

Text2Scene: Text-driven Indoor Scene Stylization with Part-aware Details

Inwoo Hwang¹, Hyeonwoo Kim¹, and Young Min Kim^{1,2}

¹Department of Electrical and Computer Engineering, Seoul National University

²Interdisciplinary Program in Artificial Intelligence and INMC, Seoul National University

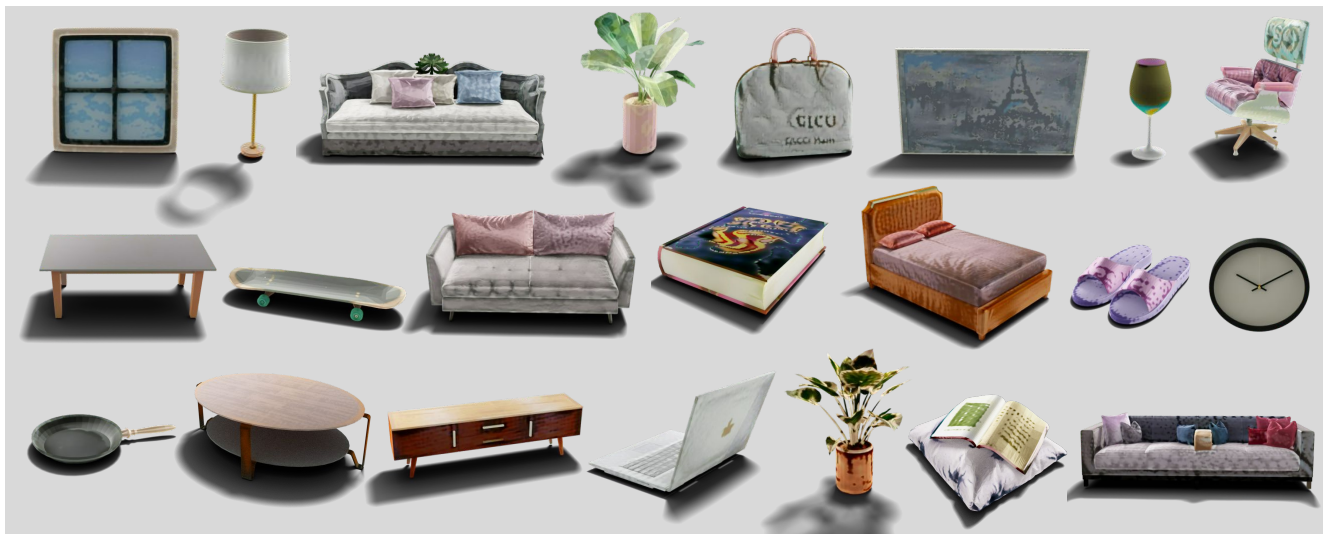


Figure 1. Stylization results of diverse objects. Text2Scene creates more realistic and part-aware textures for various categories of 3D objects without any dedicated datasets for training, which can constitute high-quality virtual scenes.

Abstract

We propose *Text2Scene*, a method to automatically create realistic textures for virtual scenes composed of multiple objects. Guided by a reference image and text descriptions, our pipeline adds detailed texture on labeled 3D geometries in the room such that the generated colors respect the hierarchical structure or semantic parts that are often composed of similar materials. Instead of applying flat stylization on the entire scene at a single step, we obtain weak semantic cues from geometric segmentation, which are further clarified by assigning initial colors to segmented parts. Then we add texture details for individual objects such that their projections on image space exhibit feature embedding aligned with the embedding of the input. The decomposition makes the entire pipeline tractable to a moderate amount of computation resources and memory. As our framework utilizes the existing resources of image and text embedding, it does not require dedicated datasets with high-quality textures designed by skillful artists. To the best of our knowledge, it is the first practical and scalable approach that can

create detailed and realistic textures of the desired style that maintain structural context for scenes with multiple objects.

1. Introduction

Virtual spaces provide an immersive experience for metaverse, films, or games. With increasing demands for virtual environments, various applications seek practical methods to create realistic 3D scenes with high-quality textures. Currently, skillful artists need to manually create 3D assets and accompanying textures with careful parameterization, which is not scalable enough to account for the diverse content the industry is heading for. Scenes can also be populated with existing 3D database models or created with recent shape-generation approaches using data-driven methods [33, 54]. However, most of them lack texture information or are limited to simple coloring.

To build realistic content, we need fine details containing the artistic nuances of styles that obey the implicit correlations with geometric shapes and semantic structure. Recent works provide methods to color a single object with the help

