Fine-Grained Spatially Varying Material Selection in Images

JULIA GUERRERO-VIU, Universidad de Zaragoza - I3A, Spain and Adobe Research, United Kingdom
MICHAEL FISCHER, Adobe Research, United Kingdom
ILIYAN GEORGIEV, Adobe Research, United Kingdom
ELENA GARCES, Adobe Research, France
DIEGO GUTIERREZ, Universidad de Zaragoza - I3A, Spain
BELEN MASIA, Universidad de Zaragoza - I3A, Spain
VALENTIN DESCHAINTRE, Adobe Research, United Kingdom

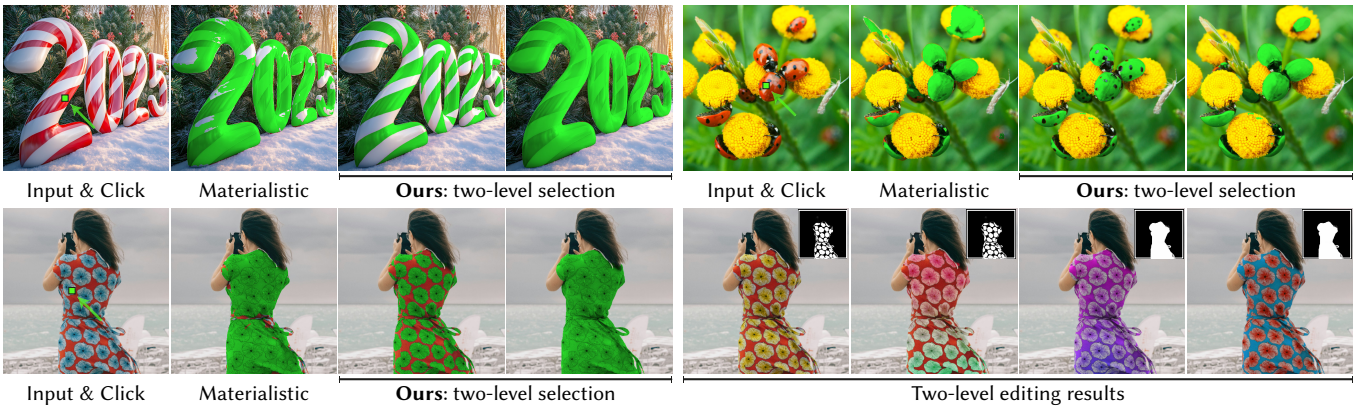

Fig. 1. Our proposed method allows fine-grained material selection in images on two different levels of granularity, significantly outperforming previous work (Materialistic [Sharma et al. 2023]) in selection accuracy and consistency. We show here results on challenging examples due to specular reflections (top left) and fine patterns outside the training data (top right, bottom left). Selection masks are shown as green image overlays. The bottom right row shows material editing results using our predicted two-level selection masks, with the masks shown as insets.

Selection is the first step in many image editing processes, enabling faster and simpler modifications of all pixels sharing a common modality. In this work, we present a method for material selection in images, robust to lighting and reflectance variations, which can be used for downstream editing tasks. We rely on vision transformer (ViT) models and leverage their features for selection, proposing a multi-resolution processing strategy that yields finer and more stable selection results than prior methods. Furthermore, we enable selection at two levels: texture and subtexture, leveraging a new two-level material selection (DuMaS) dataset which includes dense annotations for over 800,000 synthetic images, both on the texture and subtexture levels.

CCS Concepts: • **Computing methodologies** → **Image representations**.

Additional Key Words and Phrases: Material Selection

Authors' addresses: Julia Guerrero-Viu, Universidad de Zaragoza - I3A, Spain and Adobe Research, United Kingdom, juliagviu@unizar.es; Michael Fischer, Adobe Research, United Kingdom, mifischer@adobe.com; Iliyan Georgiev, Adobe Research, United Kingdom, igeorgiev@adobe.com; Elena Garces, Adobe Research, France, elenag@adobe.com; Diego Gutierrez, Universidad de Zaragoza - I3A, Spain, diegog@unizar.es; Belen Masia, Universidad de Zaragoza - I3A, Spain, bmasia@unizar.es; Valentin Deschaintre, Adobe Research, United Kingdom, deschain@adobe.com.

ACM 0730-0301/2025/12-ART185
https://doi.org/10.1145/3763332

ACM Reference Format:
Julia Guerrero-Viu, Michael Fischer, Iliyan Georgiev, Elena Garces, Diego Gutierrez, Belen Masia, and Valentin Deschaintre. 2025. Fine-Grained Spatially Varying Material Selection in Images. *ACM Trans. Graph.* 44, 6, Article 185 (December 2025), 11 pages. https://doi.org/10.1145/3763332

1 INTRODUCTION

Selection in images is an ubiquitous operation, enabling numerous downstream editing tasks. Given an image and a user input, such as a clicked pixel, selection aims to identify other pixels in the image that share a particular property with the user input. Various modalities of selection exist, for example by color [Belongie et al. 1998], object [Ravi et al. 2024], or material [Sharma et al. 2023]. The latter, in particular, facilitates selecting parts of an object or multiple objects easily, allowing further editing of regions that share the same material. Understanding which materials are the same in images also provides key information for inverse rendering and scene understanding tasks [Nimier-David et al. 2021].

Recent work [Sharma et al. 2023] has proposed to extract features from a pre-trained vision transformer (ViT) [Caron et al. 2021] and spatially process them for selection. However, the ViT's tokenization patch size and small operating resolution limit the selection precision, especially around edges and thin structures. Moreover, to fine-tune the ViT for material selection, Sharma et al. used a synthetic dataset which defined materials as *textured* surfaces (e.g., a wallpaper with a repeating pattern texture); this does not allow

selecting image elements sharing similar *subtexture* properties (e.g., a part of the repeating pattern with a similar appearance).

In this work, we tackle these limitations and propose a model for more precise and robust selection, capable of selecting at both texture and individual subtexture levels. To extract more discriminative features for material selection we leverage DINOv2 [Oquab et al. 2024]. Its refined architecture and use of larger patch sizes to extract better contextual information outperforms other existing models like DINOv1 [Caron et al. 2021] or Hiera [Ryali et al. 2023]. Unfortunately, the ViT's native resolution (518x518 for DINOv2) typically requires downscaling the input images before extracting features. This translates into poor performance when selecting boundaries or areas with fine details. To mitigate this, we devise a multi-resolution approach, splitting the original input image into tiles, and processing both a downscaled version of the original image and the tiles at the native ViT resolution, before concatenating them. This allows our aggregated features to capture both the high-level context provided by the downscaled image and finer-grained details from the better-resolved individual tiles, significantly enhancing selection quality.

We also enhance the training process by sampling multiple pixels from various materials in each training image, instead of a single pixel, improving both training stability and material selection consistency. Finally, we have designed a new synthetic dual-level material selection (DuMaS) dataset which comprises 800,000+ images of indoor and outdoor scenes. It is over 16× larger than the recent Materialistic dataset [Sharma et al. 2023] and includes annotations at two selection levels: texture and subtexture. The texture level is the same as in the Materialistic dataset and targets the selection of surfaces which belong to the same texture – e.g., on a chess board, black and white squares would be selected together. Our new subtexture level adds a finer-grained option to enable the selection of parts of textures with similar appearance, grouping together individual texture components – in the chess board example, black and white squares would be selected separately (see also Fig. 1). This two-level approach also improves selection quality and consistency.

We evaluate the quality and consistency of our model both qualitatively and quantitatively, under different scenarios: varying the selected pixel, the image's field of view, or its lighting. We ablate our multi-resolution processing and training schemes as well as the impact of our dataset on selection accuracy. Finally, we compare our method to two state-of-the-art methods, namely Materialistic [Sharma et al. 2023] and the Segment Anything Model 2 (SAM2) fine-tuned for materials [Fischer et al. 2024; Ravi et al. 2024], showing significant improvement over both.

