MTScan: Material Transfer from Partial Scans to CAD Models

Xiangyu Su
Shenzhen University
Shenzhen, China
xiangyv.su@gmail.com

Sida Peng
Zhejiang University
Hangzhou, China
pengsida@zju.edu.cn

Oliver van Kaick
Carleton University
Ottawa, Canada
Oliver.vanKaick@carleton.ca

Hui Huang
Shenzhen University
Shenzhen, China
huihuang@szu.edu.cn

Ruizhen Hu*
Shenzhen University
Shenzhen, China
ruizhen.hu@gmail.com

Abstract

We introduce a method for transferring material information from a partial scan to a CAD model by establishing a dense correspondence between the scan and the CAD model. Our method is enabled by a pipeline composed of a material decomposition network, a geometry mapping network, and material completion networks. Specifically, given a single RGB-D source image and a target CAD model aligned to the scan, we employ a material decomposition network to extract material and illumination parameters from the image. Next, we sample point clouds from the image and CAD model, and establish a dense correspondence between the two point clouds with a geometry mapping network, which maps the point clouds to a shared template space where correspondences can be derived from closest points and aligned UV maps can be obtained. Finally, based on the established correspondence, we transfer the decomposed material information from the source to the target, and further perform material completion via diffusion on the point clouds and in the UV space. We demonstrate with qualitative and quantitative evaluations that our method is able to obtain more accurate material transfers than previous work in challenging input cases with imperfect shape alignment, so that the shapes with transferred materials better resemble the scanned shapes.

Keywords: Appearance transfer Relightable Materials 3D shape modeling Appearance Modeling.

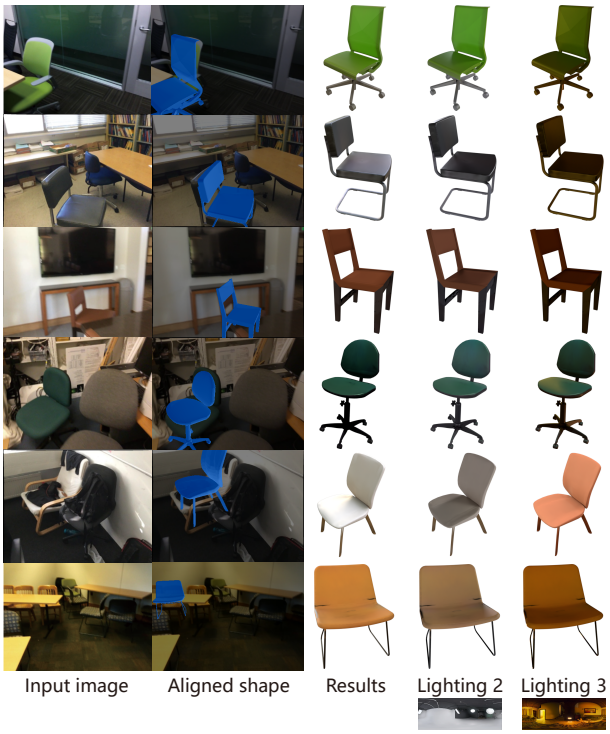

Figure 1. Our material transfer method is robust to imperfect shape retrieval and alignment. All cases in this figure have varying degrees of misalignment and occlusion, but our method is still able to get reasonable results. The figure shows the input image and aligned shape (left two columns), the rendering of the shape with transferred SVBRDF in the environment lighting estimated by the material decomposition network (middle column), and rendering in two different illumination conditions (right two columns).

1. Introduction

Many graphics applications such as AR/VR/MR, robotics, and simulation need realistic scenes with high-quality meshes. To automatically generate such a scene, various methods attempt to retrieve CAD models from CAD libraries and align them to objects in real-world captured RGB images [16, 15] and RGB-D scans [1, 2, 25]. But most of these works only focus on obtaining accurate ge-

ometric alignment, and thus generated scenes lack appearance information such as textures and material properties, which is essential scene information for most graphics and vision tasks. In this paper, we aim to transfer materials from captured RGB-D images to aligned CAD models to produce high-quality relightable objects.

Recently, some methods [30, 19] have been proposed for transferring appearance materials from images to 3D shapes. They first establish part-level correspondences between the segmented shape and the observed image. Then, based on the content of corresponding image regions, they retrieve similar materials from a pre-collected material library for each part of the target shape. Although these methods can produce photorealistic relightable objects, they rely on high-quality mesh segmentation, which is challenging due to the complexity of shape understanding. Moreover, they project 3D shapes to 2D space and then establish 2D correspondences with images through image alignment and translation. Thus, these methods require a relatively complete observation of the target shape. If the object is occluded by other objects in the scene or captured incompletely, the structural information in the images may be disrupted, resulting in inaccurate transfer results. However, such occlusion and incomplete observation are common occurrences in real-world scans.

In this paper, we propose a novel framework to transfer materials from a real-world RGB-D image to an aligned CAD model. Our key contribution is to establish dense correspondences between scans and shapes, which can help us easily transfer 2D materials to 3D models without any segmentation of the shapes. To achieve this, we establish a shared template space for shapes of the same category by learning a deformation neural field that automatically aligns intra-category shapes. Then, the input RGB-D image and CAD model are mapped to this space for establishing their point-wise correspondences. The shared template space eliminates shape variations between the observed object and retrieved CAD model, effectively improving the matching accuracy. In contrast to implicitly correlating images with 3D shapes in latent space [29, 37], our strategy produces explicit correspondences and thus provides more direct guidance for appearance transfer.

Based on the estimated correspondences, we develop a pipeline to effectively transfer materials from the captured image to the target CAD model. Our pipeline firstly estimates materials from images. Then, the materials are transferred to the target shape based on predicted correspondences, so that we obtain target shapes with partial materials. Finally, the materials of the shape are completed in point and UV space. A challenge for this pipeline is that existing UV mapping techniques tend to produce inconsistent material maps in UV space, thus decreasing the performance of the material completion network. To solve this issue, we design a semantically aligned UV mapping technique to learn a regular UV space that exhibits semantic consistency across intra-category shapes, thereby helping the network in learning the material distribution.

With these contributions, we are able to transfer materials from scans to aligned 3D shapes. The generated meshes have materials closer to the scans when compared to the results of previous work. We demonstrate this improvement with visual and quantitative evaluations of our method, which include a comparison with state-of-the-art methods. We also show the effect of the different components of our method on the results.

