EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2024)
M. Corsini, D. Ferdani, A. Kuijper, and H. Kutlu (Editors)

Short Paper

SemanticSplatStylization: Semantic scene stylization based on 3D Gaussian splatting and class-based style transfer

S. N. Sinha[1] and H. Graf[1] and M. Weinmann[2]

[1]Fraunhofer IGD, Germany
[2]Delft University of Technology, Netherlands

Abstract

We propose a novel approach for 3D Semantic Style Transfer in 3D Gaussian Splatting (3DGS) that applies style transfer to specific segments of a 3D scene using 2D style images. Our method leverages a finetuning of 3D Gaussian splats and fast 2D class-based style transfer to achieve targeted stylization with superior fidelity and multi-view consistency compared to existing state-of-the- art methods. By incorporating a semantic understanding, our approach ensures precise, context-aware stylization, aligning the visual characteristics of segments with their intended style. The application of 3D Semantic Style Transfer in cultural heritage preservation and restoration holds significant potential. By accurately capturing and transferring styles onto specific segments of cultural heritage objects, our approach demonstrates the potential of providing more accurate and visually appealing stylization results that preserve the integrity and historical significance of cultural heritage artifacts.

CCS Concepts

• ***Computing methodologies*** → ***Rasterization; Artificial intelligence; Image manipulation;***

1. Introduction

The conservation and preservation of cultural heritage rely on capturing digital heritage objects in terms of various media items such as photos, videos, text, or 3D reconstructions. Semantic systems [CD17] in cultural heritage utilize semantic technologies to organize and analyze cultural heritage items. Iconography[Tay20], a key discipline in art history, is essential for comprehending the connections between representations, their historical context, and social significance, making it integral to the development of a system facilitating research, clustering, and comparison of visual items in digital heritage. In the field of cultural heritage already proposals have been presented for using digital analogues for re-colorization and restoration of 3D objects [Øst19]. Neural style transfer [GEB15] is a powerful technique that combines content and style images to create high-quality artistic representations. Style transfer can play a vital role in digitally restoring and preserving cultural heritage by faithfully recreating unique artistic styles, textures, and colors. Moreover, incorporating historical image styles into 3D presentations enhances the immersive experience, surpassing traditional representations and promoting the exploration and appreciation of history through VR/AR technologies [ZGMJ23].

The advancements in radiance fields [MST*20] for 3D reconstruction have opened up new possibilities for immersive exploration of the 3D world. Radiance fields, which map 3D coordinates to color and density values, can be implemented using implicit MLPs, explicit voxels, or a combination of both. Recently, 3D Gaussian Splatting (3DGS) [KKLD23] has been demonstrated to surpass existing implicit neural representation methods in terms of both quality and efficiency, thereby representing the current state-of-the-art in novel view synthesis. Improved 3DGS methods like Gaussian grouping [YDYK23] support open-world and fine-grained scene understanding, enabling a variety of downstream scene editing applications with improved flexibility and effectiveness. The objective of 3D style transfer is to generate stylized representations of novel views in a 3D scene while maintaining consistency across multiple viewpoints. While 3D Gaussian Splatting based stylization methods, such as StyleGaussians proposed by Liu et al. [LZX*24], have shown success in achieving zero-shot style transfer and maintaining multi-view consistency, there are still limitations to consider. These methods often struggle to transfer style in a semantic manner, meaning that the stylization may not align with the underlying semantic content of the scene. Additionally, the geometry is not reconstructed based on the new styles, which can lead to a lack of preservation of style distortions.

We propose a 3D style transfer algorithm based on 3DGS that can transfer style to specific regions of the scene segmented by semantic Object-IDs. Although not representing a zero-shot approach, our method yields semantic stylization results that is not possible with the current state-of-the-art zero-shot methods. Moreover, our results are more faithful and consistent across multiple views, as supported by our quantitative and qualitative analyses. We incorporate an advanced shading function and differentiable environment light map for more accurate and realistic stylization.