In summary, we propose a material selection model with improved accuracy and support for both texture- and subtexture-level selection in images thanks to the following contributions:

- A material selection architecture that offers two-level control of the selection granularity;
- An improved training scheme for multi-pixel sampling and multi-resolution processing;
- A large synthetic dataset comprising annotations on both texture and subtexture levels.

We will release our model and test code, as well as a significant subset of our dataset[1].

2 RELATED WORK

Object segmentation and selection. Recent research in object segmentation and selection has significantly advanced both 2D and 3D image understanding. In 2D images and videos, models such as SAM and SAM2 [Kirillov et al. 2023; Ravi et al. 2024] enable selection/segmentation (and tracking) of objects. In 3D representations, methods for radiance fields [Bhalgat et al. 2023; Fu et al. 2022; Kim et al. 2024], point clouds [Tchapmi et al. 2017] and inter-surface mappings [Morreale et al. 2024] leverage geometric or multi-view cues to achieve segmentation and selection in a consistent manner.

These techniques, however, operate on an *object* level, targeting the selection of distinct objects in an image. In contrast, we target fine-grained material selection, where our goal is to identify regions sharing the clicked pixel's material, irrespective of the object(s) onto which the material is applied: a single object can contain multiple materials and a single material can appear on multiple objects.

Material segmentation and selection. Material selection from 2D images is a non-trivial problem, since a surface's perceived appearance can be influenced by numerous factors beyond its reflectance properties, such as its geometry [Boyaci et al. 2003], illumination [Fleming et al. 2003], or the surrounding surfaces and object identity [Sharan 2009; Sharan et al. 2014]. Early material selection algorithms were built around low-level features such as color- and texture descriptors [Belongie et al. 1998; Haralick et al. 1973] or hand-crafted filters and heuristics [Leung and Malik 2001; Malpica et al. 2003]. Additionally, selection is simplified when the image can be decomposed into (potentially disjoint) regions of varying reflectance properties, as shown by Lensch et al. [2003] or, more recently for radiance fields, by Verbin et al. [2022]. However, this often requires additional information such as multi-view images or specialized capture hardware [Xue et al. 2020], and single-image decomposition into physical components remains challenging [Kocsis et al. 2024; Zeng et al. 2024; Zhu et al. 2022].

Leveraging deep networks' strong classification capabilities, methods targeted material semantic classification [Bell et al. 2015; Cimpoi et al. 2014; Sumon et al. 2022]. Others proposed to train material perceptual similarity metrics for classification of complete photographs containing materials [Lagunas et al. 2019; Sharan et al. 2013].

Closest to our method is Materialistic [Sharma et al. 2023], a material selection method leveraging large vision models' features [Caron et al. 2021]. While we also target material selection, we differ from this work in multiple ways: we improve feature processing through a multi-resolution approach, preserving more signal throughout the pipeline and hence improving selection accuracy and consistency. Further, we extend their proposed definition of pixels with similar materials (i.e., pixels belonging to the same texture) by adding a finer-grained selection level we call subtexture. This enables the selection of texture sub-elements which share the same appearance. Additionally, despite the existence of several datasets [Bell et al. 2013; Deschaintre et al. 2018; Eppel et al. 2024; Murmann et al. 2019a; Sharan et al. 2014; Upchurch and Niu 2022; Vecchio and Deschaintre

[1]Project website: https://graphics.unizar.es/projects/MatSelection/

2024; Wang et al. 2016] for semantic material classification or material selection, they typically provide coarse semantic annotations. For material selection, Sharma et al. [2023] used 50,000 synthetic renderings with ground-truth annotations. However, their dataset contains few, strongly textured materials with only texture-level annotations. In contrast, we design a significantly larger (800,000+ images) synthetic dataset containing many spatially varying textures with annotations both at the texture and subtexture levels.

Image encoders. Most modern object and material selection methods utilize the features of big vision models, ViTs or masked autoencoders (MAEs) pre-trained on large collections of natural images, as backbones. ViTs like DINO and DINOv2 [Caron et al. 2021; Oquab et al. 2024] learn priors encoded in these images (e.g., the appearance of shadows and reflections), making them well-suited for generalization to vision tasks such as object or material selection. A defining feature of both ViTs and MAEs is the use of self-attention [Vaswani 2017], enabling global information sharing across the image.

However, attention is a memory-intensive operation, limiting both architectures in terms of input patch size and resolution. The recently introduced Hiera architecture [Bolya et al. 2023; Ryali et al. 2023] mitigates this via hierarchical feature extraction and smaller kernels, leading to sharper feature boundaries. Scaling-on-scales (S^2, [Shi et al. 2024]), proposes to upscale the input to different resolutions and concatenate the resulting features, obtaining comparable (and occasionally superior) performance to larger vision models.

In this work, we also leverage the features of big vision models, evaluate various options [Caron et al. 2021; Oquab et al. 2024; Ryali et al. 2023] and adjust the scaling proposed by S^2 to fit our selection context, significantly improving selection quality and consistency by preserving more of the input image's available information.

3 METHOD

We design a new state-of-the-art model for material selection in images that addresses the precision and robustness limitations of prior approaches, and build a large synthetic dataset with material annotations at both texture and subtexture levels.

Our model takes as input an image and a query pixel, and outputs per-pixel material similarity to this query at both texture and subtexture levels, as demonstrated in Fig. 1. This similarity can then be thresholded into a binary selection mask. We build our model on the architecture of Materialistic [Sharma et al. 2023], to which we make three key modifications, described in detail in Section 3.1: (1) multi-resolution processing and feature aggregation to improve precision, (2) two-level representation to allow for texture- and subtexture-level selection, and (3) multiple query sampling during training to improve robustness.

An overview of our pipeline is shown in Fig. 2. We first extract features from the input image with a ViT encoder at different resolutions. Our model can be used with different encoders, and we evaluate alternatives in Section 4; for convenience, unless explicitly stated, our explanations and results in the rest of the paper use DINOv2. The extracted features are then processed through our proposed multi-resolution aggregation scheme, and fed into a material selection head akin to that of Materialistic, modified for two-level selection. This head takes the aggregated features, and computes the cross-similarity between the features and the image patch centered around the user-provided query pixel, yielding query-conditioned features. Both in Materialistic and in our work, the feature extraction by the encoder – and therefore the subsequent processing – is done from four different transformer blocks. After the cross-similarity feature weighting layer, the query-conditioned features at different blocks are combined and bilinearly upsampled via convolutional layers. Finally, a channel-wise sigmoid is applied to yield the final per-pixel similarity. The model is trained minimizing a binary cross-entropy (BCE) loss on our DuMaS dataset (Section 3.2).

3.1 Two-Level Multi-Resolution Material Selection

We now describe here the three key components that enable our model to perform fine-grained, spatially varying material selection (§3.1.2 to §3.1.4). Prior to that (§3.1.1), we discuss our feature extractor of choice and its advantages over previous alternatives.

