SAMa: Material-aware 3D selection and segmentation

Michael Fischer[1,2*], Iliyan Georgiev[1], Thibault Groueix[1], Vladimir G. Kim[1],
Tobias Ritschel[2], Valentin Deschaintre[1]

[1]Adobe Research [2]University College London

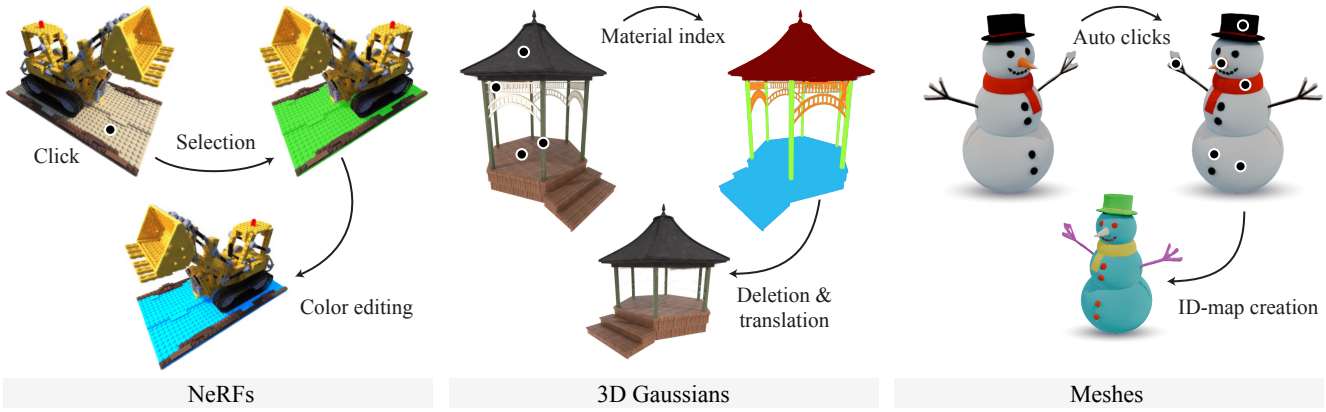

Figure 1. 3D material selection on three different representations using our SAMa method. Our approach enables several applications; from left to right: color editing on NeRFs, decomposition and editing on Gaussians, automatic material-ID-map creation on meshes.

Abstract

Decomposing 3D assets into material parts is a common task for artists and creators, yet remains a highly manual process. In this work, we introduce Select Any Material (SAMa), a material selection approach for various 3D representations. Building on the recently introduced SAM2 video selection model, we extend its capabilities to the material domain. We leverage the model's cross-view consistency to create a 3D-consistent intermediate material-similarity representation in the form of a point cloud from a sparse set of views. Nearest-neighbor lookups in this similarity cloud allow us to efficiently reconstruct accurate continuous selection masks over objects' surfaces that can be inspected from any view. Our method is multiview-consistent by design, alleviating the need for contrastive learning or feature-field pre-processing, and performs optimization-free selection in seconds. Our approach works on arbitrary 3D representations and outperforms several strong baselines in terms of selection accuracy and multiview consistency. It enables several compelling applications, such as replacing the diffuse-textured materials on a text-to-3D output with PBR materials, or selecting and editing materials on NeRFs and 3D-Gaussians.

*Corresponding author. Work done during an internship at Adobe Research. Contact: m.fischer@cs.ucl.ac.uk.

1. Introduction

Understanding the materials around us is an extremely common task for humans, but remains challenging for machine vision approaches. In this paper we focus on material selection on 3D objects.

Existing work on material understanding has mostly focused on the 2D image domain, addressing tasks like segmentation [4, 40, 57, 59], reconstruction [14, 15, 25, 56], generation [52, 61, 62] or, more recently, material selection [55]. In material selection, the task is to select all image pixels that share the same material, given a user prompt (typically a click on the material of interest), regardless of variations in shading or object boundaries. This is an important distinction from semantic or object selection, which aim to identify individual objects or semantic object groups. For instance, when clicking on a wood material, we neither want just the current object that the material is applied on (*e.g.*, a single chair), nor the semantic entity or group that the material belongs to (*e.g.*, all chairs). Instead, we want to return *all parts* (of the asset) made of that specific type of wood. We differentiate selection and segmentation, where the former aims to identify one material based on a user click prompt, while the latter targets a complete decomposition into different material regions. We follow the material definition of established works [13, 55] and consider two materials similar if they share the same texture and re-

flectance properties.

Material selection becomes especially relevant in the light of recent generative 3D asset creation and image/text-to-3D workflows. Current methods either provide non-parametric implicit representations (*e.g.*, Neural Radiance Fields (NeRFs)) or unstructured output (as in triangle soups and baked textures produced by image/text-to-mesh methods [20, 24, 37, 44]), both of which are challenging to use for artists and downstream tasks. Material selection, in this context, has a wide range of downstream applications, *e.g.*, enhancing the X-to-3D workflow with material masks, improving the editability of 3D reconstructions (*e.g.*, through material replacement), or extracting areas of similar materials as a prior for observation sharing in inverse rendering.

However, most models targeting material-related tasks, including selection, do not trivially extend to the 3D domain, as they are trained on 2D images and therefore have no incentive for producing multiview-consistent predictions [17]. Moreover, the 3D domain contains inherent challenges and ambiguities like self- or dis-occlusions and view-dependent effects, and requires accurate propagation of the model predictions into novel, unseen views. Recent research has therefore developed algorithms to address the problems from multiview-inconsistent predictions that arise when lifting 2D (object) selection to 3D, predominantly via pre-processing noise-consolidation steps such as feature-field distillation [31] or contrastive (similarity) learning [29], both of which are time-consuming.