DOI: 10.2312/gch.20241256

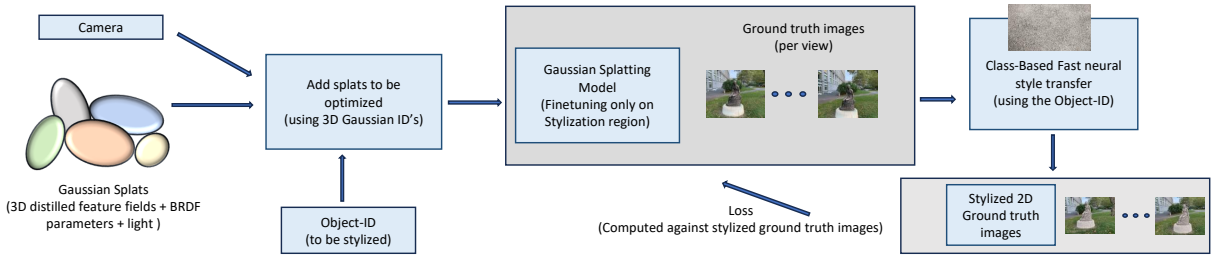

Figure 1: *Training pipeline for semantic stylization of Gasussian splats using our network*

2. Related work

2.1. Learning based methods and scene understanding

3D Gaussian Splatting (3DGS) [KKLD23] has emerged as a powerful technique for real-time 3D scene rendering, known for its fast reconstruction and high-quality results. It employs rasterization for rendering, enabling real-time performance and surpassing previous ray casting-based radiance field methods [MST*20]. Advanced methods like Gaussian grouping [YDYK23] and LangSplat [QLZ*23] go beyond appearance and geometry modeling, supporting open-world and fine-grained scene understanding. These methods outperform NeRF-based approaches [MST*20; ZLLD21] in terms of scene modeling capabilities. However, Gaussian-based methods struggle with scenes that have specular and reflective surfaces, as they do not explicitly model appearance properties like specular highlights. Recent methods, such as GaussianShader [JTL*23], have made progress in enhancing neural rendering for scenes with reflective surfaces while maintaining efficiency.

2.2. Neural Style transfer

Gatys et al. [GEB15] introduced neural style transfer, which separates and recombines content and style to create high-quality artistic images. Subsequent advancements, including feed-forward networks [JAF16; HB17; LLKY19], improved the speed of the optimization process. Style transfer has also been extended to the 3D domain [HTS*21; MWWL22], but many methods lack generalizability to new styles [NLX22; WJC*23]. Gaussian splatting-based models like StyleGaussian [LZX*24] achieve real-time transfer but do not support semantic style transfer or preserve stylization distortions.

3. Methodology

We present a method to semantically stylize a 3D representation in terms of Gaussian splats using the semantic object-IDs generated during the optimization process. As illustrate in Figure 1, we employ a fast 2D style transfer in combination with a semantic segmentation of a scene represented in terms of 3DGS to achieve a semantic style transfer for the Gaussian splats using fine-tuning in the selected region of interest.

3.1. Preliminaries

Gaussian splatting: In 3DGS [KKLD23], the scene is represented using a set of 3D Gaussians, with parameters including the center position, size, opacity, and color. These properties are differentiable and can be projected to 2D splats for scene optimization. Our implementation incorporates recent advancements in 3DGS algorithms to enhance scene understanding [YDYK23] and appearance modeling [JTL*23].

Semantic scene understanding: In our work, we utilize the Gaussian grouping method [YDYK23] to generate consistent mask identities across views of the scene and group 3D Gaussians with the same semantic information. The Segment anything model (SAM) [KMR*23] in combination with a zero shot tracker [COP*23] is used to automatically generate masks for each image, ensuring that each 2D mask corresponds to a unique identity in the 3D scene. Gaussian grouping introduces an additional attribute called Identity Encoding to each Gaussian, which efficiently distinguishes different objects in the scene. The final rendered 2D mask identity feature is computed as a weighted sum over the Identity Encoding of each Gaussian. A grouping loss to group the 3D Gaussians based on their object mask identities is utilized, which includes a 2D Identity Loss λ_{2d} and a 3D Regularization Loss λ_{3d}. The total training loss ($\mathcal{L}_{render}$) is represented as a combination of these losses, along with the conventional 3D Gaussian Loss, in the following equation:

$$\mathcal{L}_{render} = (1-\lambda)L_1 + \lambda \cdot \mathcal{L}_{D-SSIM} + \lambda_{2d}\mathcal{L}_{2d} + \lambda_{3d}\mathcal{L}_{3d} \quad (1)$$

where λ, λ_{2d} and λ_{3d} are weighting factors.