3.1.1 Image encoder. Different encoders can be used to extract features from the input image. In this work, we rely on DINOv2 [Oquab et al. 2024], a pre-trained self-supervised ViT that builds upon the foundation of DINO [Caron et al. 2021], incorporating improvements in architecture and training strategies. In particular, it utilizes a larger and more diverse dataset, better augmentations, and a more refined distillation process to enhance feature extraction capabilities with richer representations. DINOv2 also adopts a larger patch size ($ps = 14$, almost twice as large as DINO), which captures higher contextual information but results in lower-resolution tensors at $1/14$ of the input image resolution. In our experiments, using DINOv2 as an encoder yields more discriminative features, particularly for disentangling appearance from lighting variations, and more confident selection predictions compared to using the original DINO. Its bigger patch size slightly degrades performance on edges and small details on cluttered scenes when using a single resolution approach. However, this is mitigated by our multi-resolution approach, where a bigger patch size does not hamper performance. In the following, we refer to DINOv2 as the pre-trained embeddings from variant ViT-B/14 with $ps = 14$ and feature dimension $d = 768$, our best performing configuration. Following previous work [Sharma et al. 2023], we extract four intermediate features (both local and global tokens) from transformer blocks at indices 2, 5, 8, and 11.

3.1.2 Multi-resolution processing and feature aggregation. Despite the impressive performance of ViTs as feature extractors, their tokenization significantly reduces the resolution of the input images, degrading precision on edges and thin structures. Therefore, we propose to extract features from regions of the input image at multiple resolutions, improving the sharpness and robustness of our material selection results. Given an input image I, we construct a pyramid of n resolutions $\{I_1, I_2\}$, where I_1 represents the image at the input resolution $r \times r$ of the image encoder ($r = 518$ for DINOv2), and I_i represents 2^{i-1} higher-resolution versions, ensuring that their resolution remains divisible by the ViT patch size ps (to maintain alignment between the feature maps across resolutions). Each I_i is split into 2^i non-overlapping tiles, each of them with the resolution of I_1, $r \times r$. In practice, our 1024×1024 training images are downsampled to the image encoder's native resolution for I_1, while

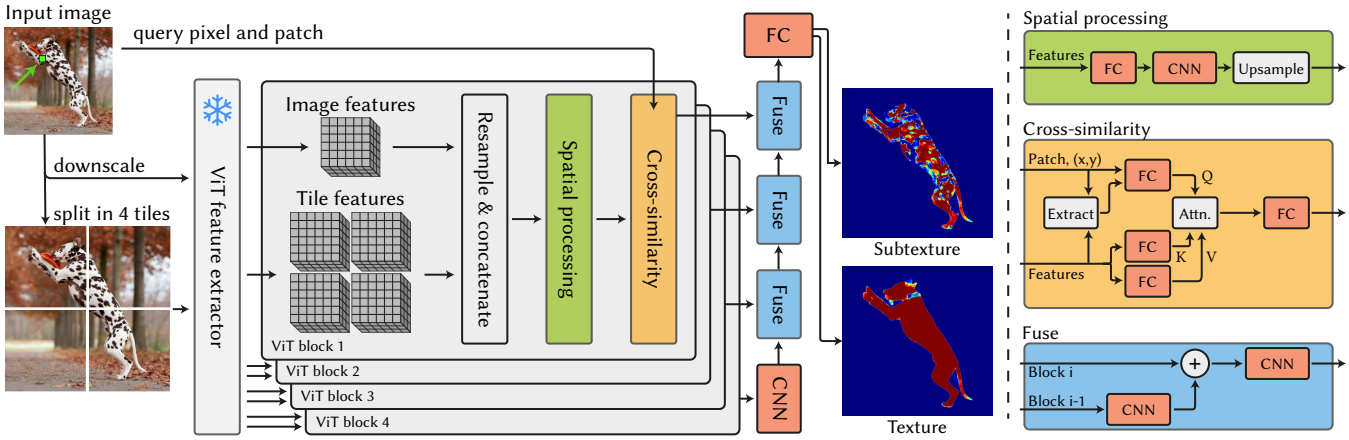

Fig. 2. **Model architecture.** Our model extracts ViT features at different resolutions, for the full image and for each separate tile, leading to a multi-resolution feature encoding. The subsequent spatial processing layers upscale the features before the cross-similarity computes the attention with respect to the clicked image pixel and patch. We exploit the information encoded at different depths by repeating this process across four ViT levels, before fusing the output with a residual CNN and feeding it to our two-level selection head, producing selections at both subtexture and texture level. The ViT is frozen, the red blocks are trained. FC is short for fully-connected network.

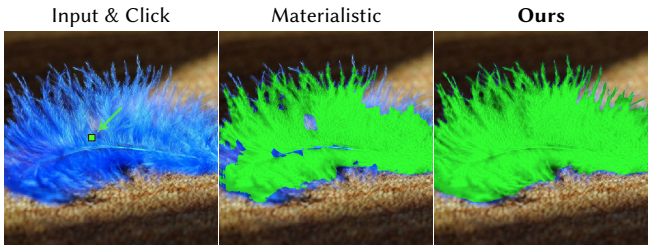

Fig. 3. **Fine-grained material selection.** Our multi-resolution aggregation allows to recover thin structures, overcoming a limitation acknowledged in previous work [Sharma et al. 2023].

I_2 makes use of the full image resolution, slightly resized and split into four tiles. These tiles are fed into the image encoder, which, as explained in §3.1.1, extracts features at four different scales (from four transformer blocks, $j = 1, ..., 4$). This yields feature maps F_{ij}, extracted at image resolution i and block j. In our experiments, we evaluated two and three resolution levels and found that three levels did not provide tangible benefits given our training data resolution, while requiring significantly more compute for the additional tiles' features. We therefore use two levels $n = 2$, effectively achieving twice higher resolution than the single resolution approach.

We then introduce a feature aggregation module to integrate information across resolutions (see Fig. 2). We first resample all feature maps F_{ij} to match the target resolution (with different target resolutions for each block j) and concatenate the features along the channel dimension, making sure to preserve the original image layout when re-arranging the tiles:

$$F_{\text{agg},j} = \text{Concat}(\{\text{Resample}(F_{ij})\}_{i=1}^{n}), \quad (1)$$

where Resample$(\cdot)$ denotes bilinear upsampling and area downsampling to match the target resolution, and Concat$(\cdot)$ represents concatenation along the channel axis.

Previous related work on multi-resolution aggregation [Shi et al. 2024] always downsamples the higher-resolution features before concatenating them. In contrast, we adapt our up- or down- sampling strategy to the various target feature resolutions per block, where we make sure to preserve the highest-resolution spatial details (see supplemental document for resampling details). For aggregation, concatenation outperforms averaging as it preserves more information by retaining both fine-grained and broader contextual features that are later processed by the material selection head. We illustrate the benefits of this module in Fig. 3, where our multi-resolution processing allows to sharply select the thin structures of the feather.

3.1.3 Two-level representation. Our model outputs material similarity representations at two levels, texture-level and subtexture-level, in the form of two different similarity maps. To achieve this, we modify the output per-pixel similarity score of our model to have two output channels, each of them using a sigmoid activation function. We train both channels jointly, minimizing the BCE loss on our DuMaS dataset.

3.1.4 Multi-query sampling during training. To improve training stability and robustness, we sample multiple query pixels per image during training, covering diverse materials, which has proven helpful in the object selection context in previous work [Ravi et al. 2024]. This strategy is similar in spirit to using a higher effective batch size, re-using the image encoder computation and making the optimization more stable. We show in Section 4 and the supplemental document how the multiple query sampling benefits the robustness and accuracy of our material selection results.