In this work, we close the gap between material selection in 2D and 3D and propose an efficient and accurate material selection and segmentation method for 3D objects. One of our core insights is that we can draw parallels between video selection and accurate 3D selection, since in both video and 3D, the selected elements have to be consistent across frames (or views), regardless of object and camera movement or differences in shading and occlusions. We thus propose to re-purpose SAM2's recent progress in object selection across video [48] for materials. We achieve this by fine-tuning parts of the model for (video) material selection on a custom-made video dataset with dense per-pixel, per-frame annotations. We show that the use of videos for fine-tuning is key to achieving high quality in 3D.

Our approach is inherently multiview-consistent thanks to its video training. Therefore, once our SAMa is trained, it enables selection in less than two seconds from the initial click on an arbitrary 3D object. Selection visualization from different views can then be performed in under 10 milliseconds, enabling interactive visualization and editing of our selection results.

Importantly, our approach supports selection on any 3D representation that can be rendered to an image and queried for depth. We show selection results on meshes, radiance fields and 3D Gaussians. We evaluate our method qualitatively and quantitatively, both in terms of selection quality and 3D consistency, and show that it improves significantly over existing work and several strong baselines. Finally, we demonstrate multiple applications such as object segmentation into material IDs and NeRF/Gaussian editing.

In summary, we make the following contributions:

- Adaptation of a video-object-selection model to material selection on 3D shapes, by training on a novel rendered video dataset.
- Fast and efficient 3D projection, enabled by cross-frame consistency.
- Multi-modality support, segmentation and editing.

Throughout this paper, we will show the user-provided input clicks with respect to which we select materials as black circles. We will show the material similarity to these clicks in false colors, with blue and red indicating low and high, respectively.

2. Related work

Most related to our work are approaches for material selection on images and approaches that lift a 2D signal defined on renderings into a 3D representation.

Material segmentation datasets. Several semantic material datasets with material segmentation annotations exist. The Multi-Illumination dataset [42], the Light-Field Material [63] and Flickr Material [54] datasets respectively contain 1k, 1.2k, and 1k images, segmented with 35, 12 and 10 materials respectively. Of greater size, the OpenSurface [3], Material in Context [4], Dense Material Segmentation [59] datasets respectively contain 19k, 437k and 45k images annotated with respectively 37, 23 and 52 types of materials. The Local Material Database [53] further annotates 16 kinds of materials on images sources for the previously mentioned datasets. These datasets only contain coarse material categories, *e.g.*, two types of metal would have the same "metal" label, creating false positives where pixels are marked as sharing the same superclass material but do not share the same appearance.

Materialistic [55] provides a synthetic dataset of 50k HDR images, path-traced from 100 indoor scenes from the Archinteriors collection [1] and 3k materials. Complementing this data, Eppel et al. [18] extract textures from the Open Images v7 dataset [32] and apply them to random parts of 3D objects from the ShapeNet repository [10]. The resulting dataset has the advantage of having fine-grained annotations for each material, such as dirt and paint splashes.

Importantly, these datasets [18, 55] contain only static renderings, making it challenging to learn multiview consistency. In contrast, our video dataset has dense, fine-grained per-pixel material annotations, enabling the fine-tuning of video selection models.

Material selection in 2D. Most prior works in material segmentation rely on hand-crafted features [5, 23, 39, 47, 50] or focus on images of flat surfaces [11, 25, 33, 34, 43]. Recently, Sharma et al. [55] proposed Materialistic, a model based on DINO-ViT [7] features, trained to predict the material similarity between a query pixel and all other pixels in a natural image. We find that Materialistic struggles with accurate material selection on 3D objects for two reasons: (1) it is trained for full-scene photographs, leading to limited selection precision on objects, and (2) its selections are not sufficiently consistent to be lifted to varying 3D views.

Closely related, the Segment Anything Model (SAM) [30] uses a ViT trained to predict similarity between pixels. As its training data is object-selection specific and not material-aware, SAM requires many separate clicks to perform even moderately well on materials. The more recent SAM2 [48] also targets object selection, but introduces support for temporally coherent predictions across video frames. We find that neither SAM nor SAM2 perform well on material selection, except in the special case of an object made of a single material. However, once fine-tuned for materials, SAM2's cross-frame selection consistency enables our fast selection lifting to 3D.

Lifting 2D features to 3D. Due to the scarcity of annotated 3D data and the increased computational complexity compared to 2D, many approaches have attempted to lift available 2D predictions from multiple views to a shared 3D representation. We here discuss approaches that tackle selection and segmentation.

The core issue is that the underlying 2D vision models like SAM or DINO are not multiview-consistent. That is, they provide differing predictions for the same 3D point viewed from different positions, making aggregation in 3D challenging. Neural Feature Fusion Fields [58] and Feature Field Distillation [31] propose to equip NeRFs with an auxiliary feature space, rendered volumetrically to match DINO [7] or CLIP [46] features. Even though the 2D feature maps are not multiview consistent, the shared 3D representation acts as a regularizer and consolidates the quality of the features in rendered novel views. This approach has been extended to 3D Gaussian splatting [27] and other image models such as SAM [21, 28, 36, 38, 45, 66].

Other approaches use contrastive learning to lift segmentation to 3D by pushing closer rendered features of pixels belonging to the same segment, and vice versa [6, 8, 9, 12, 19, 22, 29, 35, 49, 65]. Our approach differs from this line of work in several ways. First, we lift 2D *material* similarity (rather than object similarity) to 3D. Second, as opposed to previous work, our similarity maps are already multiview consistent thanks to our fine-tuning of a video selection model [48]. Using this property, we propose a 3D representation-agnostic, lightweight 2D-to-3D lifting approach that does not require any pre-processing. Contrary to prior work, this allows us to process arbitrary 3D representations (*e.g.*, NeRFs, Gaussians, meshes) and reduces the initial click-to-selection time from 2 hours [58] or 20 minutes [29] for existing methods to around two seconds.