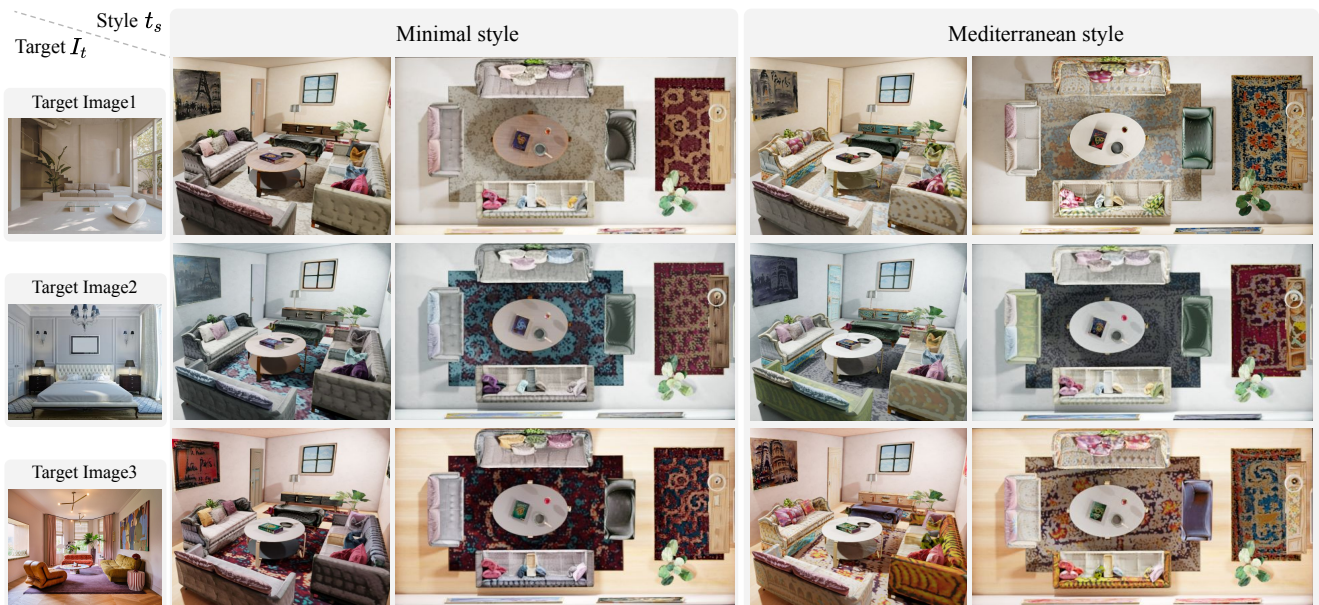


Figure 2. Our scene stylization results. Given a target image I_t and the style text t_s , Text2Scene can produce the stylized results for the entire scene.

of differentiable rendering [5], but often they are limited to single texture or blurred boundaries [6, 21, 27, 29, 52]. More importantly, the 3D objects are often textured in isolation, and only limited attempts exist to add visual appearances for large-scale scenes with multiple objects [14, 15, 19]. The biggest challenge is adding consistent style for an entire scene, but still accounting for the boundaries of different materials due to the functional and semantic relationship between parts, as observed within real-world scenes.

Our proposed Text2Scene adds plausible texture details on 3D scenes without explicit part labels or large-scale data with complex texturing. We take inspiration from abundant 3D shape and image datasets and decompose the problem into sub-parts such that the entire scene can be processed with a commodity memory and computation. Given scenes of multiple objects of 3D mesh geometry, we separately handle walls and individual objects. Specifically, the stylization of walls is formulated as texture retrieval, and the objects are initialized with base colors. From the base color assignment, we can deduce the part-level relationship for stylization and further refine them in later stages, such that their rendered images are close to the input text within the joint embedding space of foundational models.

Our coarse-to-fine strategy keeps the problem tractable yet generates high-quality texture with clean part boundaries. We first create segments of input mesh such that the segment boundaries align with low-level geometric cues. Then we start with the simplified problem of assigning a color per segment. Interestingly, the prior obtained from large-scale image datasets assign similar colors for the parts with similar textures, reflecting the semantic context or

symmetry as shown in Figure 1. We add the detailed texture on individual objects as an additional perturbation on the assigned base colors by enforcing constraints on the image features of their projections. The additional perturbations are high-frequency neural fields added to the base color.

In summary, Text2Scene is a new method that

- can easily generate realistic texture colors of the scene with the desired style provided by text or an image;
- can add detailed texture that respects the semantic part boundaries of individual objects; and
- can process the entire scene without a large amount of textured 3D scenes or an extensive memory footprint.

We expect the proposed approach to enable everyday users to quickly populate virtual scenes of their choices, and enjoy the possibility of next-generation technology with high-quality visual renderings.

2. Related works

3D Shape Understanding and Segmentation Our proposed method creates a realistic texture that abides by implicit rules of the material composition of different object parts. Different materials or textures are often assigned to different objects' functional parts, whose boundaries align with geometric feature lines. A handful of previous works find segments that constitute a 3D model with geometric information [7, 12, 20, 34], while others train 3D features to distinguish part labels provided in datasets [9, 30, 38]. Recently, PartGlott [24] suggested discovering part segments



Figure 3. Overall pipeline of Text2Scene. Given a 3D scene \mathcal{S} with optional text description, we generate textures guided by a target image I_t and the appearance style provided as a text t_s . We first stylize the structure by texture retrieval, and each object is pre-processed for part discovery for stylization. Then, we assign base colors to object parts and add local details for each object with the designated LNSF.

from language references, which should contain functional information. However, the geometric or functional distinction does not always clarify the texture boundaries, with possible additional diversity originating from the designers’ choice. The intricate rules are multi-modal distributions composed of a mixture of discrete part assignments and continuous texture details, whose results are contained within visual datasets of scenes.