In summary, our contributions include the introduction of:

- A geometry mapping network that maps a point cloud in object space to a shared template space, and infers an aligned UV mapping simultaneously;
- A material completion network that combines point and UV diffusion to generate material maps from coarse to fine based on a point cloud with partial materials;
- A material transfer method, which includes material decomposition, material transfer, and material completion, that can handle imperfect retrieval and alignment.

2. Related work

2.1. Alignment of scans to CAD models

With the availability of large-scale 3D shape datasets [7, 12], the recomposition of 3D scenes using CAD models has made significant progress in recent years. Several approaches have been introduced to perform CAD retrieval and alignment to images [21, 16, 15] or scans [1, 2, 14, 25]. For image-based CAD retrieval and alignment, Izadinia et al. [21] iteratively optimize the position and scale of the object to best match the input image. Gumeli et al. [16] establish a correspondence between 2D and 3D and then use a differentiable robust Procrustes method to continuously optimize the alignment. More recently, Gao et al. [15] propose a weakly-supervised method for this task, where they use a diffusion model to model the probability of a CAD model's shape, pose, and scale based on the input image. For scan-based CAD retrieval and alignment, Avetisyan et al. [1] learn a joint embedding between real and synthetic shapes to compute corresponding heatmaps. These maps represent the likelihood that an input key point in the scan matches a voxel of the CAD model. Based on the heatmaps, the alignment of shapes is optimized. Furthermore, Avetisyan et al. [2] predict layout elements jointly, enhancing the global consistency of the predicted scene. Di et al. [14] learn retrieval and deformation of shapes in an unsupervised manner. Despite the continuous improvement in enhancing the

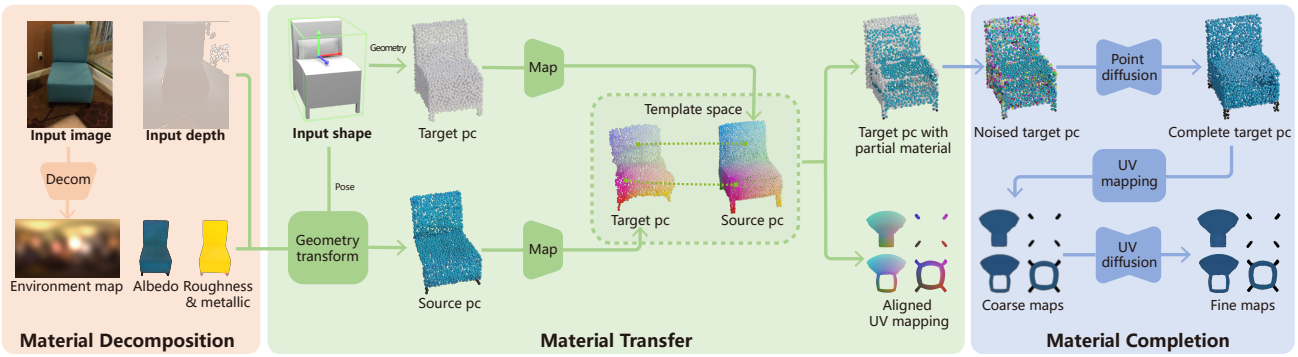

Figure 2. Overview of our method for transferring materials from a single scanned RGB-D frame to an aligned shape. Given the single frame and shape with 9 DoF pose as input, we first extract an SVBRDF (including albedo, roughness, and metallic properties) and an illumination map from the RGB-D image using a material decomposition network. Next, we recover a point cloud from the depth image and transform it to object space based on the input shape's pose. Simultaneously, we sample a point cloud from the target shape. Then, given the source point cloud from the scan and the target point cloud from the shape, we use a geometry mapping network to map the two point clouds to a shared template space, where a point-wise correspondence can be established and an aligned UV mapping can be obtained. Based on the correspondence, we transfer the SVBRDF from the image to the target point cloud. Finally, we complete the point cloud with partial materials in a coarse-to-fine manner to obtain the final material map.

accuracy of alignment in this series of works, imperfect retrieval and alignment are still unavoidable. Therefore, a robust material transfer method is needed.

2.2. Appearance transfer from 2D to 3D

The goal of 2D-3D appearance transfer is to transfer textures or materials from images to shapes. Texture transfer only handles RGB information. Wang et al. [36] transfer textures from an input image to a colletion of 3D shapes. Their method extracts texture from an image and transfers it to similar shapes based on a 2D-to-3D correspondence. Then, using the selected shape as an intermediary, the texture is transferred to other 3D shapes. Similarly, Huang et al. [20] manually create a proxy model that resembles the image as an intermediary domain for texture transfer. More recently, a series of texture generation methods have been proposed. When conditioned on a given image, these methods can also be used for texture transfer. Oechsle et al. [29] learn an implicit texture field for a category of shapes, which can predict points' color directly. Yu et al. [37] use point diffusion and UV diffusion to generate texture maps for meshes. However, such methods establish implicit correspondences between images and shapes, and may not yield accurate transfer results. By utilizing powerful pre-trained image diffusion models, significant progress has been made in text-based texture generation for 3D shapes [8, 6, 24]. These methods typically use pre-trained image models to inpaint shapes from different views. When adapting these pipelines to texture transfer tasks [34], 3-5 image conditions are needed to fine tune a pre-trained diffusion model. Material transfer handles multidimensional appearance properties, including albedo and other physical components. Nguyen et al. [28] transfer materials from an image or video to 3D geometry. Their method uses global optimization to process the entire 3D scene. Rematas et al. [33] align a target 3D shape to an image and extract materials for each part in 3D. Park et al. [30] take an image and shape with material segmentation as input, and align a projection of the 3D shape onto the image to establish part-level correspondence. Afterward, a material perdition network is used to predict materials from a pre-collected high-quality material dataset. The image alignment method requires the input images and shapes to be similar in structure. To address this limitation, Hu et al. [19] use an image translation network to establish more structurally robust semantic correspondences. However, their method requires 3D models with detailed semantic segmentation. In contrast to previous methods, our method does not require the 3D model to have any geometric segmentation. Furthermore, our method also does not rely on a pre-collected high-quality material dataset since it directly extracts material information from an image.