3. Method

Our approach targets material selection on 3D representations. Existing methods focus mostly on selection in 2D images [55], and their extension to 3D is not trivial as the underlying vision models are not consistent across views. They also do not cope well with selected material regions going out of frame or changes in the frames' background.

Instead of enforcing 3D consistency through per-asset optimization and feature consolidation [22, 29, 31, 65], we take inspiration from memory priors in recent video models [48, 60] which show good cross-frame consistency. Since renders of a 3D object from a smooth camera trajectory are not markedly different from a video, we propose to adapt SAM2 [48] to material selection by fine-tuning it, including its memory bank components, on a material-specific video dataset that we design. Once fine-tuned, given an image of a rendered object and a (clicked) pixel, the model outputs a floating-point map that encodes the similarity between the clicked pixel's material and all other pixels. Thresholding such a similarity map yields a binary selection mask.

However, fine-tuning alone does not yet lead to *interactive* selection in 3D. While it enables material selection from novel views with good consistency, it is not yet efficient, as it requires querying the model for each such view. Furthermore, cross-frame consistency can still exhibit artifacts in challenging cases and long frame sequences. To alleviate inconsistencies, we consolidate the 2D similarity maps of a sparse set of key-frames into a 3D similarity point cloud. Using this point cloud and nearest-neighbor queries, we can recover (and display) selections from any viewpoint on the 3D shape in a few milliseconds.

In summary, we train a multiview-consistent 2D selection model and project selection similarities from multiple viewpoints to 3D using a simple, yet efficient, point cloud that can be queried from new cameras. Figure 2 provides an overview of our method.

3.1. Fine-tuning for 2D material selection

We re-purpose the recently introduced SAM2 model [48] to material selection in the video domain. SAM2 uses an efficient Vision Transformer (ViT) image encoder [51] to produce a per-frame image embedding, and infers a per-pixel object similarity value for each frame. The key novel component in SAM2 is the memory attention module, which conditions the current frame embedding on the embeddings of past and future frames in the sequence, allowing the

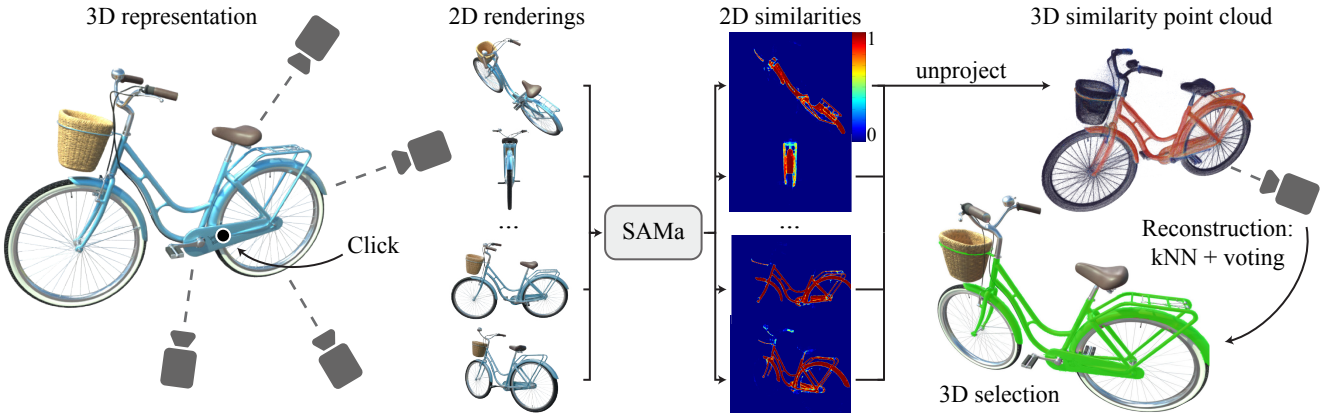

Figure 2. Overview over our method. Starting from a 3D asset and a user click, we sample cameras and create a set of renderings covering the object, which we subsequently process with our similarity network SAMa to compute dense per-pixel similarity values. We then back-project these values to 3D and store them in a point cloud than can be efficiently queried and interpolated for novel views.

model to reason both spatially and temporally. These embeddings are then combined with the encoded conditioning query (*e.g.*, a click on a pixel) in the mask decoder, producing per-frame similarity masks. As we will show in our experiments, correctly fine-tuning this memory module is key to achieving multiview-consistent selection results.

While our initial experiments confirmed SAM2's good cross-view consistency, they also revealed a tendency to select object (sub-)parts instead of materials. We therefore fine-tune the model for the task of material selection. Specifically, we freeze the encoder throughout our fine-tuning to preserve the rich priors learned from millions of images and tune the remainder of the architecture (see Fig. 3). We find that training solely on *image* data for material selection (*e.g.*, the Materialistic dataset [55]), performs reasonably well on clicked frames, but leads to a significant drop in cross-frame selection consistency, as shown in Fig. 4. We attribute this to the fact that on unseen frames, the model must infer the material selection from memory, but fine-tuning on images does not adapt the memory module since memory is never queried for a single image.

However, for 3D selection, cross-frame consistency is particularly important. We therefore design an object-centric video dataset with material-segmentation annotation by randomly sampling objects, materials and environment maps, combining them into simple scenes containing one to a few objects. We allow the same materials to appear multiple times and in different locations within a scene, to clearly disambiguate material and object selection. We render 30 frames for each video using a random choice of four possible camera trajectories: zoom-in, zoom-out, spherical turntable and fly-over. Finally, to reduce the domain gap between natural and single-object images, we alpha-composit the environment map into the background (for additional dataset and training details, see Appendix A.1). We find that 500 videos are sufficient to adapt the model to the material-selection domain.