Neural Fields and Texture Traditional texture atlas of 3D geometry is represented as a mapping from a planar image to a manifold embedded in 3D space and involves complex parameterization. With the increasing popularity of neural representation, TextureFields [31] represented texture as a mapping from 3D surface points to RGB color values without separate parameterization. It is deeply connected with graphics pipelines that adapt coordinate-based functions to depict SDF shapes [26, 32, 45] or novel-view synthesis using implicit volumes [28]. Various works show neural implicit representation is highly flexible and free from domain structure or resolution [2, 22, 25, 37, 49]. Recently, Text2Mesh [27] generated the deformation and coloring of input mesh with neural fields guided by the joint embedding of rendered image and input text. Interestingly, the generated mesh deformation and coloring also contain semantic part information. Our Text2Scene framework further extends the ability and observes distinct part boundaries, which could not be captured in previous works. We also increase the reality of resulting scenes with high-frequency details. We encode the input to the neural network with a high-order basis of intrinsic features [23] and apply the coarse-to-fine strategy as shown with geometric details of Yifan *et al.* [51].

Text-Driven 3D Stylization Recently neural networks trained with large-scale images and texts demonstrated powerful performance in many tasks with their extensive representation power. Here we primarily focus on 3D stylization attempts using image features. CLIP [39] learns latent space with a large amount of image text pairs, and the additional text input allows semantic manipulation for various generative tasks, including images [8, 35, 40–42], videos [3, 17], motions [46, 47], and 3D assets [18, 21, 36, 43]. However, for text-driven 3D stylization, it is still difficult to define clean texture boundaries and correlation with instance-level subtleties [6, 27, 29] or focus only on the specific type, such as human [16, 53]. As another way to stylize a 3D scene [4, 50], Yeh *et al.* [50] matches the input image features with CAD input using differentiable material graphs [44]. However, the representation is inherently limited to a combination of material libraries and cannot handle non-homogeneous details such as painting. On the other hand, our approach can generate texture beyond the repetitive low-level patterns of input materials and introduce part-aware texture with fine details that maintain geometric and semantic consistency.

3. Method

We stylize a 3D indoor scene without sophisticated techniques or software tools for 3D modeling or texturing. The input 3D scene $\mathcal{S} = \{\mathcal{W}, \mathcal{O}\}$ is a set of structure components \mathcal{W} and a set of objects \mathcal{O} . The structural components \mathcal{W} are walls, ceilings, and floors, whereas the objects $\mathcal{O} = \{M_i\}$ are the 3D mesh models M_i . We assume that all the components have their corresponding class labels and optionally have text descriptions.

The desired color distribution is provided as a target image I_t , which could be retrieved from the web. Also, a specific appearance style description t_s is provided as input to enforce style consistency among a set of objects \mathcal{O} . The color distribution compares the color histogram of the input target image I_t against the rendering of the current stylized scene I . Specifically, the histogram loss $\mathcal{L}_{\text{hist}}$ is defined as below

$$\mathcal{L}_{\text{hist}}(I, I_t) = \|H^{1/2}(I) - H^{1/2}(I_t)\|_2, \quad (1)$$

where $H(\cdot)$ indicates differentiable color histogram operator [1]. Also, we augment the losses derived from the joint embedding of text and images [39] by generating text descriptions T of the context using semantic labels and comparing them against the rendered image I . If we denote the pre-trained encoder for image and text as E_1 and E_2 , respectively, the CLIP similarity loss is defined as

$$\mathcal{L}_{\text{clip}}(I, T) = 1 - \text{sim}(E_1(I), E_2(T)), \quad (2)$$

where $\text{sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$ is the cosine similarity.

The overall pipeline is described in Fig. 3. We obtain the texture for the structure \mathcal{W} by texture retrieval, which is described in Sec. 3.1. The objects are stylized with additional decomposition to respect local part boundaries and, simultaneously, to practically handle multiple entities with details (Sec. 3.2).

3.1. Structure Stylization

We assign one coherent texture per structural element of \mathcal{W} . Compared to objects, the structural elements, such as walls or ceilings, are of simple planar geometry, and their textures are not heavily dependent on the relationships between different functional parts. For structural elements, it suffices to pick texture from the texture set of an existing material library of MATch [44], which contains homogeneous material. If we instead utilize visual features or CLIP embeddings for them, the resulting stylization exhibits undesired artifacts of various sizes instead of constant patterns, as shown in the supplementary.

We randomly initialize the texture from the texture set and render the structure image I_s , a bare room only containing the structural elements \mathcal{W} without objects. Then the materials are compared to the target image I_t for the histogram losses of Equation (1). The additional text prompt T_s is given as ‘a structure of a room’ to provide the context with $\mathcal{L}_{\text{clip}}(I_s, T_s)$. In summary, the texture of the structural element is retrieved to have the lowest score on the following criteria:

$$\mathcal{L}_{\text{hist}}(I_s, I_t) + \lambda_1 \cdot \mathcal{L}_{\text{clip}}(I_s, T_s). \quad (3)$$

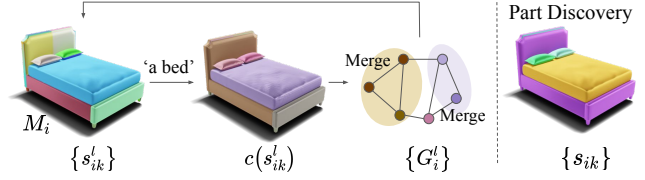


Figure 4. Overall pipeline of part discovery. We discover parts for object stylization from the super-segments of 3D object mesh. Given a 3D object mesh M_i with segments $\{s_{ik}^l\}$ at l^{th} iteration, we assign a color $c(s_{ik}^l)$ per segment and generate a graph \mathcal{G}_i^l . The pair of neighboring nodes is merged if the distance between assigned colors $c(s_{ik}^l)$ is within a threshold, then, move to $(l+1)^{\text{th}}$ iteration.