3. Overview

Figure 2 shows an overview of our method. The inputs to our method are an RGB-D image and an aligned 3D shape. We assume that the image is given with an object mask [22] and that the 3D shape is provided with a 9-DoF pose, where the retrieval and alignment of the shape can be obtained automatically with existing methods [23, 16, 15]. Note that our method does not necessitate any shape segmentation as input.

Our method starts by extracting environment and material information from the RGB image. We use a material

decomposition network to estimate an illumination map and SVBRDF from the image, including albedo, roughness, and metallic properties.

Next, we transfer the estimated material properties from the source image to the target shape to obtain a 3D shape with partial material definitions. To be more specific, given the object mask of the input RGB-D frame, we obtain a source point cloud in the world space, which is then transformed to object space by using the target shape's pose. Simultaneously, we sample a target point cloud in object space from the input 3D shape by farthest points sampling. Then, our geometry mapping network takes the source and target point clouds as input, and outputs two point clouds transformed to a shared template space, along with their UV coordinates in an aligned UV space. In the shared template space, we establish point-wise correspondences based on closest points measured by Euclidean distances. Note that the use of template space correspondences is a key component of our method to address an imperfect 3D shape retrieval and alignment. Based on the correspondences, we transfer the SVBRDF from the source to the target point cloud, obtaining a point cloud with partial material definitions.

Finally, we generate complete material maps for the target mesh based on the point cloud with partial material definitions and aligned UV mappings. Our material completion network combines 3D and 2D diffusion models to generate material maps in a coarse-to-fine manner. Specifically, taking the target point cloud with partial materials and UV coordinates as input, the network first completes the material properties in the point cloud with a 3D diffusion step. Then, a coarse material map is generated by mapping materials from the point cloud to the aligned UV space, which will be refined in an image diffusion stage.

4. Method

In this section, we explain the details of the three key components of our method: material decomposition, material transfer, and material completion.

4.1. Material decomposition

The goal of material decomposition is to extract an SVBRDF from the input image. In addition, we predict the illumination of the scene as a sub-task to help SVBRDF estimation. The SVBRDF is represented as albedo, roughness, and metallic per-pixel properties [3, 27], while the illumination map is represented with 12 spherical Gaussians (SG), where each SG is defined by amplitude, axis, and sharpness. Similarly to Boss et al. [4], we only estimate the amplitude and set the axis and sharpness to cover a unit sphere.

The architecture of our material decomposition network $\mathcal{D}$ is inspired by Collins et al. [10]. Taking the RGB image I_{rgb} and object mask I_{mask} as input, a UNet-based model with a ResNet-34 backbone estimates the SVBRDF $\hat{I}_{\text{svbrdf}}$. The UNet has a common encoder and multi-head decoder to predict each component of the SVBRDF separately. For the environment lighting $\hat{I}_{\text{light}}$, we use another encoder network followed by 3 fully-connected layers. The loss function for the network training is defined as:

$$\mathcal{L}_{\mathcal{D}} = \alpha_1 MSE(\hat{I}_{\text{svbrdf}} - I_{\text{svbrdf}}) + \alpha_2 MSE(\hat{I}_{\text{light}} - I_{\text{light}}), \quad (1)$$

where MSE is the mean squared error loss, and I_{svbrdf} and I_{light} are the ground truth SVBRDF and lighting, respectively.

4.2. Material transfer

The goal of this module is to transfer the estimated SVBRDF from the source image to the target 3D shape. Since we only have the observation of the scanned object from a single viewpoint, and this observation may even be incomplete due to occlusions in the scene, we first obtain a point cloud with partial materials in this module, which will be completed in the next stage (Section 4.3).

The key challenge here is how to establish a dense correspondence between the two shapes to guide the material transfer, given a single-view (possibly incomplete) observation with imperfect shape retrieval and alignment. To achieve this, first, two point clouds in object coordinates are sampled from the input RGB-D scan and 3D shape. Then, our geometry mapping network takes the two point clouds as input, and transforms them to a shared 3D template space and aligned 2D UV space. In the shared template space, we establish point-wise correspondences based on closest points. Finally, based on the correspondence, we directly transfer the predicted SVBRDF from the image to the shape.

4.2.1 Obtaining point clouds in object space

First, we obtain two point clouds in object space from the inputs. For the input scan, given the depth image I_{depth} and object mask I_{mask}, we extract a point cloud in world coordinates by taking pixels covered by the mask and defined by the pixel coordinates and depth. Then, we transform the point cloud to object space using the 9-DoF pose of the target shape S_{pose}, to obtain the source point cloud P_s. For the target shape, we sample a fixed number of points on the surface of the shape S_{geometry} by farthest point sampling (FPS) to obtain the target point cloud P_t.

4.2.2 Learning a shared template space and aligned UV space

The next step of the method is to obtain a dense correspondence between the two point clouds in object space. Sim-

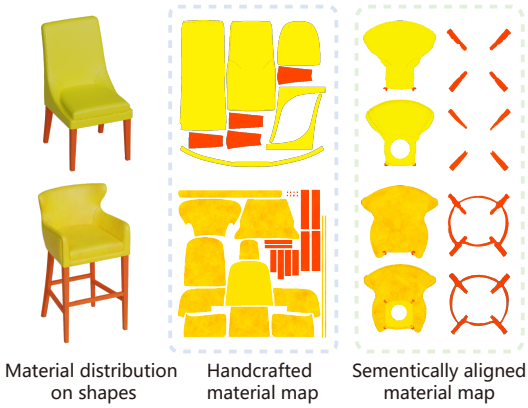

Figure 3. Exemplars of semantically aligned material maps for two shapes, showing the material distribution on the shapes (left column), the handcrafted material maps (middle column), and the semantically aligned material maps (right column). Note that only roughness and metallic distributions are shown in the figure, and the color map is represented as [R: nothing (always set to 1), G: roughness (0-1), B: metallic (0-1)].

ilarly to Yu et al. [13], we employ a shared template space for shapes in the same category, and learn how to map specific shapes to this template field. Correspondences can then be easily found after mapping the two point clouds to the shared template space.