Our new video material dataset with dense per-frame material annotations enables us to jointly fine-tune SAM2's memory attention module and the mask decoder. This way, we maintain multiview-consistency and obtain significantly better inferred selections (right column in Fig. 4). We will release our dataset upon publication.

3.2. Lifting 2D similarity to 3D

Given a click on one image, our goal is to obtain a selection in 3D of all object parts that share the same material. A 3D selection is not only view-consistent by design, it also enables downstream applications (*e.g.*, editing) that naturally operate in 3D on the object (surface). An entirely image-based pipeline would require running our 2D selection model for every new viewpoint, completely relying on the model's cross-frame consistency. Such a workflow would not be interactive (2–5 sec per frame for simple selection visualization), would suffer from flickering due to residual multiview inconsistencies in long frame sequences, and would not be compatible with many downstream applications (*e.g.*, mesh material replacement). We therefore instead consolidate similarity maps from multiple viewpoints into a lightweight 3D similarity point cloud. From this cloud we can easily reconstruct (and display) a continuous 3D selection at interactive rates.

The initial camera, in which the click was performed, will serve to condition the memory module of our SAMa model, as it ensures that the material is not occluded. We then render RGB and depth images from multiple viewpoints; for each RGB image we use our model to estimate the selection, based on the user-provided click, given the other images as video context through the memory bank. This process yields a per-viewpoint map representing the similarity to the user-clicked material. We back-project these maps to 3D using the previously rendered depth im-

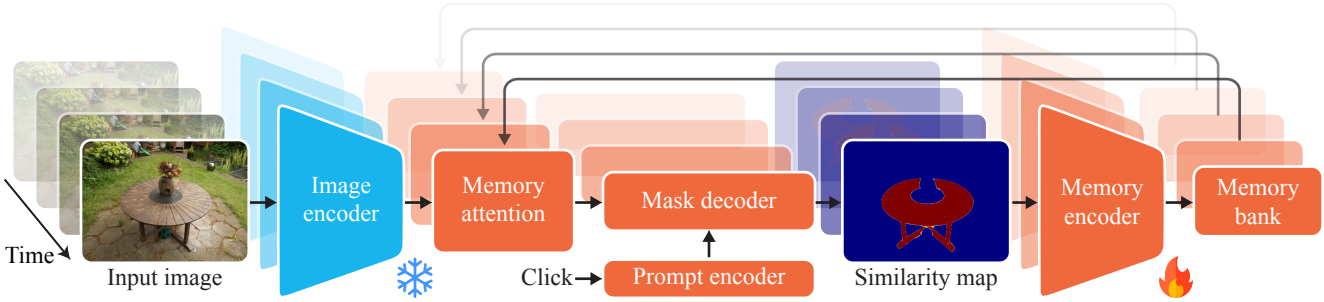

Figure 3. Schematic overview of our fine-tuned model. The image encoder (in blue) is frozen, all other blocks (in red) are fine-tuned. Given an input image and a clicked pixel, the model outputs a material similarity map. Figure adapted from Ravi et al. [48].

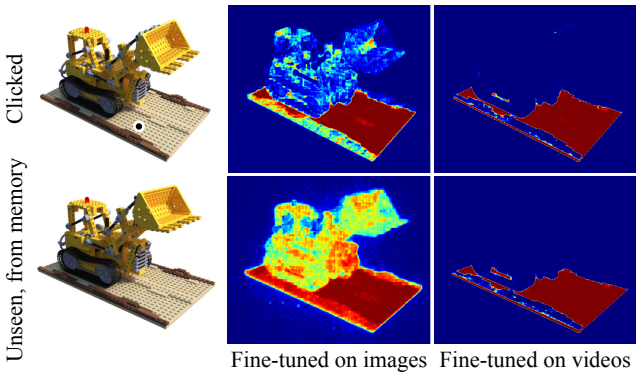

Figure 4. Effects of fine-tuning on images vs. videos. Top row shows the clicked frame. Bottom row shows an unclicked frame for which the similarity map is inferred from the model's memory.

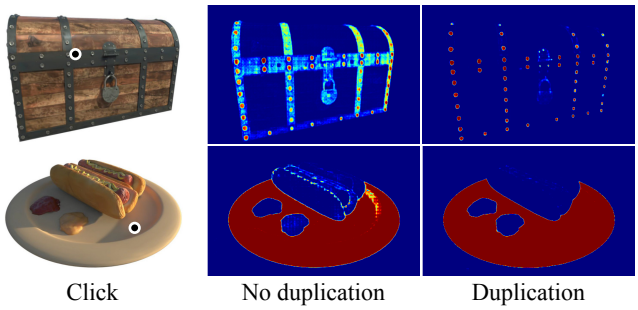

Figure 5. Effects of duplicating the clicked frame in the sequence. Similarity after frame duplication is significantly cleaner, as the model is forced to use the memory module.

ages, to obtain a 3D similarity point cloud.

Our approach works on any 3D representation that can be rendered from a given viewpoint and queried for depth. For NeRFs and 3D Gaussians we use the training views, while for meshes we use spherical Fibonacci sampling for the camera positions (looking at the object's center). To ensure good performance of our fine-tuned video model, we arrange those views along a smooth trajectory via greedy iterative camera sorting laid out in Appendix Algorithm 1.

From our point cloud, we can reconstruct a continuous 3D similarity field via k-nearest neighbour (kNN) lookups. For novel views, we use FAISS [16] for performant large-scale, GPU-accelerated approximate nearest-neighbor queries [26] at the camera rays' 3D hitpoints. We cache and reuse the acceleration structure built by the library; we need to rebuild it only when the selected material changes. With this approach, a new user selection from a novel viewpoint takes around 2 sec (including 0.5 sec for the structure construction), while querying the point cloud from a new viewpoint takes 10–20 ms at 512–1024p image resolution. Appendix A.2 provides additional details.