3.2. Object Stylization

Object stylization involves understanding the semantic structure hidden behind the mesh representation. As a pre-processing, we first subdivide individual objects M_i in \mathcal{O} into part segments $\{s_{ik}^l\}$ as described in Sec. 3.2.1. Then the scene is stylized in two steps. First, we assign base colors into individual parts to minimize the style loss for the entire scene $\mathcal{S} = \{\mathcal{W}, \mathcal{O}\}$ (Sec. 3.2.2). Here we are optimizing for the discrete set of colors assigned to subdivided parts obtained from the pre-processing. Then the textures for individual objects are further optimized to generate fine details (Sec. 3.2.3).

3.2.1 Part Discovery for Object Stylization

We first decompose individual objects into parts such that each part is composed of the same material or texture. The distinctive part boundaries are critical in providing semantic consistency and, therefore, perceptual realism toward the scene. Similar to 3D part segmentation methods, we first find super-segments based on geometric features which provide the granularity to define textural parts. For a given 3D object mesh M_i , the initial segments $\{s_{ik}^0\}$ are the decomposition applying the method by Katz *et al.* [20]. The decomposition is designed to be an over-segmentation of our aim. We generate a graph \mathcal{G}_i^0 where each node is the segment and edges connect neighboring segments. Then we incrementally merge segments that belong to the same texture until convergence. Note that, our part discovery method operates robustly regardless of the initial composition, but we use [20] which preserves the original geometry and details.

The challenge here is that, unlike semantic segmentation approaches, no large-scale public dataset exists that provides the ground truth for ‘texture similarity’ as the segmentation labels. We find a supervision signal from the large-scale pre-trained model, and create a simple text prompt $T_{i,c}$ using the class name, such as *a bed* or *a chair*. At l^{th} iteration, we assign a color $c(s_{ik}^l)$ to each segment $\{s_{ik}^l\}$, which is optimized to minimize the distance

between the rendered images I_{M_i} of multiple viewpoints and the text $T_{i,c}$ in the joint embedding space of CLIP, or $\mathcal{L}_{\text{clip}}(I_{M_i}, T_{i,c})$ as defined in Equation (2). If the resulting colors assigned to two adjacent segments are similar, the two parts are likely to be the parts with the same texture source. Therefore we merge the two segments for the next iteration $\{s_{ik}^{l+1}\}$. In particular, we merge segments if the assigned color has a distance of less than a threshold λ_{th} in the CIE color space which is known to be related to human perception. Note that, while the initial color is gray for the assignment, merging segments happen with the optimized color. We repeat the process until the number of segments does not decrease anymore and empirically found that it usually converges within 2-3 steps. The overall pipeline for part discovery is also described in Fig. 4.

3.2.2 Part-level Base Color Assignment

After the objects $M_i \in \mathcal{O}$ are decomposed into parts $\{s_{ik}\}$, we assign a solid color per part, namely $c(s_{ik})$. The base color assignment handles a low-dimensional optimization space with a coarse set of discrete parts but still observes the holistic distribution of the entire scene. Then the base color is combined with the output of designed neural style fields in Sec. 3.2.3, which generate high-frequency local texture.

We optimize the base color using the combined loss as before, $\mathcal{L}_{\text{color,scene}} + \mathcal{L}_{\text{clip,scene}}$. The color loss again considers the similarity with the target image $\mathcal{L}_{\text{hist}}(I, I_t)$. The scene being optimized I is rendered with the stylized structure \mathcal{W} and the current estimates of the base colors. The clip loss considers both individual objects and the global scene and is calculated as the sum of object clip loss and global clip loss.

$$\mathcal{L}_{\text{clip,scene}} = \lambda_2 \cdot \sum_i \mathcal{L}_{\text{clip}}(I_{M_i}, T_i) + \lambda_3 \cdot \mathcal{L}_{\text{clip}}(I, T). \quad (4)$$

Unlike Sec. 3.2.1, text description T_i for object M_i could be a simple text prompt using the class name, or a detailed text prompt based on user choice. We render individual objects from various angles I_{M_i} and compare them with the text description T_i . The embedding for the scene is also compared against the text embedding T represents the type of scene, for example, ‘a bedroom’. By jointly applying the loss, the base color $c(s_{ik})$ is selected as a representative color that harmonizes nicely with the global context.