3.2 DuMaS training dataset

Most existing material datasets with dense image-space annotations [Bell et al. 2013; Murmann et al. 2019a; Upchurch and Niu 2022] contain semantic annotations of material classes, such as "wood" or

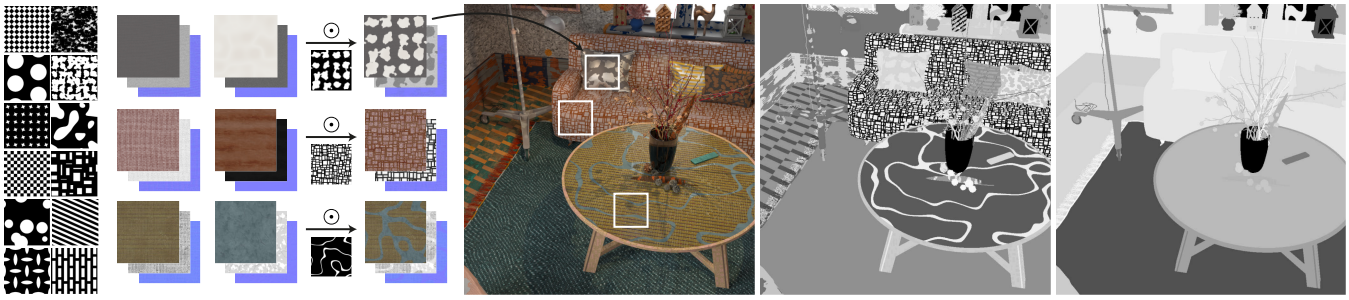

Fig. 4. **DuMaS training dataset creation.** Starting from twelve base noise patterns from Substance Designer, we vary their parameters to create 1,571 binary masks which we use to randomly combine 3,026 stationary reflectance maps to generate 10,881 materials (see text for more details). We assign these material maps to objects of our 132 scenes under different combinations, and render videos by varying camera viewpoints, resulting in 816,415 individual images. For each image, our dataset includes dense annotations at both subtexture and texture level; annotation IDs are mapped to gray levels for visualization.

"metal", but lack fine-grained annotations of variations within a class. The dataset of Materialistic [Sharma et al. 2023] does contain fine-grained per-pixel material annotations for 50,000 synthetic images; however, such annotations do not distinguish between different individual texture components within a single texture (e.g., black and white squares on a checkered pattern). Such distinction is necessary to enable our two-level material selection. Therefore, we render a new large-scale synthetic dataset of over 800,000 images including fine-grained material annotations at both texture- and subtexture-level, named DuMaS (Dual-level Material Selection) dataset.

We do so by rendering 132 different synthetic scenes (119 indoor, 13 outdoor) from the Evermotion collection [eve 2024] with random materials alongside texture and subtexture annotations. We first create the material maps which will be applied to objects in different scenes. We select a set of 3,026 stationary[2] source materials from the Adobe 3D Assets library, constituting our initial set of *reflectance* maps. We then create 1,571 different binary masks by randomly sampling generative parameters from twelve noise patterns in Substance Designer. Combining a binary mask (used as alpha channel) with a pair of reflectance maps leads to a new *texture* map (see Fig. 4, left). In particular, for each binary mask we sample five pairs of reflectance maps without repetition, for a total of 7,855 different texture maps. Together, the reflectance and texture maps yield 10,881 unique material maps. Each reflectance and each texture map are assigned a unique ID, which is then used for image annotation at texture and subtexture levels.

For each scene, we create five different material assignments by randomly sampling from our set of material maps, for a total of 660 different scene configurations. We maintain the original relationships between objects, so that objects (or parts of them) with the same material in the original scene will also share the same material after our assignment. Transparent and emissive materials in the original scenes are left unmodified. The final annotations, in the form of per-pixel material IDs at two levels, subtexture and texture, are as follows: if an object is assigned a material from the initial reflectance set, both levels will share the same ID. If it is assigned a material from the texture set, the texture-level annotation will store its ID, and the subtexture level will store the ID of its assigned constituent reflectance map (see Fig. 4, right).

We render videos of up to one minute for every scene, following a camera trajectory that mimics a first-person exploration of the scene, at 30 fps. Each frame is rendered at 1024×1024 resolution with 256 samples per pixel using Blender Cycles 4.2. Our full DuMaS dataset contains 816,415 frames (around 250 days of GPU rendering time). Including videos instead of independent images allows us to use our dataset to fine-tune video selection models like SAM2.

3.3 Implementation Details

We train our model on our DuMaS dataset for 10 epochs, using the Adam optimizer with learning rate 1e-4 on four A100-40GB GPUs, using the DDPS distributed strategy and batch size 4 images per GPU. During training, we sample random crops at ViT resolution and apply random exposure, saturation, and brightness augmentations. For our multi-query sampling, we uniformly sample 10 pixels within the crop. See the supplemental document for implementation details.

4 RESULTS

In this section we present qualitative and quantitative evaluations, as well as a robustness analysis of our model with respect to user inputs, zoom levels, illumination, and sensitivity to the selection threshold. Finally, we ablate several aspects of our architecture and show application examples for material editing at texture and subtexture levels. We will publicly release our evaluation framework's code and create a benchmark for both material selection quality and robustness, to facilitate future work.

4.1 Real-World Test Datasets

We evaluate our method mainly on two test datasets that contain *in-the-wild*, real-world images: (i) the Materialistic Test dataset [Sharma et al. 2023], containing 50 images annotated at *texture* level; and (ii) the Two-Level Test dataset, our new, manually annotated test set containing 20 images with annotations at both *subtexture* and *texture* levels. This new dataset contains challenging real-life scenarios with strong lighting variations, indoor and outdoor instances, cluttered scenes with thin structures, and

[2]A class of materials consisting of plain colors, or structures that (randomly) repeat over the surface [Aittala et al. 2015].