3.3. Refinement

Frame duplication. We observe that the frame where the user clicks exhibits significantly higher selection noise. This is due to the memory module not being queried for this selection, meaning that the model does not have access to the information in the other frames. To improve selection quality on this frame, we simply duplicate it. The first copy is used for conditioning the selection without memory module, and the second copy is included with the other frames in the sequence, using the memory module. We show the effect of click-frame duplication in Fig. 5 and on the original SAM2 model in the supplemental material.

kNN-based voting. Thresholding our kNN-reconstructed 3D similarity field yields a binary selection field. To ensure a clean selection, we use a binary voting scheme: we consider a 3D point as selected if more than half of its nearest neighbors pass the selection threshold. The threshold can be set by the user to adjust the selection, as in prior work [55]. We show the effect of this aggregation strategy in Fig. 6.

4. Evaluation

4.1. Datasets

We evaluate our method on three datasets: (1) the eight scenes in the NeRF dataset [41], (2) five real-world scenes from the MIPNeRF-360 dataset [2], and (3) twelve objects from our own object-centric dataset. For synthetic objects we render material-ID maps which provide ground-truth an-

Figure 6. kNN 3D voting significantly reduces noise and improves selection quality, as seen from the insets.

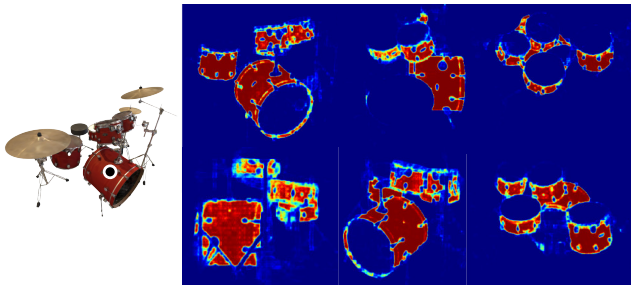

Figure 7. Conditioned by an initial click on the bass drum, our selection model achieves remarkable multiview-consistency in the presence of severe occlusions and perspective changes.

notations per view. For real-world assets we hand-annotate five images per scene (examples in the supplemental).

4.2. Baselines

We compare our method against three baselines. The first is the original Materialistic method [55] for which we query for different views by re-projecting the initial click into the new view. In its default version, this can only be done for views where the original click is not occluded. We can still query this baseline from new views thanks to our 3D-point-cloud lookup, but the results will be patchy as it cannot process all of the input views. However, Sharma et al. [55] show that selection can work across two frames by computing the cross-attention between the initial click's Q values and the KV values of the other views. We extend this scheme to n frames to process all unseen viewpoints, and refer to it below as "Materialistic MV" (multi-view).

Additionally, we compare against the multiview-consistent, but not material-aware, SAM2 model. Finally, "Ours" denotes our full method, including our fine-tuned network. For all methods, we lift the results to 3D using our point-cloud representation.

4.3. Results

We evaluate each method in 3D (after the point cloud lookup), along three axes: selection quality, robustness and multiview consistency. We report metrics on binary selection masks obtained by thresholding similarities against 0.5. Metrics are normalized to $[0, 1]$; see Appendix B for details.

Dataset	NeRF [41]		MIPNeRF-360 [2]		Our Dataset	
	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑
Ours	$\mathbf{0.48 \pm .2}$	$\mathbf{0.58 \pm .3}$	$\mathbf{0.60 \pm .3}$	$\mathbf{0.72 \pm .3}$	$\mathbf{0.69 \pm .2}$	$\mathbf{0.78 \pm .2}$
SAM2	$0.33 \pm .2$	$0.43 \pm .3$	$0.51 \pm .3$	$0.65 \pm .3$	$0.36 \pm .2$	$0.47 \pm .2$
Mat.	$0.24 \pm .1$	$0.36 \pm .2$	$0.31 \pm .3$	$0.44 \pm .3$	$0.47 \pm .2$	$0.59 \pm .2$
Mat. MV	$0.27 \pm .2$	$0.32 \pm .2$	$0.32 \pm .3$	$0.47 \pm .3$	$0.51 \pm .2$	$0.62 \pm .2$

Table 1. Selection accuracy across datasets (columns) for several methods (rows), with 95% confidence intervals. For the per-scene measurements and precision and recall, we refer to Appendix B. Mat. is short for Materialistic [55].

	Dataset	NeRF [41]	MIPNeRF-360 [2]	Our Dataset
Consistency	Ours	$\mathbf{2.2 \pm 0.2}$	1.4 ± 0.2	$\mathbf{1.7 \pm 0.1}$
	SAM2	$\mathbf{2.2 \pm 0.2}$	$\mathbf{1.2 \pm 0.1}$	1.9 ± 0.2
	Materialistic	5.5 ± 0.3	4.4 ± 0.3	5.9 ± 0.4
	Material. MV	3.9 ± 0.2	4.1 ± 0.4	4.9 ± 0.3
Robustness	Ours	$\mathbf{1.1 \pm 0.8}$	$\mathbf{1.2 \pm 1.3}$	$\mathbf{0.3 \pm 0.2}$
	SAM2	1.3 ± 0.9	2.9 ± 3.8	0.7 ± 0.6
	Materialistic	3.2 ± 0.6	7.1 ± 4.5	1.8 ± 1.0
	Materialistic MV	3.9 ± 1.4	3.5 ± 1.5	2.1 ± 1.0

Table 2. Multiview consistency (top) and robustness (bottom) of our selection across unseen test views. We report Hamming distance ($\times 100$) with 95% confidence intervals. Lower is better.