3.2.3 Detailed Stylization

The base color is combined with additional details regressed from a neural network to express the detailed local texture. We define *local neural style field (LNSF)* for each object, which generates the local textures added to the base color. The color for point p is defined by the following equation,

$$c(s_p) + \alpha \cdot \mathcal{F}_i(\gamma(p), \phi(p), s_p). \quad (5)$$

We train a LNSF \mathcal{F}_i per object, which outputs the color to be added to the base color $c(s_p)$, where s_p indicates the part segment id. The color range α maintains the final color to be similar to the base color. $\gamma(\cdot)$ is the positional encoding of the xyz coordinate to capture high-frequency details, and $\phi(\cdot)$ represents the coefficients of the eigenfunctions for the intrinsic geometry using the Laplace-Beltrami operator. Therefore the object-specific neural fields respect the part boundaries and local geometric details.

LNSF for each object are trained with additional style context $\mathcal{L}_{\text{clip}}(I_{M_i}, T_i^+)$. The text prompt T_i^+ augments the object description of T_i with appearance style description t_s , such as ‘minimal style’ or ‘Mid-Century style’. By enforcing the same appearance style description, we can weakly bind the styles of individual objects in the same scene while optimizing separately. We could also use T_i^+ instead of T_i for optimizing the base color, and it shows slightly better results.

Part-aware Geometric Deformation Even though our main focus is to add colors to objects with local part-aware details, we could also concurrently produce part-aware geometric deformation with the same architecture. We can slightly change the LNSF \mathcal{F} to have two branches of output that estimate color and displacement. For each point on the surface, we assign color by Eq. 5 and adjust displacement along the vertex normal direction. To learn the effective deformation, we add geometric loss $\mathcal{L}_{\text{clip}}(I_{M_i}^{geo}, T_i^+)$, where $I_{M_i}^{geo}$ is an image rendering textureless geometry as [27].

Rendering and Implementation Details To render each object, the individual objects M_i are scaled to fit a unit box. The predicted color of each vertex allows the entire mesh to be differentiable rendering through interpolation using [5]. We render individual objects with random augmented backgrounds (white, black, random Gaussian, chess board), which helps the pipeline focus on the foreground object [16, 18]. Inspired by [36], since CLIP has a bias for the canonical pose, we augment the text prompt with the azimuth angle of view, namely ‘front view of’, ‘side view of’ or ‘back view of’. Finally, in Sec 3.2.3, random perspective transformation and random crop boost learning local details. To render a scene, we randomly sampled from pre-define 20 camera poses to evenly cover the entire scene \mathcal{S} .

4. Experiments

Our stylization is first evaluated for individual objects in Sec. 4.1. We evaluate the quality of textured object meshes and also assess that our pipeline can discover part segments to assign realistic stylization. Then we demonstrate the stylization results of room-scale scenes with multiple objects in Sec. 4.2.