Appearance modeling: In our stylization framework, we incorporate an enhanced representation of appearance by utilizing spherical harmonic coefficients to capture reflections, similar to the approach employed in Gaussian shader [JTL*23]. We compute the appearances of the 3D Gaussian spheres using a shading function that incorporates various factors such as diffuse color, roughness, specular tint, normal vector, and a differentiable environment light map. This approach allows us to achieve a more realistic and elaborate stylization of the scene.

2D class-based style transfer: Our objective is to achieve real-time stylized image generation by employing the fast stylization method (FSM) [JAF16]. FSM trains a fully convolutional neural network (FCN) to produce a stylized image $T_S = M(I)$ that harmonizes the content of the input image I with the desired style S. Notably, this approach eliminates the need for optimization during inference and hence was the choice of algorithm for 2D semantic style transfer. The loss function given by

$$\mathcal{L}(I) = \underbrace{\frac{1}{C_2H_2W_2}||F_2(M(I)) - F_2(I)||_2^2}_{\textit{content loss}} + \underbrace{\sum_l^L \frac{1}{C_l}||G(M(I),l) - G(S,l)||_F^2}_{\textit{style loss}} \quad (2)$$

Figure 2: *Qualitative Analysis: (a) comparison with the state-of-the-art methods (b) Stylization of cultural heritage asset (Statue dataset[MAD*23]) using different stones.*

with

$$\underbrace{G_{i,j}(X,l)}_{\textit{gram matrix}} = \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} F_{l_{(h,w,i)}}(X) \cdot F_{l_{(h,w,j)}}(X) \tag{3}$$

where, $F_l(\cdot)$ is a feature extraction network (typically chosen as VGG16 [SZ14]) that outputs a feature map at the l-th level. The dimension of the level l feature map (in VGG16 it is 4 levels) is given by $C_l \times H_l \times W_l$, where C_l represents the number of channels, H_l represents the height, and W_l represents the width of the feature map at level l. The content loss preserves the spatial structure of the input image, while the style loss preserves the style defined by the style image. By combining these losses, the input image undergoes a style change while retaining its structure. To achieve a semantic style transfer in 2D [KVL19], we use the object-ID from Gaussian splats to get the object mask (O_m) and then apply it to the style image (T_S) to get the ground-truth stylized image (G_S) according to:

$$G_S = T_S * O_m + (1 - O_m) * I \tag{4}$$

where I is the input image.

3.2. Semantic stylization of splats

The steps involved in our proposed method for semantic 3DGS stylization are:

- Training the 3DGS with semantic scene understanding [YDYK23] and appearance modelling [JTL*23] as explained in Section 3.1
- Using the style images to train the fast neural transfer algorithm [JAF16]
- The 3D stylization process is achieved using the architecture depicted in Figure 1. The object ID of the segment to be stylized is used to identify the corresponding 3D Gaussian IDs associated with that object. These Gaussian IDs determine the set of splats belonging to the specified object. The region of interest is removed, and the subset Gaussian model, including the region to be stylized, is optimized. Finally, the new splats are generated using the semantic stylized 2D multi-view images as ground truth. During the fine-tuning process, only LPIPS [ZIE*18] loss is employed for the region to be stylized while using L1 loss outside this region.

4. Evaluation

In our evaluation, we present both a quantitative and qualitative analysis of our method, including cpmparisons to state-of-the-art methods. For the comparison, we utilize results directly taken from the respective papers.

Dataset For both the quantitative and the qualitative analysis, we utilized the statue dataset [MAD*23]. Additionally, for qualitative analysis, we used datasets from the Mip-Nerf 360 dataset (truck and train) [BMV*22] and style images from Wikiart ([Wik]).

Implementation details Our training process were conducted on an Nvidia RTX 3090 graphics card for 30,000 iterations for the Gaussian splatting model. During the stylization process, fine-tuning was performed for 10,000 iterations.