Moreover, given that objects typically have contiguous regions with the same material, e.g., the soft pillow of a chair seat, or the hard surface of a chair's wooden legs, differently from existing UV mapping methods that are oblivious to such information [5], we would like to maintain this material distribution in the mapping to reduce the difficulty of subsequent material completion. Thus, inspired by Chen et al. [9], we also learn a semantically aligned UV mapping. As shown in Figure 3, this aligned UV mapping maintains the material distribution of the shapes.

To learn the geometry and UV mappings, we introduce a geometry mapping network. Our network jointly learns a latent shape space for a collection of shapes from the same category, a shape mapper to map a shape to the template space based on the shape latent code, and a UV mapper to map a template space shape to the aligned UV space. Note that the UV mapping is based on the template space, so that the mapping of different shapes naturally results in a semantic alignment.

Figure 4 shows a simplified diagram of the geometry mapping network $\mathcal{G}$. The shape mapper takes the shape code z and a point in object space as input, and outputs the offset v of the point to the template space. The UV mapper takes a point in the template space as input, and outputs the point's UV coordinates. The loss function that guides the training of the network is defined as:

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{latent} + \mathcal{L}_{map} + \mathcal{L}_{uv}, \tag{2}$$

where $\mathcal{L}_{latent}$ is the latent space loss, $\mathcal{L}_{map}$ is the shape mapper loss, and $\mathcal{L}_{uv}$ is the UV mapper loss.

Specifically, $\mathcal{L}_{latent}$ is a regularization term for the shape latent space. $\mathcal{L}_{map}$ is used to ensure a robust mapping from object space to template space, and is defined as:

$$\mathcal{L}_{map} = \beta_1 \mathcal{L}_{sdf} + \beta_2 \mathcal{L}_{normal} + \beta_3 \mathcal{L}_{smooth}, \tag{3}$$

where $\mathcal{L}_{sdf}$ is a Signed Distance Function (SDF) reconstruction loss, since the correspondence is learned in terms of an SDF, $\mathcal{L}_{normal}$ is a key loss term which ensures that the normal of a surface point is highly correlated with its semantic information, and $\mathcal{L}_{smooth}$ encourages the smoothness of the SDF field.

$\mathcal{L}_{uv}$ is used for learning the semantically aligned UV mapping:

$$\mathcal{L}_{uv} = \beta_4 \mathcal{L}_{prior} + \beta_5 \mathcal{L}_{dist}, \tag{4}$$

where $\mathcal{L}_{prior}$ is a prior loss to guide the UV mapping and $\mathcal{L}_{dist}$ is a distortion loss to minimize the generated map's distortion when mapping to 3D Shapes.

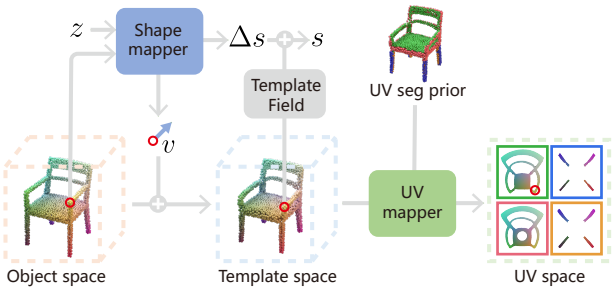

Figure 4. Geometry mapping network used in our work. Given a shape code z and a point in object space, the shape mapper predicts the offset v of the point to map it to the learned template space and a correction SDF value Δs . Then, the point is offset to map it to the template field and predict its SDF value $\tilde{s}$, which is added to Δs to provide the final SDF value s. Note that SDF values are only used for training. Next, the UV mapper maps the point to the semantically-aligned UV space based on the UV segmentation prior in template space.

4.2.3 Optimizing the shape code

To map the point clouds to the template space, we first obtain a shape code for each input. For the target point cloud P_t, we obtain its shape code z_t during the training of the geometry mapping network. Moreover, since the learned shape embedding is in a continuous space and the inference of the geometry mapping network can be performed for an arbitrary number of samples, for the source point cloud P_s, we optimize a shape code z_s that best explains the single view observation via a maximum-a-posteriori estimation. The objective function is defined as:

$$\arg\min_{\boldsymbol{z_s}} \sum_{(\boldsymbol{x}_j, \boldsymbol{s}_j) \in P_s} \beta_6 \mathcal{L}_{sdf}(\mathcal{G}(\boldsymbol{z_s}, \boldsymbol{x}_j), s_j) + \beta_7 ||\boldsymbol{z_s}||_2^2, \tag{5}$$

where x_j are the point coordinates, s_j is the SDF value in P_s, which is always set to 0 since all points are sampled from the surface, and $\mathcal{G}$ is the trained geometry mapping network.

4.2.4 Establishing correspondences

Given the two point clouds in template space, point-wise correspondences can be established based on point distances. We establish the correspondence in two stages. Firstly, for each point in the source point cloud, we find its nearest point in the target point cloud. Since the number of samples in the source point cloud is much larger than in the target, a point in the target point cloud may be the closest point to several points in the source. Thus, in the second stage of the matching, for each point in the target point cloud, we select the source point with the shortest distance as the corresponding match. Based on the point-wise correspondence between the source and target point clouds, along with the pixel-wise correspondence between the input image and source point cloud, we transfer the estimated SVBRDF from the image to the target point cloud.

4.3. Material completion

After material decomposition and material transfer, we obtain a point cloud of the target shape with partial materials, where the materials were transferred directly from the input image. Then, the goal of the material completion is to generate a complete material map for the target 3D mesh while preserving the transferred materials. Inspired by Yu et al. [37], we use a material completion network for this task, which combines point diffusion and UV diffusion models to generate SVBRDF maps from coarse to fine.

4.3.1 Point diffusion

In the coarse stage, given the point cloud with partial materials, we use a point diffusion model to complete the materials. The point cloud P is defined as $\{C, M^{pc}\}$, where $C \in \mathbb{R}^{4096\times3}$ are the point coordinates and $M^{pc} \in \mathbb{R}^{4096\times5}$ are the material properties on the point cloud. During training, given a point cloud sample $\{C,M_0^{pc}\}$, where $M_0^{pc} = (M_{known}^{pc}, M_{unknown}^{pc})$ and M_{known}^{pc} correspond to the material properties derived from the material transfer, we add noise on M_0^{pc} to obtain the noised material M_t^{pc}. The noise adding process is defined as:

$$M_t^{pc} = (M_{known}^{pc}, \sqrt{\alpha_t} M_{unknown}^{pc} + \sqrt{1-\alpha_t}\epsilon), \quad (6)$$

where $t \in \{0, 1, ..., T\}$ is the time step, and α_t is the noise level which is dictated by t and $\epsilon \in \mathcal{N}(0, 1)$.