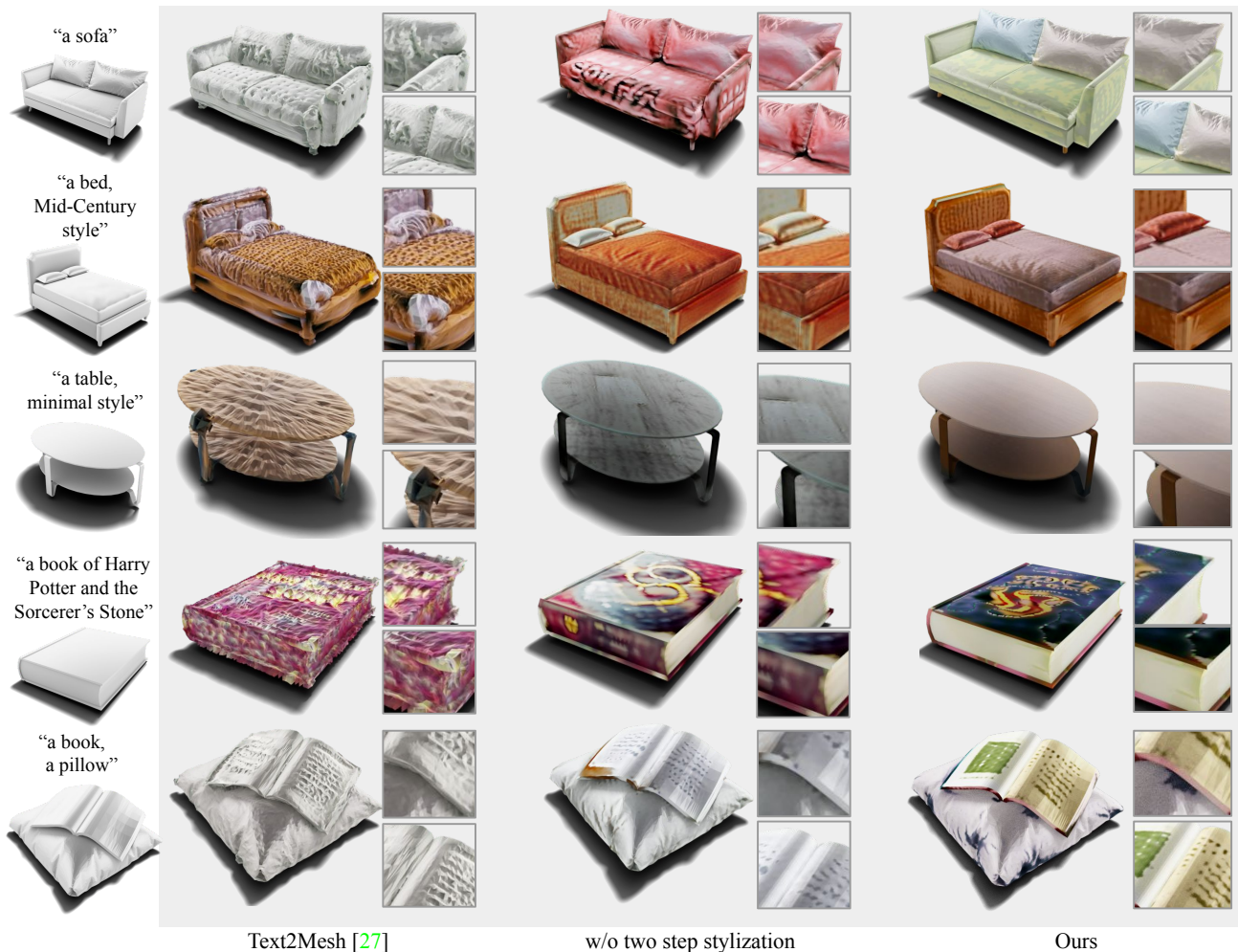


Figure 5. Visual comparison with baselines. Our method utilizes the found part information and generates detailed textures through the designed network in combination with a coarse-to-fine stylization scheme. As a result, it generates a globally harmonious 3D style that clearly classifies part segments. We can handle diverse text such as simple categories or styles, detailed descriptions, or text that includes different two object categories.

4.1. Object Stylization

We show that our method can stylize various objects and produce realistic 3D assets to populate virtual scenes. We use object meshes of various types and sizes, collected from Turbo Squid [48], 3D-FUTURE [13], and Amazon Berkeley Objects [10]. For detailed stylization, we subdivide the object into an average of 119769 faces and 60437 vertices. For large general objects such as beds, sofas, tables, etc., we use class labels and the text input for the specific style as ‘a [class label], [specific] style’. However, we need a detailed explanation for small objects that cannot be explained only by class labels, such as books. For these objects, we configure detailed text individually, for example, ‘a book of Harry Potter and the Sorcerer’s Stone’.

Figure 5 shows the renderings of stylized results. We render objects in Blender [11] with a fixed lighting setup.

The input mesh and the text are also provided in the left column. Since our method utilizes the obtained part information and coarse-to-fine stylization scheme with a two-step approach, it creates a more realistic texture with clear boundaries for each part of the 3D assets. Text2Mesh [27] is another text-driven 3D stylization approach, which adds the RGB color and deformation fields on the vertices of mesh. While it augments the detailed variations on the input mesh, the deformation map can occasionally introduce undesired artifacts and the part boundaries are only roughly estimated. We also provide the results of an ablated version that directly uses LNSF without a two-step stylization scheme. Our part discovery module and coarse-to-fine stylization scheme play a critical role in producing realistic assets with high-quality stylization. Figure 1 contains more results on diverse objects.

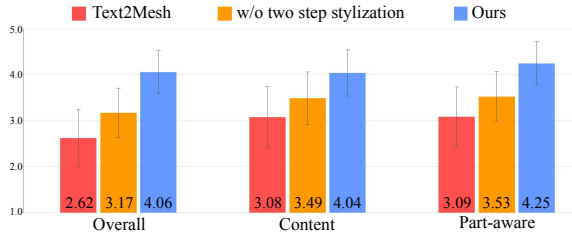


Figure 6. Results of user study for object stylization

Because stylization is a subjective task, and no ground truth or metrics exist, we conduct a user study for quantitative evaluation. We ask 98 users to rate the quality of generated outputs on a scale of 1 (worst) to 5 (best) in response to the following questions: (Q1) ‘How natural are the output results?’, (Q2) ‘How well does the output contain text information?’, and (Q3) ‘How well does the output reflect part information?’ Figure 6 contains the mean and the standard deviation of the scores. Our method outperforms competing methods in all aspects. Therefore we conclude that our method generates a realistic texture that abides by input text description and semantic part information.

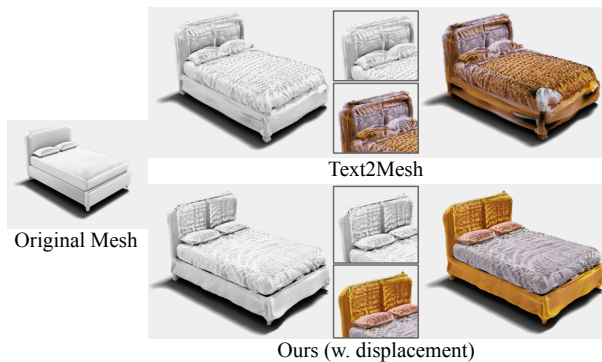


Figure 7. With a slightly modified LNSF \mathcal{F} , we can simultaneously generate the texture with both color and displacement fields, and all of the generated style information respects the part information discovered.

We also provide a parallel comparison against Text2Mesh by incorporating deformation fields with our approach. Our original method preserves the original shape of the mesh, while Text2Mesh deforms vertices along the normal direction in addition to generating texture. We can make a similar version of our LNSF and produce additional deformation as described in Sec. 3.2.3. Figure 7 shows the deformed results from the source mesh compared to Text2Mesh. Since our approach explicitly considers the part information with the two-stage approach, our results with deformation fields also respect different semantic parts of the object.

