealistic video style transfer. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 1089–1098.
24. Luan, F.; Paris, S.; Shechtman, E.; Bala, K. Deep photo style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4990–4998.
25. Risser, E.; Wilmot, P.; Barnes, C. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv* **2017**, arXiv:1701.08893.
26. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M.H. Diversified texture synthesis with feed-forward networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3920–3928.
27. Heitz, E.; Vanhoey, K.; Chambon, T.; Belcour, L. A sliced wasserstein loss for neural texture synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 9412–9420.
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
29. Chen, D.; Liao, J.; Yuan, L.; Yu, N.; Hua, G. Coherent online video style transfer. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1105–1114.
30. Ruder, M.; Dosovitskiy, A.; Brox, T. Artistic Style Transfer for Videos and Spherical Images. *Int. J. Comput. Vis.* **2018**, *126*, 1199–1219. [CrossRef]
31. Wu, Z.; Zhu, Z.; Du, J.; Bai, X. Ccpl: Contrastive coherence preserving loss for versatile style transfer. *arXiv* **2022**, arXiv:2207.04808.
32. Huang, H.; Wang, H.; Luo, W.; Ma, L.; Jiang, W.; Zhu, X.; Li, Z.; Liu, W. Realtime neural style transfer for videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 783–791.
33. Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; Ding, E. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 6649–6658.
34. Yin, K.; Gao, J.; Shugrina, M.; Khamis, S.; Fidler, S. 3dstylenet: Creating 3d shapes with geometric and texture style variations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 12456–12465.
35. Michel, O.; Bar-On, R.; Liu, R.; Benaim, S.; Hanocka, R. Text2mesh: Text-driven neural stylization for meshes. *arXiv* **2021**, arXiv:2112.03221.
36. Huang, H.P.; Tseng, H.Y.; Saini, S.; Singh, M.; Yang, M.H. Learning to stylize novel views. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 13869–13878.
37. Mu, F.; Wang, J.; Wu, Y.; Li, Y. 3d photo stylization: Learning to generate stylized novel views from a single image. *arXiv* **2021**, arXiv:2112.00169.
38. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 405–421.
39. Liu, L.; Gu, J.; Zaw Lin, K.; Chua, T.S.; Theobalt, C. Neural sparse voxel fields. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 15651–15663.
40. Yu, A.; Ye, V.; Tancik, M.; Kanazawa, A. pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4578–4587.
41. Zhang, K.; Riegler, G.; Snavely, N.; Koltun, V. Nerf++: Analyzing and improving neural radiance fields. *arXiv* **2020**, arXiv:2010.07492.
42. Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; Su, H. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 14124–14133.
43. Zhang, K.; Kolkin, N.; Bi, S.; Luan, F.; Xu, Z.; Shechtman, E.; Snavely, N. Arf: Artistic radiance fields. In *Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022*; Springer Nature: Cham, Switzerland, 2022; Proceedings, Part XXXI.
44. Liu, K.; Zhan, F.; Chen, Y.; Zhang, J.; Yu, Y.; El Saddik, A.; Xing, E.P. StyleRF: Zero-shot 3D Style Transfer of Neural Radiance Fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
45. Zhang, Y.; He, Z.; Xing, J.; Yao, X.; Jia, J. Ref-NPR: Reference-Based Non-Photorealistic Radiance Fields for Controllable Scene Stylization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.

46. Zhang, S.; Peng, S.; Chen, T.; Mou, L.; Lin, H.; Yu, K.; Zhou, X. Painting 3D Nature in 2D: View Synthesis of Natural Scenes from a Single Semantic Mask. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
47. Xu, S.; Li, L.; Shen, L.; Lian, Z. DeSRF: Deformable Stylized Radiance Field. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
48. Jain, J.; Li, J.; Chiu, M.; Hassani, A.; Orlov, N.; Shi, H. OneFormer: One Transformer to Rule Universal Image Segmentation. *arXiv* **2022**, arXiv:2211.06220.
49. Porter, T.; Duff, T. Compositing digital images. In Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques, Minneapolis, MN, USA, 23–27 July 1984; pp. 253–259.
50. Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Newcombe, R. The Replica dataset: A digital replica of indoor spaces. *arXiv* **2019**, arXiv:1906.05797.
51. Fu, H.; Cai, B.; Gao, L.; Zhang, L.X.; Wang, J.; Li, C.; Zhang, H. 3d-front: 3d furnished rooms with layouts and semantics. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021.
52. Knapitsch, A.; Park, J.; Zhou, Q.Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph. (ToG)* **2017**, *36*, 1–13. [CrossRef]
53. Lai, W.S.; Huang, J.B.; Wang, O.; Shechtman, E.; Yumer, E.; Yang, M.H. Learning blind video temporal consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
54. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
55. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
56. Wang, Z.; Zhang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Xing, W.; Lu, D. AesUST: Towards aesthetic-enhanced universal style transfer. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022.