Selection accuracy. We perform a click in one view, then for each of 50 random novel views we compare the obtained selection mask against a rendered ground-truth mask. We compute mean intersection over union (mIoU), a classical selection metric measuring the overlap between the masks. We also report F1 score which is the harmonic mean of precision and recall, and is more robust than either alone. We average each metric over the views and over five random clicks on each material. We report the averages and 95% confidence intervals across the datasets in Tab. 1, higher is better. Appendix B provides per-scene breakdown.

Multiview consistency. We demonstrate our method's multiview consistency in Fig. 7, and measure it numerically in Tab. 2 top as follows. We perform a click in one view, then sample 50 novel views for which the clicked 3D point is unoccluded. For each view, we average the difference between the binary selection value at the point and the reference selection value of 1 in the clicked view. Perfect multiview consistency means zero average difference, *i.e.* all values are 1. Note that this metric does not quantify mask correctness; if a returned mask is wrong, but consistently so, the reported score will still be high. Both our method and SAM2 show similar consistency, while both Materialistic baselines, which do not benefit from the cross-frame memory mechanism, achieve lower consistency scores.

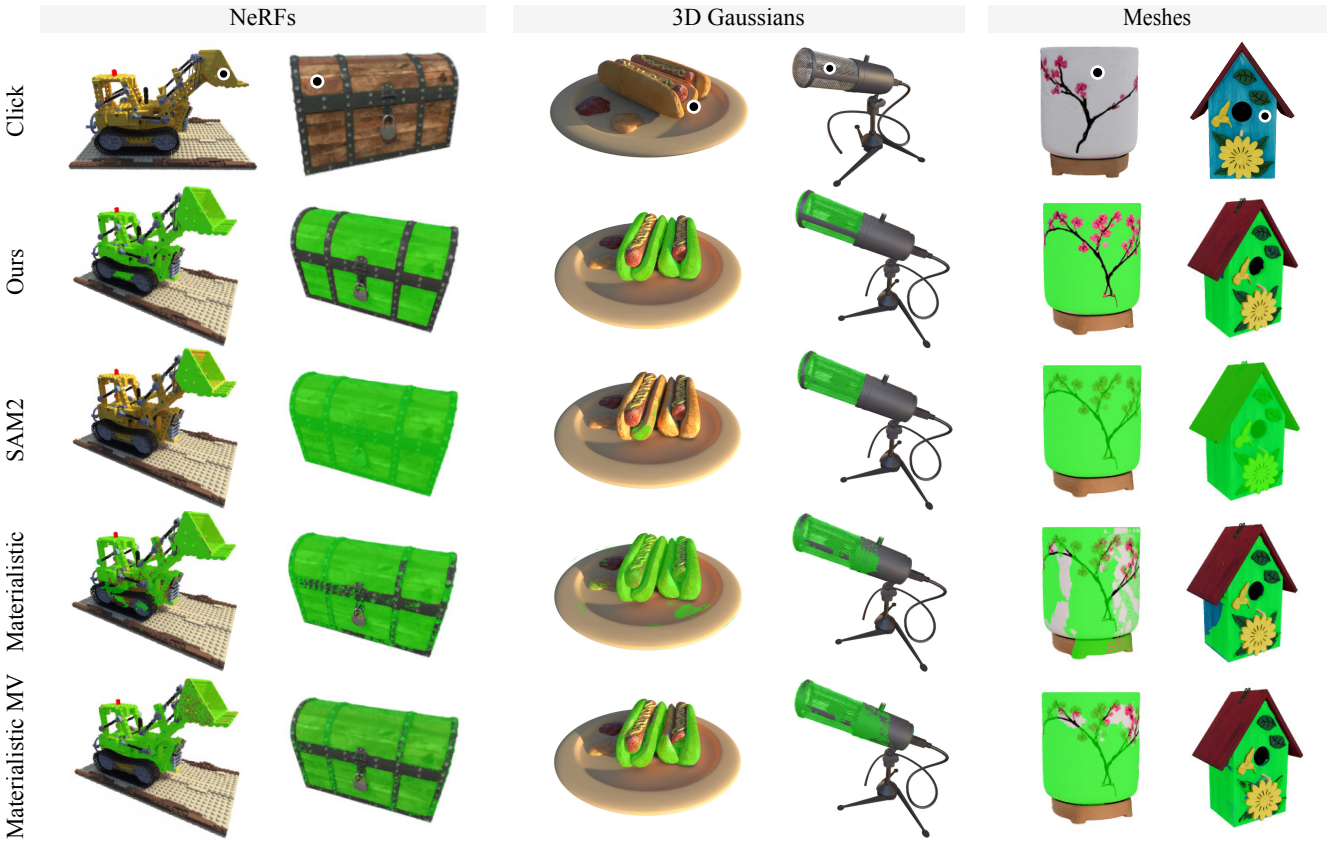

Figure 8. Selection results on NeRFs, 3D Gaussians and meshes across objects (columns) and methods (rows). Leftmost column shows the clicked view and the click. Selection results, from a novel view, are overlaid as green masks.

Robustness to click location. On a random view, we perform 5 random clicks within a single material. We then average the pairwise Hamming distances between the 5 selection masks. We report results in Tab. 2 bottom, lower is better. Like the multiview metric above, this metric does not quantify mask correctness. We see that architectures that benefit from multi-frame context show better robustness.

Qualitative evaluation. We show selection results on all our evaluated modalities (NeRFs, 3D Gaussians, meshes) in Fig. 8. We see that Materialistic and Materialistic MV do not work well on high-frequency boundaries of objects and that SAM2 cannot be trivially used for material selection, with parts of or entire objects selected at once. In contrast, our method creates sharp boundaries, including around thin elements, selecting parts of the object with materials similar to that clicked in the first view, and is robust to lighting variations (see Fig. 9).