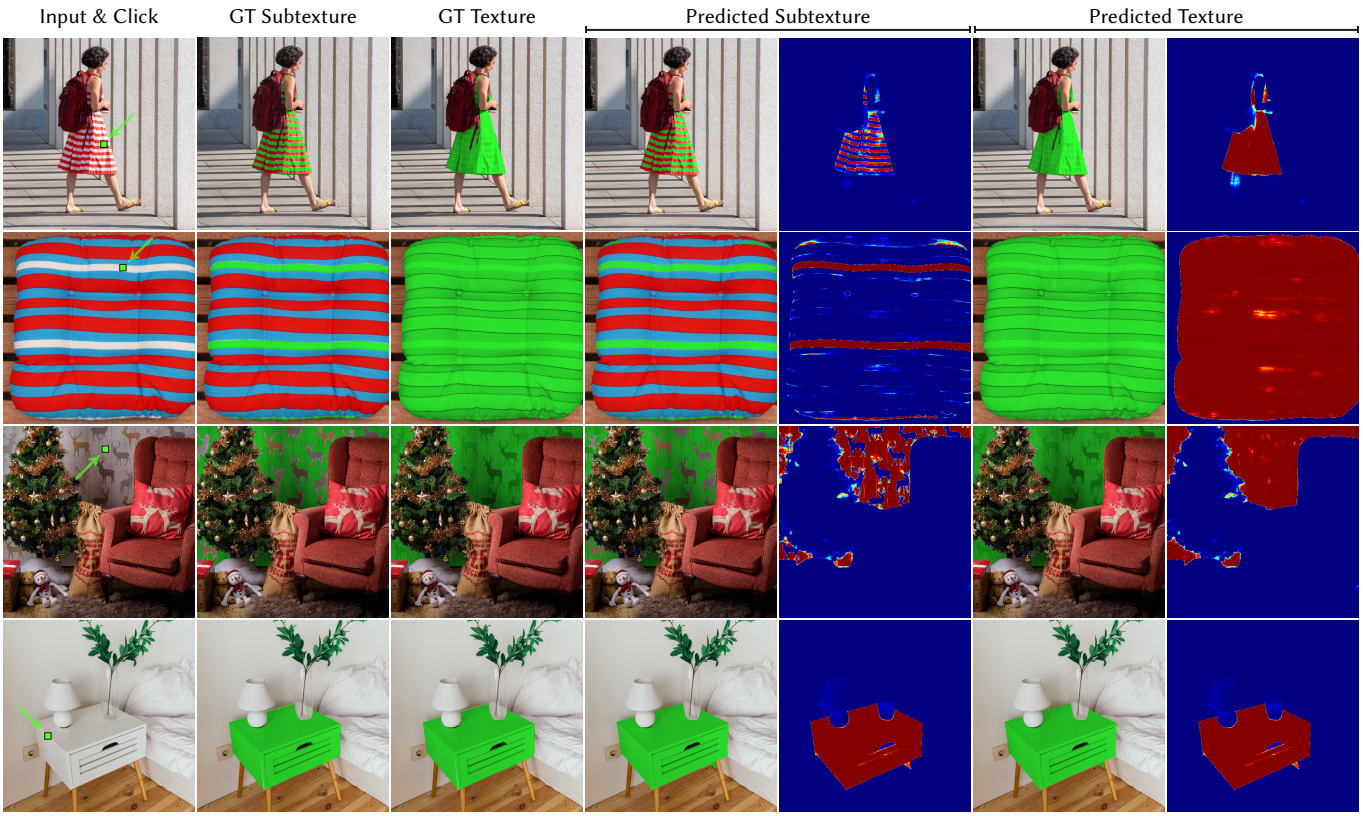

Fig. 5. **Qualitative results**. We show results of our method (two-level selection) for images in our real-world test datasets. For each example and level, we show the predicted, binarized selection masks as green image overlays and the similarity score before thresholding in false color (blue low, red high). Our method works well at both subtexture and texture levels even for cluttered scenes and patterns very different from the ones in the training set (third row, wallpaper), and with challenging examples where objects in the scene share the same color (bottom row). In non-spatially varying materials (bottom row) our model's predictions for both levels are consistently the same.

high-frequency appearances. We show this new test set in the supplemental document and will release it upon publication. In both datasets, we sample 10 query pixels per image for evaluation, generating 500 test cases for the Materialistic Test dataset, and 200 test cases for our Two-Level Test dataset.

4.2 Evaluation

Qualitative results. Figure 5 presents the results of our two-level selection method in challenging scenarios. These images contain a diverse range of cases, demonstrating the robustness of our approach. Our method accurately selects textured areas and discriminates subtextures, as shown in rows one to three. The third row shows a highly cluttered scene, with several spatially-varying materials, in which our method successfully makes the right selection at both levels, despite the pattern being very different from the training ones. The last row showcases a challenging, mostly-white scene with varying reflectances, where our method correctly identifies the table material at both texture and subtexture levels. We show additional results in the supplemental document.

Comparisons. We compare our model with two recent state-of-the-art methods: Materialistic [Sharma et al. 2023], and SAM2 [Ravi et al. 2024]. Since SAM2 was originally trained for *object* selection, we fine-tune it with our DuMaS dataset to the material selection task (see supplemental document for details on the fine-tuning). For completeness, we also report results of Materialistic trained on our DuMaS dataset, as well as using DINOv2 as encoder. For the SAM2 model evaluations we duplicate the first frame and use the output of the second frame [Fischer et al. 2024], which significantly improves its overall confidence and accuracy. We also include results without this frame duplication in the supplemental document.

Figure 6 presents *qualitative* results for texture-level selection on real images, from a single query pixel, comparing our method with Materialistic and SAM2 fine-tuned with our DuMaS dataset. Overall, our method is more accurate and results in fewer false positives. Compared to Materialistic, it consistently produces sharper and cleaner predictions across all examples; this is probably due partly to our DuMaS dataset, but also to Materialistic's reliance on features at a single resolution, which constrains the precision of the output, and results in blurry edges and a diminished capacity to accurately select fine details. In contrast, the fine-tuned SAM2 model, using the Hiera

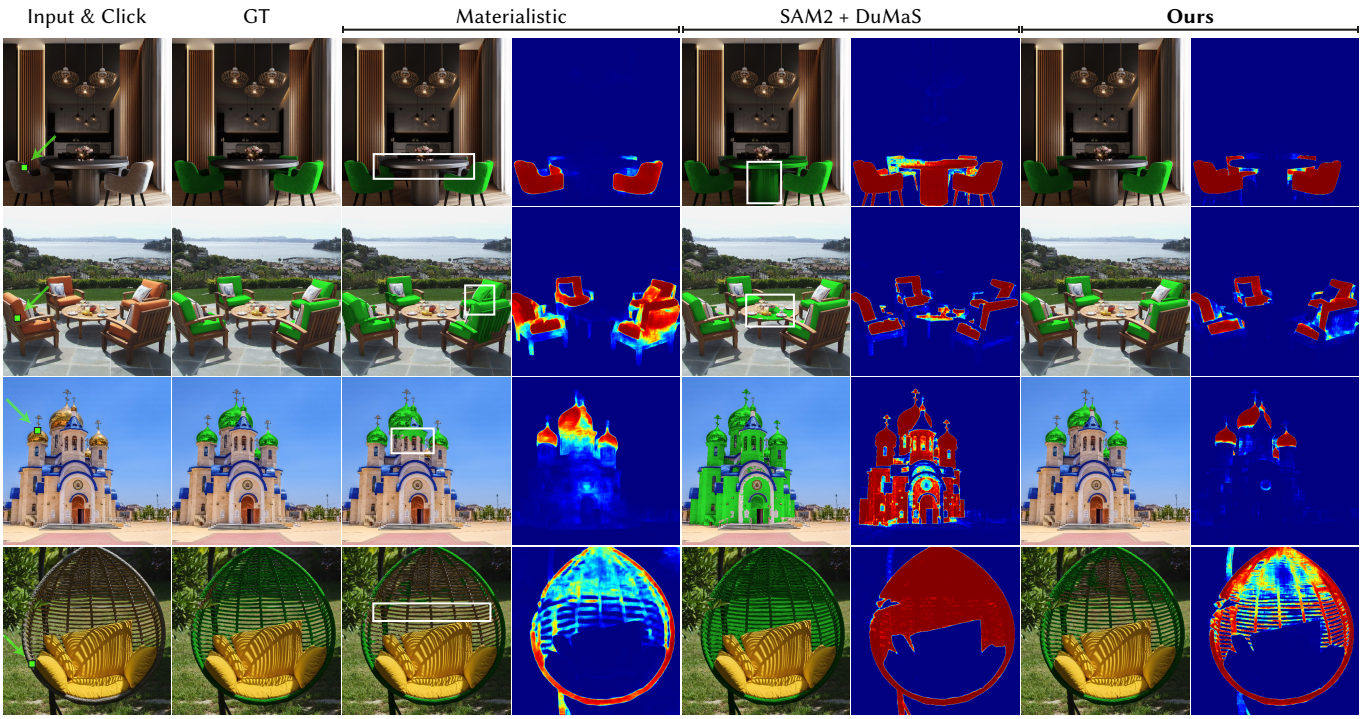

Fig. 6. **Qualitative comparison.** Texture-level selection results for Materialistic, SAM2 fine-tuned on our DuMaS dataset, and our method. The white squares highlight areas where the methods struggle. Materialistic fails to select relevant areas, especially in the presence of small or thin structures (first and last rows), or to produce sharp, clear selections (second and third rows). SAM2 has improved sharpness, but produces many false positives in areas of similar appearance.

encoder (with a smaller patch size and twice the input resolution), produces sharp edges and handles thin structures, but tends to over-select areas of the image of relatively similar appearance. We also show additional comparisons in the supplemental document.