Part Discovery As a side product of object stylization, we can discover different parts with distinguished boundaries

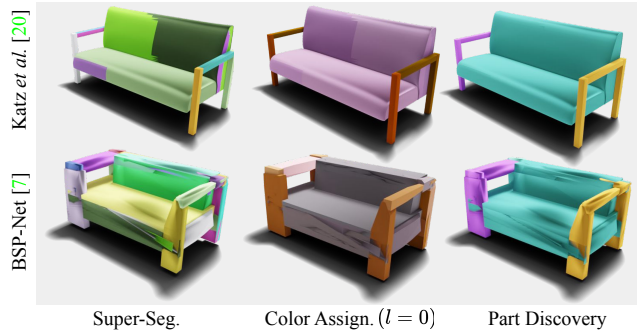


Figure 8. The robustness against initial super-segmentation. The super-segments of the top and the bottom rows are generated by [20] and [7], respectively (left). Starting from the initial color assignment (middle), our approach stably finds part decomposition to assign different base colors and therefore different texture information (right). The text *a chair* is used.

that can guide a realistic color assignment (Sec. 3.2.1). Our part discovery combines geometric super-segments and implicit clues from image-text embeddings. The initial super-segments guide the algorithm to follow geometric feature lines, and the final results stably find part information despite different initializations. Figure 8 shows examples initialized with Katz *et al.* [20] (top) and BSP-Net [7] (bottom). Even for input mesh with bad topology or different initialization, our iterative part discovery quickly finds visually coherent parts without any training with segmentation labels.

4.2. Scene Stylization

Now we demonstrate that our text2scene framework can quickly generate a realistic texture for a room with multiple objects. Recall that the 3D scene $\mathcal{S} = \{\mathcal{W}, \mathcal{O}\}$ is composed of the structure components \mathcal{W} and a set of objects \mathcal{O} . We use the same objects as described in Sec. 4.1 and arrange them to constitute scenes. As there is no existing dataset composite of complete object meshes with labels, we built a total of four scenes: two bedrooms and two living rooms. Each scene contains an average of 20 objects of various sizes and classes. Additionally, we provide a target image I_t for the color distribution, and a text prompt describing the desired style t_s . While it can be daunting to define the desired style for the entire scene, images and texts can provide a simple way to deliver the information.

Figure 2 shows stylized results of the same geometry, but observing various target images and style prompts. The generated textures respect the semantic labels of various furniture and different parts and contain localized diverse details. This is in contrast to many prior stylization methods where the fine perturbations are spread throughout the scenes. We also show scenes with different types of rooms containing different objects, but stylized based on the same

target image in Fig. 9. Additional results of various input configurations are available in the supplementary material. Note that our target image I_t does not need to restrict to an indoor image, and can be replaced such as natural photographs. And by changing random seeds, diverse results can be obtained from the same input conditions. Also, since we stylize the entire object, we can easily edit the 3D scene through object relocation. These results can also be found in the supplementary.



Figure 9. Result for the different arrangement of objects. We used target image number 3 in Fig. 2 in both spaces.

	<i>-retrieval</i>	<i>-hist,glo</i>	<i>-detail</i>	Ours
(Q1): Realistic	2.23(±0.48)	3.68(±0.49)	4.02(±0.49)	4.22(±0.42)
(Q2): Color	2.09(±0.48)	2.63(±0.59)	3.95(±0.47)	3.86(±0.49)

Table 1. Results of user study for scene stylization



Figure 10. Ablation results for the scene. We used target image number 1 in Fig. 2 in all spaces.

We also provide users’ evaluation of the quality of our stylization in Table 1. Users evaluate the results by answering the following questions: (Q1) ‘How realistic is the output result?’ (Q2) ‘How similar is the color distribution of the scene to the given images?’ Users assess the overall quality of the outputs, and how well they match the target image. Since there are no previous works that use the same setting, we compare the results of ablated versions: *-retrieval* replaces the separate texture retrieval module (Sec. 3.1) with a stylization network of objects for structure components; *-hist,glo* removes the color loss and the global clip loss for the base color assignment (Sec. 3.2.2); and *-detail* removes the detailed stylization step for objects (Sec. 3.2.3). The responses of (Q1) indicate that Text2Scene generates high-quality scenes and each of the components plays a crucial role to achieve reality. The effect of texture retrieval is the most prominent. The color distribution results (Q2) indicate that the base color assignment is critical. The added local details are more important for the realism

of the results. Figure 10 shows exemplar images of the ablated versions used for the user study.

GPU Cost and Scalability Text2Scene first assigns base colors to discovered parts for all objects in the scene, and generates details of objects individually. The cardinality for the base color assignment is only a few hundred and it does not require much memory and allows us to consider the whole scene while expanding the scale. The most memory-intensive process is the detail generation using a neural network, which only processes a single object at a time. Therefore the entire process can be trained only on a single 11 GB GPU, making it an accessible tool for casual users. In a single GPU, the base color allocation of the entire space takes 5 hours, and learning the details for each object takes 10 minutes.