We further evaluate our results in 2D without the point cloud lookup, which improves average mIoU by around 5%. However, our 3D aggregation in a point cloud lookup provides a significant advantage in efficiency, reducing per-frame inference processing times from around 5 sec in 2D to around 10 ms in 3D (500× faster), making it a more practical choice overall. This difference in quality is mainly explained by the depth estimation quality on volumetric representations, which is not always perfect.

5. Applications

5.1. Segmentation

While our method targets selection, it can be used to automatically segment an object into material subparts. Inspired by the image-level sampling in SAM [30], we propose an equivalent approach for materials on 3D objects. It involves two steps: (1) automatically choosing "selection clicks" and (2) merging similar resulting selections.

Densely sampling an entire object from multiple views is impractical (500-click sampling of the Lego asset in Fig. 4 takes ∼20 min). Instead, we select clicks based on the modes of a CIE LAB (D65) 3D histogram. To account for color differences due to shading and illumination, we discretize the L channel more coarsely than the AB channels (4 vs 16 bins). This strategy places samples in locations of different color, which often coincide with material regions. We sample 25 points in total, with stratified sampling in each color mode area, proportionally to the color area.

Given the sampled clicks, we compute the material sim-

Figure 9. Our method is robust w.r.t. shading variations on the surface, shown here for reflections, specularity and shadows.

ilarities as outlined in Sec. 3 for a random set of views, to obtain a binary selection masks per click. We compare these masks to one another and store their pairwise mIoU in a (symmetric) matrix, where each entry represents how similar the selection for two different clicks is. A high value implies that the clicks which created this entry led to similar selections, *i.e.*, were on the same material, so we can safely keep only one of them. We repeat this process until all matrix entries are below the empirically determined threshold 0.75, leaving only clicks on truly different materials. We show results of this segmentation approach in Fig. 10.

5.2. Editing

Using our material selection results, we can easily edit the selected regions. We show various edits and applications for NeRFs, 3D Gaussians and meshes in Fig. 1.

NeRFs. For NeRFs, we demonstrate color editing. We ray-march the NeRF as usual, but for each 3D point we query whether it has been selected. If yes, we adjust the color returned by the NeRF through a color shift in LAB space, to preserve relative shading and lighting information. We show an example in Fig. 1 and in the appendix.

Material-aware 3D Gaussians. For 3D Gaussians, we use our material segmentation step (described in Sec. 5.1). We then render the respective material masks for each training view and convert them to ID masks, so each pixel in the training images is associated with a material index. We then re-train the Gaussians with an extra channel for materials which is treated like the RGB channels for rasterization. This creates a clear separation between Gaussians at material boundaries, simplifying downstream edits, and provides a per-Gaussian material handle. We can now select all Gaussians that encode, *e.g.*, material number two, and edit their properties such as color, position or density. In Fig. 1 we move the gazebo's wooden base upwards and set the white painted regions' density to zero.

Meshes. For selection on meshes, we exploit their UV parametrization by writing the selected material similarities

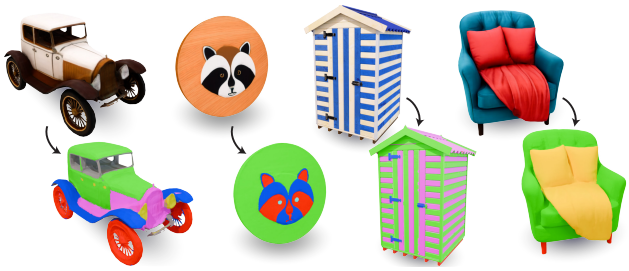

Figure 10. Automatic segmentation of meshes into materials.

to a 2D UV map. This enables trivial creation of material-ID maps, or changing the selected material. Here, because the similarities are directly projected to pixel values, we find it beneficial to use the hole-filling and sprinkle-removal techniques described in the original SAM2 paper [48].

We show results on the output of text-to-mesh generated assets and other meshes [64]. Using our automatic segmentation we can easily replace the diffuse textures on a text-to-3D generated asset with PBR materials.

Future work

We find our method to significantly improve material selection in 3D, however some limitations remain to be addressed. Selecting materials on objects like glass and mirror remains a challenge as it is unclear if a user would prefer to select the transparent/mirror material or what is behind/reflected. Our method also depends on precise 3D reconstruction for accurate material selection. Errors in depth reconstruction can cause noise in our point clouds and inaccurate lookups in novel views. Improving depth estimation in volumetric reconstruction will help mitigate this issue.

Conclusion

We present SAMa, a material selection model for 3D, leveraging a video model for cross-view consistency and a simple yet efficient projection to 3D in the form of a similarity point cloud. Our approach enables interactive material selection, visualization and downstream manipulation of the 3D assets. As we specialize the SAM2 video model to a new modality, we find that finetuning using videos is important and that 500 varied videos are enough to change the modality. We believe this opens interesting opportunities to explore selection across various modalities.

References

[1] Evermotion arch interior, 2021. https://evermotion.org/shop/cat/397/archinteriors. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 5, 6

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Opensurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on graphics (TOG)*, 32(4):1–17, 2013. 2

[4] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015. 1, 2

[5] Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. Color-and texture-based image segmentation using em and its application to content-based image retrieval. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 675–682. IEEE, 1998. 3

[6] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[8] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. *Corr*, 2023. 3

[9] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Chen Yang, Wei Shen, Lingxi Xie, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. In *NeurIPS*, 2023. 3

[10] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[11] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(9):2175–2188, 2013. 3

[12] Seokhun Choi, Hyeonseop Song, Jaechul Kim, Taehyeong Kim, and Hoseok Do. Click-gaussian: Interactive segmentation to any 3d gaussians. *ECCV*, 2024. 3

[13] Kristin J Dana, Bram Van Ginneken, Shree K Nayar, and Jan J Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics (TOG)*, 18(1):1–34, 1999. 1