Given the noised point cloud $\{C, M_t^{pc}\}$, time step t, and shape condition maps x_{shape} as input, where $x_{shape} = [x_{normal}, x_{coor}, x_{mask}]$ is the concatenation of pre-computed maps, which is generated by mapping 3D normal and coordinates to 2D through the aligned UV mapping, we employ a point denoising network to predict the material $\hat{M}_0^{pc}$, as shown in the top of Figure 5:

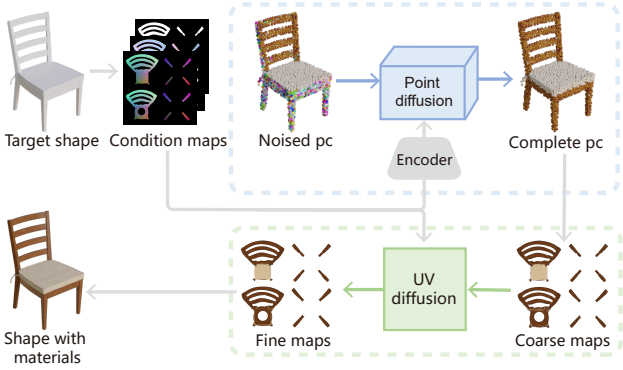

Figure 5. Material completion network used in our work, where the core components are based on the method of Yu et al. [37]. Given a point cloud with partial materials as input, we generate material maps for the 3D mesh from coarse to fine. In the coarse stage (top of the figure), we complete the material on the point cloud. Then, we map the generated point cloud to UV space to obtain a coarse material map. This map is refined in UV space in the fine stage (bottom of the figure).

$$\hat{M}_0^{pc} = \mathcal{C}_{\theta_1}^{coarse}(\{C, M_t^{pc}\}, t, \mathcal{E}_\phi(x_{shape})), \quad (7)$$

where $\mathcal{C}_{\theta_1}^{\text{coarse}}$ is the point denoising network with parameters θ_1, and $\mathcal{E}_\phi$ is a light shape encoder to extract a global shape embedding from x_{shape}.

The loss function for the training of the point denoising network is defined as:

$$\mathcal{L}_{\mathcal{C}}^{coarse} = \gamma_1 \left\| M_0^{pc} - \hat{M}_0^{pc} \right\|^2. \quad (8)$$

That is, we directly measure the difference between the denoised result $\hat{M}_0^{pc}$ and the original input M_0^{pc} because we found that this is more stable during training. Additionally, since the known materials in the point cloud are always fixed during the forward process, we can mask away the known materials and only compute the loss for unknown materials. In addition, we use an augmented PVCNN [26, 37] as point denoising network backbone and an MLP network as shape encoder. During inference, the materials transferred from the image are fixed, and we perform denoising only in the ramaining areas.

4.3.2 Material refinement

In the point diffusion stage, we obtain a point cloud with complete materials. Based on the UV mapping generated in Section 4.2.2, we can map and interpolate materials to UV space to obtain coarse material maps M_{coarse}^{map}. To further

refine the coarse maps, we employ an additional 2D diffusion model in UV space, as shown in the bottom of Figure 5. The main difference from Yu et al. [37] is that we use a semantically aligned UV mapping, which has the advantage of maintaining the material distribution patterns that appear in the 3D models, significantly reducing the difficulty of material generation. Furthermore, as Yu et al. [37] mentioned, their method struggles to generate seamless results when there are too many fragmented cuts of the UV map. Our new UV mapping also addresses this limitation since there are only 4 cuts for most of the chairs.

During training, given the material map $M_0^{map} \in \mathbb{R}^{512\times512\times5}$, the noised map M_t^{map} is defined as:

$$M_t^{map} = \sqrt{\alpha_t} M_0^{map} + \sqrt{1-\alpha_t}\epsilon, \tag{9}$$

where $t \in \{0, 1, ..., T\}$ is the time step, and α_t is the noise level which is dictated by t and $\epsilon \in \mathcal{N}(0, 1)$. In order to maintain the information in the coarse material map M_{coarse}^{map}, we take M_{coarse}^{map} as an input condition for the 2D denoising network. The denoising process is defined as:

$$\hat{M}_0^{map} = \mathcal{C}_{\theta_2}^{fine}\left(M_t^{map}, M_{coarse}^{map}, x_{shape}, t\right), \tag{10}$$

where $\mathcal{C}_{\theta_2}^{fine}$ is the image denoising network with parameters θ_2. The loss function for the training is defined as:

$$\mathcal{L}_{\mathcal{C}}^{fine} = \gamma_2\mathcal{L}_{MSE} + \gamma_3\mathcal{L}_{smooth} + \gamma_4\mathcal{L}_{render}, \tag{11}$$

where $\mathcal{L}_{MSE}$ is the MSE loss between the predicted denoising results $\hat{M}_0^{map}$ and the original input M_0^{map}, $\mathcal{L}_{smooth}$ is a smoothed reconstruction loss to help the network learn the distribution of materials more effectively, and $\mathcal{L}_{render}$ is a rendering loss to measure the rendering difference between the prediction and ground truth.

We use a 2D U-Net combined with self-attention modules as the backbone of the image denoising network. During inference, we map the SVBRDF from the point cloud to UV space so that we obtain coarse material maps M_{coarse}^{map}. Then, we denoise the input to get fine material maps M_{fine}^{map}.

5. Experiments and evaluation

In this section, we first present the implementation details. Then, we show results generated by our method. After that, we show the comparisons with baselines to demonstrate the superiority of our method. Finally, ablation experiments on the main modules proved the effectiveness of the modules.

5.1. Implementation details

5.1.1 Datasets

We use two types of datasets with our method: shape and environment lighting. The environment lighting collection is used to render high-quality images to train our material decomposition network. To ensure diverse realistic backgrounds and lighting conditions, we collect 165 indoor HDRIs from [17]. For the shape collection, we use 3D shapes from ABO [10], where the shapes have fine, realistic materials. We use the ABO dataset to generate training data for all three networks. Moreover, since most of the existing scan2CAD methods retrieve CADs models from ShapeNet [7], we use ShapeNet to supplement the shape collection.