Table 1 reports *quantitative* results for three different metrics on both test datasets. The L1 metric measures the pixel-wise difference between the predicted values and the ground truth mask; lower values indicate better agreement, which can be interpreted as higher prediction confidence. The Intersection over Union (IoU) measures the overlap between the ground truth mask and the predicted mask, and the F1 score is the harmonic mean of precision and recall, providing a single metric that balances both aspects of a model's accuracy. IoU and F1 are computed by binarizing the outputs with a threshold, which we fix to 0.5. Metrics for tasks outside a method's original training scope are shown in the table for reference, but marked in gray (e.g., our subtexture level is not supported by previous works). Table 1 supports our qualitative comparisons; when trained on DuMaS, Materialistic slightly improves its performance, due to the larger and more diverse dataset. Still, our method consistently achieves the best performance for both subtexture and texture level selection. Interestingly, when changing the image encoder of Materialistic to DINOv2 (second row), the overall quantitative performance of Materialistic is slightly degraded: despite the stronger feature representation capabilities of DINOv2, its larger patch size (14 vs. 8) yields lower resolution features, significantly impacting performance on edges. This highlights the benefits of our multi-resolution pipeline, compensating for this limitation.

Table 1. Mean results of various methods (rows) across different metrics (subcolumns) on two real-world test datasets (columns) . For single-level methods trained on our DuMaS dataset (+DuMaS), we train two separate models (one per level) and report the result of the relevant one per column. Gray text indicates cases where the model is evaluated on a different task to the one it is trained for, and boldface marks the best results.

	Materialistic Test			Two-Level Test					
	Texture Level			Subtexture Level			Texture Level		
	L1 ↓	IoU ↑	F1 ↑	L1 ↓	IoU ↑	F1 ↑	L1 ↓	IoU ↑	F1 ↑
Materialistic	0.057	0.858	0.906	0.144	0.513	0.629	0.112	0.657	0.749
Materialistic + DINOv2	0.069	0.838	0.890	0.202	0.463	0.581	0.135	0.636	0.728
Materialistic + DuMaS	0.043	0.858	0.904	0.092	0.615	0.717	0.101	0.680	0.765
SAM2 *obj.sel.	0.086	0.633	0.708	0.121	0.426	0.539	0.118	0.558	0.632
SAM2 + DuMaS	0.060	0.784	0.847	0.103	0.576	0.681	0.071	0.730	0.799
Ours	**0.030**	**0.896**	**0.935**	**0.071**	**0.673**	**0.766**	**0.069**	**0.750**	**0.823**

4.3 Robustness Evaluation

To evaluate the robustness of our method, we assess prediction *consistency* across query pixels, zoom levels, and different illuminations. Moreover, we evaluate how robust the predictions are to different thresholds, which we call prediction *confidence*.

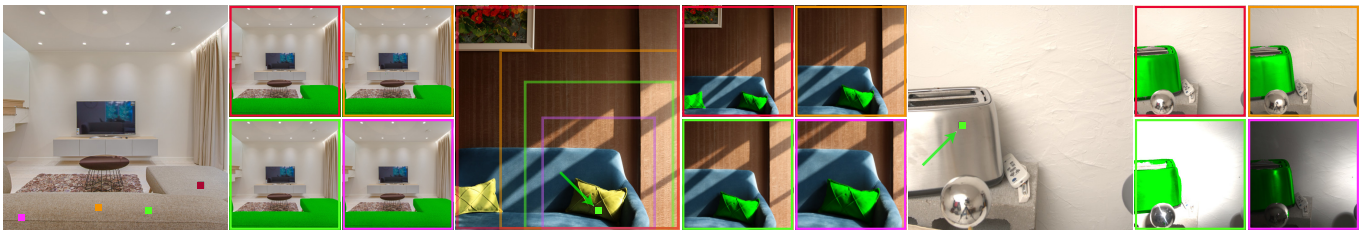

Fig. 7. **Robustness.** Robustness evaluation of our method with respect to the clicked pixel (left subset), the image crop with increasing zoom levels (middle subset), and illumination changes (right subset). The images are challenging due to similar albedo (left), strong shading variations on the cushions (middle) and specular highlights on the toaster (right). For the false-color material similarity maps and comparisons against Materialistic and SAM2, please see the supplemental document, Figs. 9-12.

To test *pixel consistency*, we randomly sample up to three materials per image, and five query pixels per material: Ideally, the selection predictions for all query pixels belonging to the same material would be identical. With regard to *zoom consistency*, we create five different crops with increasingly large zoom levels centered on each query pixel, and evaluate to what extent the same query pixel results in the same selection in all cases. For *illumination consistency*, we use the Multi-Illumination dataset [Murmann et al. 2019b], which includes thirty in-the-wild scenes captured under twenty-five different illuminations. We sample up to three materials per scene and measure consistency of the selections across illuminations.

For all consistency evaluations, we compute the average pairwise Hamming distance (after binarizing the masks with a threshold of 0.5), with lower Hamming distances indicating higher consistency.

As shown in Table 2, our method demonstrates significantly higher consistency compared to Materialistic, achieving approximately 1.8× lower Hamming distance. When Materialistic is trained on our DuMaS dataset, its consistency improves substantially, particularly across query pixels. This improvement underscores the benefits of our larger-scale dataset, which features a more diverse range of materials. Our architecture and training method further enhance consistency, especially noticeable on the more challenging Two-Level Test and Multi-Illumination datasets. Comparing to SAM2 fine-tuned on our DuMaS dataset, our method shows a slight improvement in consistency; moreover, as showed in Table 1, our results are more accurate.

Fig. 7 shows qualitative examples for pixel, zoom and illumination consistency. In difficult scenarios, such as a sofa surrounded by surfaces of similar color (left), or a cushion with sharp shadows (middle), our predictions remain consistent and highly confident, clearly outperforming previous work. Regarding illumination, our predictions remain reasonably consistent even under very strong changes, as shown in Fig. 7 (right). For qualitative comparisons to Materialistic and SAM2, as well as a video showing the consistency of our method, please refer to the supplemental material.

Finally, we evaluate the robustness of our method in terms of prediction *confidence* by measuring the sensitivity of the selection result (the final binary mask) to the threshold over the similarity score. Our approach outperforms prior works in this regard, with full metrics provided in the supplemental material.

Table 2. Consistency results of various methods (rows) across different query pixels, zoom levels, and illuminations (subcolumns) on three real-world test datasets (columns). For single-level methods trained on our DuMaS dataset (+DuMaS), we train two separate models (one per level) and report the result of the relevant one per column. All are mean Hamming distances (lower is better). Note that only consistency is measured here, without assessing accuracy.

	Materialistic Test Texture Level		Two-Level Test Subtexture Level		Two-Level Test Texture Level		Multi-Illumination Texture Level
	Pixel ↓	Zoom ↓	Pixel ↓	Zoom ↓	Pixel ↓	Zoom ↓	Illumination ↓
Materialistic	0.041	0.121	0.082	0.200	0.082	0.200	0.071
Materialistic + DuMaS	0.028	0.087	0.077	0.222	0.059	0.161	0.065
SAM2 + DuMaS	0.025	0.080	0.053	**0.165**	0.059	**0.111**	0.039
Ours	**0.024**	**0.071**	**0.052**	0.166	**0.041**	0.112	**0.034**

Table 3. Mean results of ablations of our method (rows) across different metrics (subcolumns) on the two test real-world evaluation datasets and a challenging subset (columns). For the Single Level ablation, we train two separate models (one per level) and report the result of the relevant one per column.