[14] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)*, 37(4):1–15, 2018. 1

[15] Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. Flexible svbrdf capture with a multi-image deep network. In *Computer graphics forum*, pages 1–13. Wiley Online Library, 2019. 1

[16] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv*, 2024. 5

[17] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2

[18] Sagi Eppel, Jolina Li, Manuel Drehwald, and Alan Aspuru-Guzik. Learning zero-shot material states segmentation, by implanting natural image patterns in synthetic data. *arXiv preprint arXiv:2403.03309*, 2024. 2

[19] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *ICLR*, 2023. 3

[20] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2

[21] Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and P. J. Narayanan. Interactive segmentation of radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4201–4211, 2023. 3

[22] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. *arXiv preprint arXiv:2403.18118*, 2024. 3

[23] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973. 3

[24] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[25] Yiwei Hu, Chengan He, Valentin Deschaintre, Julie Dorsey, and Holly Rushmeier. An inverse procedural modeling pipeline for svbrdf maps. *ACM Transactions on Graphics (TOG)*, 41(2):1–17, 2022. 1, 3

[26] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 5

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[28] Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[29] Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. GARField: Group Anything with Radiance Fields, 2024. arXiv:2401.09419 [cs]. 2, 3

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

head, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *ICCV*, 2023. 3, 7

[31] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022. 2, 3

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 2

[33] Jason Lawrence, Aner Ben-Artzi, Christopher DeCoro, Wojciech Matusik, Hanspeter Pfister, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Inverse Shade Trees for Non-Parametric Material Representation and Editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 25(3), 2006. 3

[34] Daniel Lepage and Jason Lawrence. Material matting. *ACM Trans. Graph.*, 30(6):1–10, 2011. 3

[35] Anran Liu, Cheng Lin, Yuan Liu, Xiaoxiao Long, Zhiyang Dou, Hao-Xiang Guo, Ping Luo, and Wenping Wang. Part123: Part-aware 3d reconstruction from a single-view image. *arXiv*, 2024. 3

[36] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023. 3

[37] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[38] Xiaoyang Lyu, Chirui Chang, Peng Dai, Yang-Tian Sun, and Xiaojuan Qi. Total-decom: Decomposed 3d scene reconstruction with minimal interaction. *arXiv preprint arXiv:2403.19314*, 2024. 3

[39] Norberto Malpica, Juan E Ortuño, and Andres Santos. A multichannel watershed-based algorithm for supervised texture segmentation. *Pattern Recognition Letters*, 24(9-10): 1545–1554, 2003. 3

[40] Satoru Masubuchi, Eisuke Watanabe, Yuta Seo, Shota Okazaki, Takao Sasagawa, Kenji Watanabe, Takashi Taniguchi, and Tomoki Machida. Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials. *npj 2D Materials and Applications*, 4(1):3, 2020. 1

[41] 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. *Communications of the ACM*, 65(1):99–106, 2021. 5, 6

[42] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4080–4089, 2019. 2

[43] Fabio Pellacini and Jason Lawrence. Appwand: Editing measured materials using appearance-driven optimization. *ACM Trans. Graph.*, 26(3):54–es, 2007. 3

[44] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[45] Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Language-driven physics-based scene synthesis and editing via feature splatting. *arXiv preprint arXiv:2404.01223*, 2024. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[47] Trygve Randen and John Hakon Husoy. Filtering for texture classification: A comparative study. *IEEE Transactions on pattern analysis and machine intelligence*, 21(4):291–310, 1999. 3

[48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3, 5, 8

[49] Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G Schwing, and Oliver Wang. Neural volumetric object selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2022. 3

[50] Constantino Carlos Reyes-Aldasoro and Abhir Bhalerao. The bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition*, 39(5): 812–826, 2006. 3

[51] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. 3

[52] Sam Sartor and Pieter Peers. Matfusion: a generative diffusion model for svbrdf capture. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 1

[53] Gabriel Schwartz and Ko Nishino. Recognizing material properties from images. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1981–1995, 2019. 2

[54] Lavanya Sharan, Ruth Rosenholtz, and Edward H. Adelson. Accuracy and speed of material categorization in real-world images. *Journal of Vision*, 14(10), 2014. 2

[55] Prafull Sharma, Julien Philip, Michaël Gharbi, Bill Freeman, Fredo Durand, and Valentin Deschaintre. Materialistic: Selecting similar materials in images. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 1, 2, 3, 4, 5, 6

[56] Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik.

Match: Differentiable material graphs for procedural material capture. *ACM Transactions on Graphics (TOG)*, 39(6): 1–15, 2020. 1

[57] Randy M Sterbentz, Kristine L Haley, and Joshua O Island. Universal image segmentation for optical identification of 2d materials. *Scientific reports*, 11(1):5808, 2021. 1

[58] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, pages 443–453, 2022. 3

[59] Paul Upchurch and Ransen Niu. A dense material segmentation dataset for indoor and outdoor scene parsing. In *European conference on computer vision*, pages 450–466. Springer, 2022. 1, 2

[60] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines, 2024. 3

[61] Giuseppe Vecchio. Stablematerials: Enhancing diversity in material generation via semi-supervised learning. *arXiv preprint arXiv:2406.09293*, 2024. 1

[62] Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. Controlmat: a controlled generative approach to material capture. *arXiv preprint arXiv:2309.01700*, 2023. 1

[63] Ting-Chun Wang, Jun-Yan Zhu, Ebi Hiroaki, Manmohan Chandraker, Alexei A Efros, and Ravi Ramamoorthi. A 4d light-field dataset and cnn architectures for material recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 121–138. Springer, 2016. 2

[64] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 8

[65] Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu Fang. OmniSeg3D: Omniversal 3D Segmentation via Hierarchical Contrastive Learning, 2023. arXiv:2311.11666 [cs]. 3

[66] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21676–21685, 2024. 3