Limitations While our approach results in scene stylization, the pipeline separately handles individual objects after the base color assignment. This is a practical choice for scalability but may lack an understanding of the context of the entire space. Instead, we rely on the text description to weakly bind the objects into a similar style. We can design the pipeline to receive an additional input with a texture map or a lightweight network and extend our model to better observe the holistic scene context within a limited GPU memory. Also, our pipeline requires a class label or optional text description per object, which can be further automated.

5. Conclusions

We introduce Text2Scene, a novel framework to generate a texture for 3D scenes. Our hierarchical framework can handle a variety of objects, including highly detailed textures for objects such as book or paintings. By leveraging the representation power of pre-trained CLIP, the framework does not require any 3D datasets with texture or part annotation. Given the 3D mesh models and the class labels or text descriptions of objects, our framework easily produces stylized results by picking a target image and a simple text description. We hope that Text2Scene to facilitate the automatic interior recommendation or realistic virtual space generation.

Acknowledgements This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00208197) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub). Inwoo Hwang is supported by Hyundai Motor Chung Mong-Koo Foundation. Young Min Kim is the corresponding author.

References

- [1] Mahmoud Afifi, Marcus A. Brubaker, and Michael S. Brown. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4
- [2] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, June 2022. 3
- [3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022. 3
- [4] Kang Chen, Kun Xu, Yizhou Yu, Tian-Yi Wang, and Shi-Min Hu. Magic decorator: Automatic material suggestion for indoor digital scenes. *ACM Trans. Graph.*, 34(6):232:1–232:11, Oct. 2015. 3
- [5] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 5
- [6] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3
- [7] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bspnet: Generating compact meshes via binary space partitioning. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 7
- [8] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 3
- [9] Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, and Hao Zhang. Bae-net: Branched autoencoder for shape co-segmentation. *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 2
- [10] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022. 6
- [11] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6
- [12] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2
- [13] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d future shape with texture. *International Journal of Computer Vision*, pages 1–25, 2021. 6
- [14] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In *ICCV*, 2021. 2
- [15] Lukas Höllein, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3d scene reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6198–6208, 2022. 2
- [16] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 3, 5
- [17] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3
- [18] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. *CVPR*, 2022. 3, 5
- [19] Jaebong Jeong, Janghun Jo, Sunghyun Cho, and Jaesik Park. 3d scene painting via semantic image synthesis. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [20] Sagi Katz and Ayellet Tal. Hierarchical mesh decomposition using fuzzy clustering and cuts. *ACM Transactions on Graphics*, 22, 06 2003. 2, 4, 7
- [21] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Popa Tiberiu. Clip-mesh: Generating textured meshes from text using pretrained image-text models. *SIGGRAPH Asia 2022 Conference Papers*, December 2022. 2, 3
- [22] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. 3
- [23] Lukas Koestler, Daniel Grittner, Michael Moeller, Daniel Cremers, and Zorah Löhner. Intrinsic neural fields: Learning functions on manifolds. In *ECCV*, 2022. 3
- [24] Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J Guibas, and Minhyuk Sung. Partglot: Learning shape part segmentation from language reference games. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [25] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 3
- [26] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3
- [27] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. *arXiv preprint arXiv:2112.03221*, 2021. 2, 3, 5, 6

- [28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3
- [29] Shailesh Mishra and Jonathan Granskog. Clip-based neural neighbor style transfer for 3d assets. In *arXiv*, 2022. 2, 3
- [30] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2
- [31] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings IEEE International Conf. on Computer Vision (ICCV)*, 2019. 3
- [32] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation, 2019. 3
- [33] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1
- [34] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [35] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 3
- [36] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3, 5
- [37] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 3
- [38] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 2
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3, 4
- [40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 3
- [41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3
- [42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arxiv:2208.12242*, 2022. 3
- [43] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv preprint arXiv:2110.02624*, 2021. 3
- [44] Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik. Match: Differentiable material graphs for procedural material capture. *ACM Trans. Graph.*, 39(6):1–15, Dec. 2020. 3, 4
- [45] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions, 2020. 3
- [46] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022. 3
- [47] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3
- [48] TurboSquid. Turbosquid 3d model repository. In <https://www.turbosquid.com/>, 2022. 6
- [49] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *CVPR*, 2021. 3
- [50] Yu-Ying Yeh, Zhengqin Li, Yannick Hold-Geoffroy, Rui Zhu, Zexiang Xu, Miloš Hašan, Kalyan Sunkavalli, and Manmohan Chandraker. Photoscene: Photorealistic material and lighting transfer for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18562–18571, June 2022. 3
- [51] Wang Yifan, Lukas Rahmann, and Olga Sorkine-hornung. Geometry-consistent neural shape representation with implicit displacement fields. In *International Conference on Learning Representations*, 2022. 3
- [52] Kangxue Yin, Jun Gao, Maria Shugrina, Sameh Khamis, and Sanja Fidler. 3dstylenet: Creating 3d shapes with geometric and texture style variations. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 2
- [53] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *ECCV*, 2022. 3
- [54] Dongsu Zhang, Changwoon Choi, Inbum Park, and Young Min Kim. Probabilistic implicit scene completion. In *International Conference on Learning Representations*, 2022. 1