5.1.2 Data preparation

To prepare the training data for the material decomposition network, we need photorealistic renderings of diverse 3D shapes in different scenes with ground truth SVBRDF components and illumination. We parameterize the environment maps with 12 SGs firstly. Then, we render all shapes in ABO [10] from 30 different random views and 3 different random lightings.

To get the training data for the geometry mapping network, we need point clouds with SDF values from the shape surfaces and surrounding space. We use shapes of the same category from ABO [10] and ShapeNet [7]. First, we normalize each ground truth mesh into a unit sphere. Then, for surface points, we render depth maps from 100 different views and record rendering parameters to avoid sampling invisible inner points. The surface points are recovered from depth maps. Normals of these points are sampled with the same method. For surrounding points, we sample points in a unit cube uniformly. We compute the distance to the nearest surface point as the SDF value. The sign of the SDF value is decided by checking the depth buffer. For each shape, we randomly sample 500K surface points with normals and 500K surrounding space points with SDF values.

To get the training data for the material completion network, we need point clouds with normals and SVBRDF values for the coarse stage, and SVBRDF maps in aligned UV space for the fine stage. The point clouds are sampled using the same virtual scan method as when preparing data for the geometry mapping network. To get the aligned material maps, we first train the geometry mapping network and then map the material component to inpaint the maps. Since we need shapes with materials in this stage, only shapes from the ABO dataset are used.

5.1.3 Training details

We train all three networks with four Nvidia RTX3090 GPUs under Ubuntu 20.04.2. The material decomposition network is trained for 20 epochs with loss weight parameter set $\{\alpha_1, \alpha_2\} = \{1, 1\}$. An AdamW optimizer with initial learning rate 1e-3 is used and the batch size is set to 24. The geometry mapping net is firstly trained for 50 epochs

to learn a correspondence with loss weight parameter set $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\} = \{1, 1e2, 5, 0, 0\}$. Note that only the shape mapper and template field are trained at this stage. Then, we fix the pre-trained network, and only train the UV mapping module for 10 epochs with loss weight parameter set $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\} = \{0, 0, 0, 10, 10\}$. An Adam optimizer with initial learning rate 1e-4 is used and batch size is set to 256. The material completion network is trained for 1000 epochs in the coarse stage with loss weight parameter set $\{\gamma_1, \gamma_2, \gamma_3, \gamma_4\} = \{1, 0, 0, 0\}$. An Adam optimizer with initial learning rate 2e-4 is used and batch size is set to 20. Then, we train the fine stage for 1000 epochs with loss weight parameter set $\{\gamma_1, \gamma_2, \gamma_3, \gamma_4\} = \{0, 10, 10, 1\}$, and the batch size is set to 4.

5.1.4 Inference details

During inference, given a single scan frame from the real world [11], we align the 3D object using ROCA [16]. Note that there are other methods [2, 23, 15] that can provide the same type of alignment. Then, given the rough object bounding box generated by ROCA, we use SAM [22] to get a finer object mask I_{mask}. Given the RGB image I_{rgb} and finer object mask I_{mask}, we estimate the SVBRDF and illumination using the material decomposition network. Then, the source point cloud p_s and the target point cloud p_t are sampled from the depth image and target shape, respectively. We employ an Adam optimizer with learning rate le-5 to optimize the source shape code. After mapping the two point clouds to the template space and obtaining the target point cloud with partial materials, the portion of the point cloud without a defined SVBRDF is modified with random noise. The material decomposition network takes the noised point cloud as input, and outputs a point cloud with complete materials. Note that to ensure the material consistency in the generated results, we fix newly completed materials, add slight noise to the known materials and re-denoise the part. The materials on the point cloud are mapped to UV space to generate coarse material maps with a resolution of 512. The image diffusion model refines coarse maps to generate the final material maps.

Now, we obtained output material maps in aligned UV space. To assign the maps to the target mesh correctly, we use the pre-trained geometry mapping network to recompute a new UV mapping for the target mesh.

5.2. Results

Figure 1 shows a sample of results obtained with our method, where we show the input image and aligned shape on the left, a rendering of the input shape with transferred material under the predicted illumination in the middle, and renderings in two different environment maps on the right. When inspecting these results, we see that our method can handle imperfect shape retrieval and alignment. For inaccurate retrieval (row 1), our method successfully transfers the materials from an armchair to a side chair. For inaccurate alignment (rows 4, 5, and 6), our method also provides reasonable results. Furthermore, all source models in the input image are partially occluded by themselves or other objects in the scene. For example, chairs in row 1, 3, and 6 are occluded by tables and the chair in row 5 is occluded by clothes laid on it. The use of the fine masker SAM [22] has successfully removed the occlusion and provides cleaner object masks which may have several fragmented cuts. Thanks to our 3D correspondences pipeline, these fragmented cuts have little effect in the results. Note that previous methods [29, 30, 19] assume a complete observation to establish 2D correspondences, and would fail in most of these difficult cases. Moreover, regardless of whether the object is near (row 2) or far (row 6) from the camera, our method can get reasonable results.

5.3. Comparison

In this section, we compare our method with baselines that solve the same problem with different strategies by qualitative and quantitative evaluations.

Baseline methods. We compare our method to 3 different baselines. The first baseline, denoted as *TF-mat*, is based on Texture Fields [29]. Since the original TF just learns an implicit texture space, we expand this method to the material space and train a material field for each category conditioned on the multi-dimensional SVBRDF components. During inference, the SVBRDF condition is predicted by our material decomposition network. Since this method takes 3D points as input, we reassemble the material maps using a UV-coordination map calculated by a shape-specific UV mapping. The second baseline, denoted as *PUD-mat*, is based on Point UV Diffusion [37]. We expand this method to transfer materials and train a material Point UV diffusion network conditioned on the SVBRDF. During inference, the network takes the estimated SVBRDF as input, and directly outputs material maps in shape-specific UV space. The third baseline, denoted as *AUV-mat*, is based on the single view reconstruction pipeline in AUV-Net [9]. AUV-Net learns a semantic aligned UV mapping for shapes in the same category. Based on the aligned UV space, the method reconstructs a textured shape from a single view observation. We only expand the texture generation branch to material space since our shape is given. More specifically, we train an auto-encoder to generate material maps from the input SVBRDF directly. During inference, this network takes the estimated material as input, and outputs material maps in the aligned UV space.