	Materialistic Test Texture		Two-Level Test Subtexture		Two-Level Test Texture		Challenging Subset Subtexture		Challenging Subset Texture	
	L1 ↓	IoU ↑	L1 ↓	IoU ↑	L1 ↓	IoU ↑	L1 ↓	IoU ↑	L1 ↓	IoU ↑
Ours, DINO	0.045	0.850	0.094	0.616	0.097	0.698	0.107	0.591	0.104	0.644
Ours, Hiera	0.046	0.838	0.093	0.593	0.080	0.716	0.112	0.581	0.099	0.685
Ours, w/o Multi-Res.	0.033	0.889	0.081	0.637	0.075	0.740	0.130	0.512	0.122	0.585
Ours, w/o Multi-Sampl.	0.036	0.893	0.083	0.622	0.088	0.680	0.133	0.507	0.143	0.517
Ours, Single Level	0.037	0.888	0.077	0.643	0.081	**0.750**	0.106	0.581	0.099	0.685
Ours, Full	**0.030**	**0.896**	**0.071**	**0.673**	**0.069**	**0.750**	**0.068**	**0.694**	**0.058**	**0.763**

4.4 Ablations

We next analyze the impact of our most critical design decisions. All variants have been trained the same number of epochs with the full DuMaS dataset unless otherwise stated. Our full model includes the DINOv2 image encoder, multi-resolution feature aggregation (Multi-res., Section 3.1.2), and multiple query sampling (Multi-sampl., Section 3.1.4).

We first explore the effect of the image encoder (Table 3, first two rows). Replacing DINOv2 features with DINO or Hiera encoding produces less accurate selections and with noticeably less confidence (lower L1) in all datasets. Both DINO and Hiera seem to be more biased towards color and lighting than DINOv2, as shown

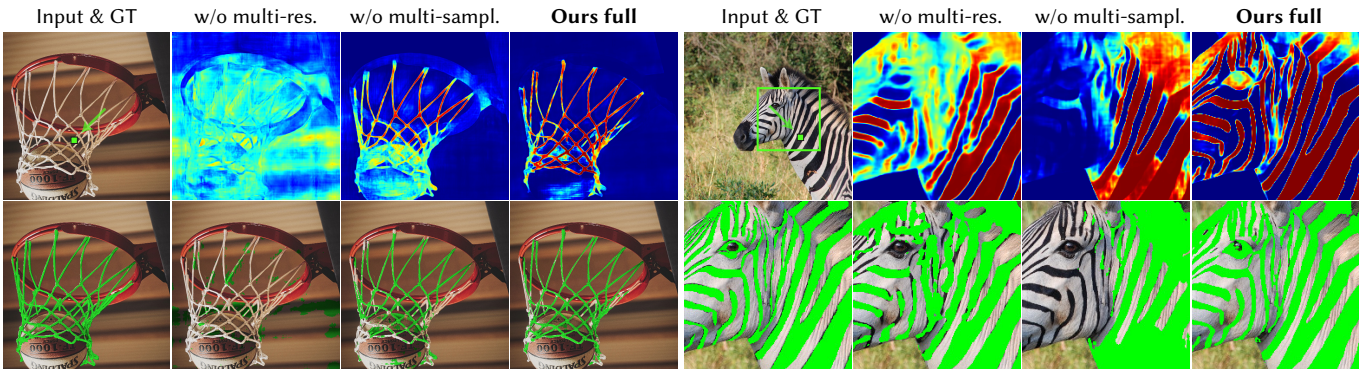

Fig. 8. **Ablations.** We ablate parts of our method (multi-resolution and multi-sampling, respectively) on two challenging examples containing thin structures (net and zebra stripes) and albedo entanglement (net has a similar albedo to the white parts of the basketball, in front of a patterned background).

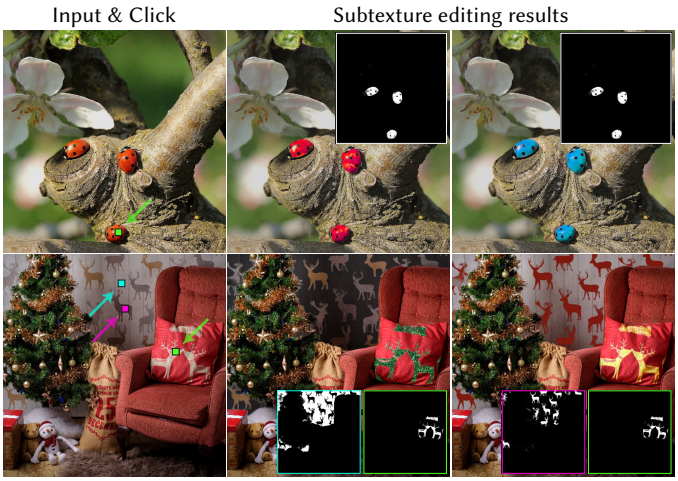

Fig. 9. **Editing.** We use our method's selection masks at subtexture level to perform fine-granular edits of the image's materials in Photoshop.

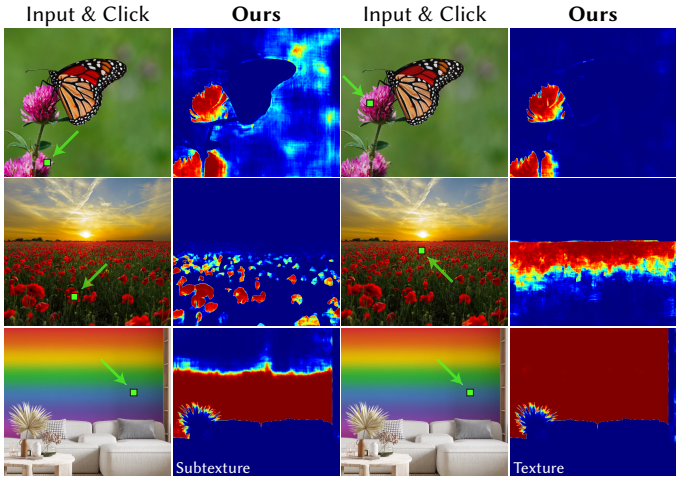

Fig. 10. **Limitations.** Our method struggles with clicks on out-of-focus regions (top) and long-horizon imagery with changing frequencies (middle). Textures without individual components (bottom) stretches our definition of subtexture.

qualitatively in the supplemental document. Then, we show results removing our multi-resolution and multi-sampling approaches, both in Table 3 and in Fig. 8. Our multi-resolution feature aggregation improves quality of the predictions (lower L1), which is particularly visible in the qualitative results, as it drastically improves performance when dealing with thin edges, like the basketball net. Our multi-sampling strategy, on the other hand, improves overall precision, in particular for texture-level selection. This may stem from computing gradients on the predicted selection for multiple materials in an image at a time, in every optimization step. Additionally, this multi-sampling significantly improves confidence by reducing the sensitivity to the selection threshold, which minimizes the need for manual adjustments (see results in the supplemental document).

We also assess in Table 3 the effect of jointly training both subtexture and texture levels in one model (Ours Full) compared to training two separate models with a single output, one per level (Ours, Single Level). Notably, training with all data concurrently does not negatively impact performance, allowing to have a single model for both selection levels, and supporting our hypothesis that jointly estimating both outputs is beneficial for accuracy.

Last, we further evaluate our ablations on a subset of 30 challenging test cases including fine structures, albedo entanglement, and strong light variations. We include this full challenging subset in the supplemental document, Figure 3. The results (Table 3, right-most columns) show the clear benefits of our multi-resolution component, which significantly improves performance on thin structures, as well as our multi-sampling strategy, which improves robustness overall and helps in scenarios with albedo entanglement and strong light variations (50% lower L1 and 20% higher IoU in Table 3). We hypothesize that by sampling multiple query pixels per image during training, it is more likely that these challenging cases (e.g., two pixels with same albedo from different materials, or two pixels from the same material in areas with strong lighting variation) appear in the same training batch, improving the gradient.

4.5 Application: Material Editing

Selection of specific regions within an image is an exceedingly common and highly useful tool for a number of applications, among which editing is one of the most representative examples. We show practical applications of our selection for material editing tasks in image space, in Fig. 1 (bottom right) and Fig. 9. Our new level of granularity at the subtexture level, as well as our improved precision and accuracy, allow users to easily edit challenging scenarios that would otherwise require significant manual intervention, including individual texture components such as the flowers of the dress (Fig. 1), or the deer on the cushion and wallpaper (Fig. 9).