Quantitative analysis

Methods	Short-range Consistency		Long-range Consistency	
	LPIPS	*RMSE*	*LPIPS*	*RMSE*
HyperNet [CTT*22]	0.036	0.043	0.076	0.078
StyleRF [LZC*23]	0.050	0.045	0.123	0.098
StyleGaussian[LZX*24]	0.026	0.031	0.072	0.073
Ours	0.019	0.042	0.028	0.055

Table 1: ***Quantitative results.*** *We evaluate the performance of our semantic stylization method against the state-of-the-art in terms of consistency, using LPIPS (↓) and RMSE (↓).*

The absence of a standard quantitative metric for evaluating 3D style transfer quality has been acknowledged in previous research [LZC*23; ZKB*22]. To address this limitation, our evaluation focuses on several key aspects, including multi-view consistency and transfer quality as shown in Table 1. For assessing multi-view consistency, we adopt a similar methodology to previous studies [LZX*24; LZC*23; CTT*22], employing optical flow [TD20] and softmax splatting [NL20] to warp one view to another. We then utilize the masked RMSE score and LPIPS score [ZIE*18] as metrics to measure the consistency of stylization across views. In our experiments, we perform style transfers using six style images and the generated results (see Figure 2b) are compared only in the semantic stylized regions between the first and second frames for short-range consistency, and between the first and end frame for long-range consistency. As can be seen from Table 1, our method outperforms the other methods in terms of long-term consistency

and demonstrates some improvement in the LPIPS score for short-term consistency.

Qualitative analysis Our method demonstrates the ability to generate plausible results when compared to other style transfer methods (Figure 2a), and works also well on cultural heritage assets as shown with the Statue dataset [MAD*23], in the Figure 2b). By applying semantic style transfer to 3D Gaussian splats, our method accurately transfers style to individual segmented splats. In contrast, the other methods[CTT*22; LZC*23; LZX*24], although zero-shot, are unable to transfer styles to specific segments.

5. Conclusion and Future work

In this paper, we propose a method for semantically stylizing a scene represented by Gaussian splatting from a 2D style image. This technique has potential applications in digital restoration of cultural heritage objects by re-colorizing 3D objects like statues using iconographic evidence. By seamlessly integrating artistic styles into 3D environments, it enhances the transformation and representation of cultural heritage artifacts, providing immersive digital experiences. Future work can focus on developing an end-to-end architecture that integrates the training of style images directly into the training pipeline to improve efficiency and effectiveness.

Acknowledgement The work in this paper was partially funded by the European Commission for the PERCEIVE project (grant agreement 101061157).

References

[BMV*22] BARRON, J.T., MILDENHALL, B., VERBIN, D., et al. "Mipnerf 360: Unbounded anti-aliased neural radiance fields". *CVPR*. 2022 3.

[CD17] CARBONI, NICOLA and DE LUCA, LIVIO. "Towards a Semantic Documentation of Heritage Objects through Visual and Iconographical Representations". *International Information and Library Review* (2017) 1.

[COP*23] CHENG, HO KEI, OH, SEOUNG WUG, PRICE, BRIAN, et al. "Tracking Anything with Decoupled Video Segmentation". *ICCV*. 2023 2.

[CTT*22] CHIANG, P.Z., TSAI, M.S., TSENG, H.Y., et al. "Stylizing 3D Scene via Implicit Representation and Hypernetwork". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022 3, 4.

[GEB15] GATYS, LEON A., ECKER, ALEXANDER S., and BETHGE, MATTHIAS. *A Neural Algorithm of Artistic Style*. 2015 1, 2.

[HB17] HUANG, XUN and BELONGIE, SERGE. "Arbitrary style transfer in real-time with adaptive instance normalization". *Proceedings of the IEEE international conference on computer vision*. 2017 2.

[HTS*21] HUANG, HSIN-PING, TSENG, HUNG-YU, SAINI, SHASHANK, et al. "Learning to stylize novel views". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021 2.

[JAF16] JOHNSON, JUSTIN, ALAHI, ALEXANDRE, and FEI-FEI, LI. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*. 2016 2, 3.

[JTL*23] JIANG, YINGWENQI, TU, JIADONG, LIU, YUAN, et al. "GaussianShader: 3D Gaussian Splatting with Shading Functions for Reflective Surfaces". *arXiv* (2023) 2, 3.

[KKLD23] KERBL, BERNHARD, KOPANAS, GEORGIOS, LEIMKÜHLER, THOMAS, and DRETTAKIS, GEORGE. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". *ACM Transactions on Graphics* (2023) 1, 2.

[KMR*23] KIRILLOV, ALEXANDER, MINTUN, ERIC, RAVI, NIKHILA, et al. "Segment Anything". *arXiv* (2023) 2.