Metric. For evaluation, we measure the difference between the input image and the rendered target shape with the transferred materials in the predicted illumination. To

evaluate how the methods capture and maintain the material distribution for a given input image, we use the Fréchet Inception Distance (FID) [18] . This metric is widely used in image generation tasks.

Analysis. Figure 6 shows a sample of results generated by our method and the baselines. Our method obtains reasonable material transfer results even when an imperfect shape alignment is given, while the baselines suffer from inaccurate correspondence or fragile UV mapping. To be more specific, we find that *TF-Mat* hardly gets reasonable correspondences when partially occluded observations are given. Besides that, since this baseline represents the distribution of materials in a 3D volume space, it tends to generate over-smooth results. That is why shapes with pure color are generated. In addition, *AUV-Mat* uses a simple auto-encoder network to generate view-independent complete material maps from a single view observation. Although all the generated material maps are in an aligned space, we find that the method still does not have the ability to handle the misalignment between observations and material maps. Furthermore, we find that this baseline easily overfits the training data, and does not generalize well to new data. For *PUD-Mat*, we also find that the computed correspondence is of low quality. Since the method uses the pre-trained CLIP [31] to embed observed material components into a latent space, this process may disrupt structural information in the observation. The impact may be even greater when the structural information is already incomplete. Furthermore, this baseline generates material maps under an object-specific UV mapping. Thus, when the object is segmented into many small pieces in UV space, unnatural results will be easily generated, as seen on the rightest column of Figure 6. We randomly sampled 100 samples from the test set to calculate the FID, which are 199.90/205.10/201.01/205.90 for "ours/TF-Mat/AUV-Mat/PUD-Mat," with lower values preferred.

5.4. Ablation studies

In this section, we evaluate the effectiveness of each module of our method. Our pipeline completes materials on the point cloud first and then refines the material maps in an aligned UV space. To evaluate the effectiveness of this strategy, we experiment with two simplified versions of our method, called *"w/o point material completion"* and *"w/o aligned UV mapping"*. The *"w/o point material completion"* configuration indicates that after establishing the correspondence in 3D space, we map the estimated SVBRDF from the images to the target UV space, so that we obtain partial material maps directly. Then, we employ an image completion network to complete the partial material maps. The *"w/o aligned UV mapping"* configuration indicates that the material completion is processed in a shape-specific UV mapping.

Figure 6. Comparison of our method to three baselines. The figure shows the input images and aligned shape (left two columns), results of our method (middle column), and results of three baselines (right three columns). All results are rendered in the estimated illumination. Note that our method can establish a robust 3D correspondence between the input images and 3D shapes to generate more reasonable results.

Figure 7 shows a sample of results generated by our full pipeline and simplified versions. In this figure, we show the input images and aligned shapes on the left, and the transfer results by 3 different methods on the right. In each column of the results, we show the rendering of the shapes with materials on the top and the generated material maps on the bottom. Note that only half of the material maps are shown in this figure. When compared to *w/o point material completion*, we find that our full pipeline provides more consistent results. Specifically, when inspecting the "leg" (right) part of the generated maps, the result generated by the full pipeline is relatively pure black which is more similar to the input image, while *w/o point material completion* provides a more random color distribution. The method maintains a smaller black part which is transferred from the image, but fills the remaining part with brown and white colors. We believe that first completing the materials on the point clouds introduces a material distribution prior from the 3D space, which improves material consistency within different components. When compared to *w/o aligned UV mapping*, we find that our full pipeline provides more reasonable material distributions. Specifically, given the input image, we observe that the target object has a white fabric-like material with higher roughness on its upper half and a black wood-like material with lower roughness on its lower half. Such distribution can be easily found in the full pipeline results (the second column), where lighter orange indicates higher roughness and darker orange indicates lower roughness. However, we do not see such distributions in the results of *w/o aligned UV mapping* (the forth column). Thus, we believe that the aligned UV mapping introduces an implicit semantic prior in UV space,

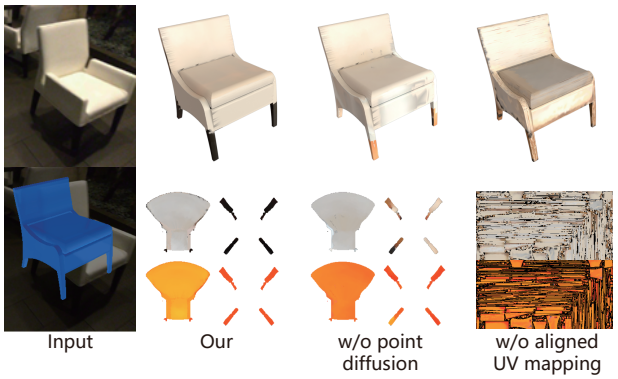

Figure 7. Ablation studies. The figure shows the input image and aligned shape (left column) and results of our full pipeline and simplified versions (right three columns). For each column, we show the rendering of the target shape on top, and the SVBRDF maps on the bottom. Note that our point cloud material completion and aligned UV mapping help us to generate more consistent and natural results.

which can help the networks to learn a reasonable material distribution. The FIDs on 100 random sampled test cases for "ours / *w/o point material completion* / *w/o aligned UV mapping*" are 199.90/201.15/225.55.

6. Discussion and future work

We presented a method for material transfer from single scanned frames to aligned shapes based on estimated dense correspondences. Our method first estimates object materials from images, then transfers them to the target shape based on predicted correspondences, and finally completes partial materials in UV space. We showed with qualitative and quantitative evaluations that, compared to other baseline methods, our method is more robust in handling imperfect alignment and partial observations. As a consequence, given scans and aligned shapes, our method can automatically transfer materials to shapes and provide realistic scenes which have both high-quality geometry and high-fidelity appearance.