5 DISCUSSION

Several avenues of future work remain open. Clicking on a pixel in an out-of-focus area of the image may lead to incorrect selections, as shown in the top row of Fig. 10: as it can be seen, our model cannot accurately distinguish the flower petals and activates inaccurate regions in the background (left). The problem mostly goes away when clicking on an in-focus area of such flowers (right). Images with long horizon lines may also be problematic, if the image exhibits strongly varying frequencies due to perspective. In the middle row of Fig. 10 the first ranks of flowers are selected but the more distant ones are not (left), or viceversa (right). We believe these limitations could be addressed by adding more training data with explicit depth of field effects and outdoor large scenes, respectively. Another idea to mitigate potential failure cases in practical scenarios could be to combine multiple selection masks, where a user could optionally select extra query pixels manually, combining the predictions and producing improved masks. This strategy could help in cases where the initial selection misses specific parts of the desired result (e.g., middle row in Fig. 10). Negative input queries, subtracting from the selection, would also be possible in cases where our selection overselects (e.g., top row, left in Fig. 10). Further, our definition of subtexture may not easily translate to continuously varying surfaces like the rainbow wall in the bottom row of Fig. 10. Future work could explore a more continuous similarity definition to enable a gradient in the similarity score, following that of the color distance on the wall. More generally, while we propose an additional selection level, we do not claim to have decisively solved the inherent ambiguity of material selection tasks, for which different definitions of what "the same material" means may be needed, depending on the intended goal and downstream applications.

6 CONCLUSION

We present a novel method for fine-grained material selection in images, which works both at texture and subtexture levels, and is more precise and robust than previous approaches. We evaluate various ViT backbones and propose a new training scheme and a large-scale dataset, significantly improving selection quality for fine structures and challenging scenarios with albedo entanglement and complex light variation.

ACKNOWLEDGMENTS

This work has been partially supported by grant PID2022-141539NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU. This work has also received funding from the Government of Aragon's Departamento de Educacion, Ciencia y Universidades through the Reference Research Group "Graphics and Imaging Lab" (ref T34_23R) and through the project "HUMAN-VR: Development of a Computational Model for Virtual Reality Perception" (PROY_T25_24). Julia Guerrero-Viu developed part of this work during an Adobe internship, and was also partially supported by the FPU20/02340 predoctoral grant. We thank the members of the Graphics and Imaging Lab, especially Nacho Moreno, Nestor Monzón, and Santiago Jiménez, for insightful discussions, help preparing the figures and final proofreading.

REFERENCES

2024. Evermotion Arch Interior. https://evermotion.org/shop/cat/397/archinteriors.

Miika Aittala, Tim Weyrich, Jaakko Lehtinen, et al. 2015. Two-shot SVBRDF capture for stationary materials. *ACM Trans. Graph.* 34, 4 (2015), 110–1.

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

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

Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. 1998. 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)*. IEEE, 675–682.

Yash Bhalgat, Iro Laina, Joao F Henriques, Andrew Zisserman, and Andrea Vedaldi. 2023. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. *arXiv preprint arXiv:2306.04633* (2023).

Daniel Bolya, Chaitanya Ryali, Judy Hoffman, and Christoph Feichtenhofer. 2023. Window Attention is Bugged: How not to Interpolate Position Embeddings. *arXiv preprint arXiv:2311.05613* (2023).

Hussein Boyaci, Lawrence T Maloney, and Sarah Hersh. 2003. The effect of perceived surface orientation on perceived surface albedo in binocularly viewed scenes. *Journal of vision* 3, 8 (2003), 2–2.

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

Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. 2014. Deep convolutional filter banks for texture recognition and segmentation. *arXiv preprint arXiv:1411.6836* (2014).

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

Sagi Eppel, Jolina Li, Manuel Drehwald, and Alan Aspuru-Guzik. 2024. Learning Zero-Shot Material States Segmentation, by Implanting Natural Image Patterns in Synthetic Data. *arXiv preprint arXiv:2403.03309* (2024).

Michael Fischer, Iliyan Georgiev, Thibault Groueix, Vladimir G Kim, Tobias Ritschel, and Valentin Deschaintre. 2024. SAMa: Material-aware 3D Selection and Segmentation. *arXiv preprint arXiv:2411.19322* (2024).

Roland W Fleming, Ron O Dror, and Edward H Adelson. 2003. Real-world illumination and the perception of surface reflectance properties. *Journal of vision* 3, 5 (2003), 3–3.

Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. 2022. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 1–11.

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

Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. 2024. Garfield: Group anything with radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21530–21539.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on*

Computer Vision. 4015–4026.

Peter Kocsis, Vincent Sitzmann, and Matthias Niessner. 2024. Intrinsic Image Diffusion for Indoor Single-view Material Estimation. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Manuel Lagunas, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. 2019. A similarity measure for material appearance. *ACM Trans. Graph.* 38, 4, Article 135 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323036

Hendrik PA Lensch, Jan Kautz, Michael Goesele, Wolfgang Heidrich, and Hans-Peter Seidel. 2003. Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics (TOG)* 22, 2 (2003), 234–257.

Thomas Leung and Jitendra Malik. 2001. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision* 43 (2001), 29–44.

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

Luca Morreale, Noam Aigerman, Vladimir G Kim, and Niloy J Mitra. 2024. Neural semantic surface maps. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15005.

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

Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. 2019b. A Multi-Illumination Dataset of Indoor Object Appearance. In *2019 IEEE International Conference on Computer Vision (ICCV)*.

Merlin Nimier-David, Zhao Dong, Wenzel Jakob, and Anton Kaplanyan. 2021. Material and Lighting Reconstruction for Complex Indoor Scenes with Texture-space Differentiable Rendering. In *Eurographics Symposium on Rendering - DL-only Track*, Adrien Bousseau and Morgan McGuire (Eds.). The Eurographics Association. https://doi.org/10.2312/sr.20211292

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal* (2024), 1–31.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

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

Lavanya Sharan. 2009. *The perception of material qualities in real-world images*. Ph. D. Dissertation. Massachusetts Institute of Technology.

Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H Adelson. 2013. Recognizing materials using perceptually inspired features. *International journal of computer vision* 103 (2013), 348–371.

Lavanya Sharan, Ruth Rosenholtz, and Edward H Adelson. 2014. Accuracy and speed of material categorization in real-world images. *Journal of vision* 14, 9 (2014), 12–12.

Prafull Sharma, Julien Philip, Michaël Gharbi, Bill Freeman, Fredo Durand, and Valentin Deschaintre. 2023. Materialistic: Selecting similar materials in images. *ACM Transactions on Graphics* 42, 4 (2023).

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When do we not need larger vision models?. In *European Conference on Computer Vision*. Springer, 444–462.

Borhan Uddin Sumon, Damien Muselet, Sixiang Xu, and Alain Trémeau. 2022. Multi-View Learning for Material Classification. *Journal of Imaging* 8, 7 (2022), 186.

Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. 2017. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*. IEEE, 537–547.

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

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

Giuseppe Vecchio and Valentin Deschaintre. 2024. MatSynth: A Modern PBR Materials Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22109–22118.

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. 2022. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5481–5490.

Ting-Chun Wang, Jun-Yan Zhu, Ebi Hiroaki, Manmohan Chandraker, Alexei A Efros, and Ravi Ramamoorthi. 2016. 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*. Springer, 121–138.

Jia Xue, Hang Zhang, Ko Nishino, and Kristin J Dana. 2020. Differential viewpoints for ground terrain material recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2020), 1205–1218.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024. Rgb-X: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. 2022. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2822–2831.