[KVL19] KURZMAN, LIRONNE, VAZQUEZ, DAVID, and LARADJI, ISSAM. "Class-Based Styling: Real-time Localized Style Transfer with Semantic Segmentation". *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019 3.

[LLKY19] LI, XIAOJUN, LIU, SIFEI, KAUTZ, JAN, and YANG, MING-HSUAN. "Learning linear transformations for fast image and video style transfer". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019 2.

[LZC*23] LIU, K., ZHAN, F., CHEN, Y., et al. "StyleRF: Zero-shot 3D Style Transfer of Neural Radiance Fields". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023 3, 4.

[LZX*24] LIU, KUNHAO, ZHAN, FANGNENG, XU, MUYU, et al. *StyleGaussian: Instant 3D Style Transfer with Gaussian Splatting*. 2024 1–4.

[MAD*23] MIRZAEI, ASHKAN, AUMENTADO-ARMSTRONG, TRISTAN, DERPANIS, KONSTANTINOS G., et al. "SPIn-NeRF: Multiview Segmentation and Perceptual Inpainting with Neural Radiance Fields". *CVPR*. 2023 3, 4.

[MST*20] MILDENHALL, BEN, SRINIVASAN, PRATUL P., TANCIK, MATTHEW, et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". *ECCV*. 2020 1, 2.

[MWWL22] MU, FANGWEI, WANG, JUE, WU, YIFAN, and LI, YU. "3D photo stylization: Learning to generate stylized novel views from a single image". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022 2.

[NL20] NIKLAUS, SIMON and LIU, FENG. "Softmax Splatting for Video Frame Interpolation". *IEEE Conference on Computer Vision and Pattern Recognition*. 2020 3.

[NLX22] NGUYEN-PHUOC, TAM, LIU, FENG, and XIAO, LI. "SNERF: Stylized Neural Implicit Representations for 3D Scenes". *arXiv* (2022) 2.

[Øst19] ØSTERGAARD, JAN STUBBE. ""Reconstruction" of the Polychromy of Ancient Sculpture: A Necessary Evil?": *Technè* (2019) 1.

[QLZ*23] QIN, MINGHAN, LI, WANHUA, ZHOU, JIAWEI, et al. "LangSplat: 3D Language Gaussian Splatting". *arXiv* (2023) 2.

[SZ14] SIMONYAN, KAREN and ZISSERMAN, ANDREW. "Very deep convolutional networks for large-scale image recognition". *CoRR* (2014) 3.

[Tay20] TAYLOR, PAUL. *Iconology and Iconography*. 2020 1.

[TD20] TEED, ZACHARY and DENG, JIA. "RAFT: Recurrent All Pairs Field Transforms for Optical Flow". *ECCV*. 2020 3.

[Wik] WIKIART. *Wikiart – visual art encyclopedia*. Website. https://www.wikiart.org/ 3.

[WJC*23] WANG, CHAO, JIANG, RUIQI, CHAI, MENGLEI, et al. "Nerf-Art: Text-Driven Neural Radiance Fields Stylization". *IEEE Transactions on Visualization and Computer Graphics* (2023) 2.

[YDYK23] YE, MINGQIAO, DANELLJAN, MARTIN, YU, FISHER, and KE, LEI. "Gaussian Grouping: Segment and Edit Anything in 3D Scenes". *arXiv* (2023) 1–3.

[ZGMJ23] ZÖLLNER, MICHAEL, GEMEINHARDT, JAN, MÖRTEL, MARINA, and JAHN, CELINA. *Style Transfer and Monocular Depth Estimation for Cultural Heritage Storytelling*. 2023 1.

[ZIE*18] ZHANG, RICHARD, ISOLA, PHILLIP, EFROS, ALEXEI A., et al. "The unreasonable effectiveness of deep features as a perceptual metric". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018 3.

[ZKB*22] ZHANG, KAI, KOLKIN, NATALIA, BI, SHIGUANG, et al. "ARF: Artistic Radiance Fields". *European Conference on Computer Vision*. 2022 3.

[ZLLD21] ZHI, SHUAIFENG, LAIDLOW, TRISTAN, LEUTENEGGER, STEFAN, and DAVISON, ANDREW J. "In-Place Scene Labelling and Understanding with Implicit Scene Representation". *ICCV*. 2021 2.