Limitations. Our method has certain limitations. Figure 8 shows example results that represent the main failures of our method. Given input images with complex material patterns, our method cannot obtain good results. Although the material decomposition step can recognize these patterns, the material completion network is unable to correctly recognize and complete these patterns during diffusion. We also find that since our material completion network requires training from scratch with 3D models that have high-quality materials, our method may encounter overfitting for other categories where such data is insufficient, resulting in an inability to generate reasonable transfer results.

Future work. One direction for future work is to address the limitations summarized in Figure 8. Training the material completion network with more shapes with diverse appearance [12] may improve the completion ability of the method. We can also use more powerful material decomposition, transfer, and completion methods to achieve better results in our whole pipeline. In addition, we can also employ pre-trained image diffusion models [32, 35] to enhance the recognition and completion capability of material patterns in UV space, which can also reduce the demand for high-quality 3D training data, thereby improving the generalization of our method across different categories.

Figure 8. Representative failure cases of our method. Our method fails to transfer materials with complex patterns to the target shapes.

Another interesting research direction is to transfer materials from multiple viewpoints of the object. There is no doubt that multi-view observations provide more geometric and appearance information about the target object, thus reducing the difficulty of material completion. However, due to different lighting conditions caused by different viewpoints, ensuring consistency in material decomposition may still be challenging.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported in parts by NSFC (62322207) and National Key R&D Program of China (2024YFB2809102).

References

[1] Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Nießner, M.: Scan2cad: Learning cad model alignment in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 2614–2623 (2019) 1, 2

[2] Avetisyan, A., Khanova, T., Choy, C., Dash, D., Dai, A., Nießner, M.: Scenecad: Predicting object alignments and layouts in rgb-d scans. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 596–612. Springer (2020) 1, 2, 8

[3] Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.: Nerd: Neural reflectance decompo-

sition from image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12684–12694 (2021) 4

[4] Boss, M., Jampani, V., Kim, K., Lensch, H., Kautz, J.: Two-shot spatially-varying brdf and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3982–3991 (2020) 4

[5] Bruno, V., Lévy, B.: What you seam is what you get (2009) 5

[6] Cao, T., Kreis, K., Fidler, S., Sharp, N., Yin, K.: Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4169–4181 (2023) 3

[7] Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 2, 7

[8] Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. arXiv preprint arXiv:2303.11396 (2023) 3

[9] Chen, Z., Yin, K., Fidler, S.: Auv-net: Learning aligned uv maps for texture transfer and synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1465–1474 (2022) 5, 8

[10] Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21126–21136 (2022) 4, 7

[11] Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017) 8

[12] Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023) 2, 10

[13] Deng, Y., Yang, J., Tong, X.: Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10286–10296 (2021) 5

[14] Di, Y., Zhang, C., Zhang, R., Manhardt, F., Su, Y., Rambach, J., Stricker, D., Ji, X., Tombari, F.: U-red: Unsupervised 3d shape retrieval and deformation for partial point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8884–8895 (2023) 2

[15] Gao, D., Rozenberszki, D., Leutenegger, S., Dai, A.: Diffcad: Weakly-supervised probabilistic cad model retrieval and alignment from an rgb image. arXiv preprint arXiv:2311.18610 (2023) 1, 2, 3, 8

[16] Gümeli, C., Dai, A., Nießner, M.: Roca: Robust cad model retrieval and alignment from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4022–4031 (2022) 1, 2, 3, 8

[17] Haven, P.: Poly haven (2024), https://polyhaven.com/hdris 7

[18] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 9

[19] Hu, R., Su, X., Chen, X., Van Kaick, O., Huang, H.: Photo-to-shape material transfer for diverse structures. ACM Transactions on Graphics (TOG) **41**(4), 1–14 (2022) 2, 3, 8

[20] Huang, H., Xie, K., Ma, L., Lischinski, D., Gong, M., Tong, X., Cohen-Or, D.: Appearance modeling via proxy-to-image alignment. ACM Transactions on Graphics (TOG) **37**(1), 1–15 (2018) 3

[21] Izadinia, H., Shan, Q., Seitz, S.M.: Im2cad. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5134–5143 (2017) 2

[22] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 3, 8

[23] Kuo, W., Angelova, A., Lin, T.Y., Dai, A.: Mask2cad: 3d shape prediction by learning to segment and retrieve. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 260–277. Springer (2020) 3, 8

[24] Le, C., Hetang, C., Cao, A., He, Y.: Euclidreamer: Fast and high-quality texturing for 3d models with stable diffusion depth. arXiv preprint arXiv:2311.15573 (2023) 3

[25] Li, C., Guo, J., Hu, R., Liu, L.: Online scene cad recomposition via autonomous scanning. ACM Transactions on Graphics (TOG) **42**(6), 1–16 (2023) 1, 2

[26] Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. Advances in Neural Information Processing Systems **32** (2019) 6

[27] Munkberg, J., Hasselgren, J., Shen, T., Gao, J., Chen, W., Evans, A., Müller, T., Fidler, S.: Extracting triangular 3d models, materials, and lighting from images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8280–8290 (2022) 4

[28] Nguyen, C.H., Ritschel, T., Myszkowski, K., Eisemann, E., Seidel, H.P.: 3d material style transfer. In: Computer Graphics Forum. vol. 31, pp. 431–438. Wiley Online Library (2012) 3

[29] Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., Geiger, A.: Texture fields: Learning texture representations in function space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4531–4540 (2019) 2, 3, 8

[30] Park, K., Rematas, K., Farhadi, A., Seitz, S.M.: Photoshape: Photorealistic materials for large-scale shape collections. arXiv preprint arXiv:1809.09761 (2018) 2, 3, 8

[31] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 9

[32] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 10

[33] Rematas, K., Nguyen, C.H., Ritschel, T., Fritz, M., Tuytelaars, T.: Novel views of objects from a single image. IEEE transactions on pattern analysis and machine intelligence **39**(8), 1576–1590 (2016) 3

[34] Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. arXiv preprint arXiv:2302.01721 (2023) 3

[35] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 10

[36] Wang, T.Y., Su, H., Huang, Q., Huang, J., Guibas, L.J., Mitra, N.J.: Unsupervised texture transfer from images to model collections. ACM Trans. Graph. **35**(6), 177–1 (2016) 3

[37] Yu, X., Dai, P., Li, W., Ma, L., Liu, Z., Qi, X.: Texture generation on 3d meshes with point-uv diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4206–4216 (2023) 2, 3, 6, 7, 8