MaRI: Material Retrieval Integration across Domains

Jianhui Wang[1*] Zhifei Yang[2*] Yangfan He[3*] Huixiong Zhang[1] Yuxuan Chen[4] Jingwei Huang[5†]
[1]University of Electronic Science and Technology of China [2]Peking University
[3]University of Minnesota-Twin Cities [4]Fudan University [5]Tencent Hunyuan3D
https://jianhuiwemi.github.io/MaRI

Figure 1. Examples from the MaRI Gallery, showcasing (a) synthetic and (b) real-world datasets we constructed. (c) MaRI: A groundbreaking framework for accurately retrieving textures from images, bridging the gap between visual representations and material properties.

Abstract

Accurate material retrieval is critical for creating realistic 3D assets. Existing methods rely on datasets that capture shape-invariant and lighting-varied representations of materials, which are scarce and face challenges due to limited diversity and inadequate real-world generalization. Most current approaches adopt traditional image search techniques. They fall short in capturing the unique properties of material spaces, leading to suboptimal performance in retrieval tasks. Addressing these challenges, we introduce MaRI, a framework designed to bridge the feature space gap between synthetic and real-world materials. MaRI constructs a shared embedding space that harmonizes visual and material attributes through a contrastive learning strategy by jointly training an image and a material encoder, bringing similar materials and images closer while separating dissimilar pairs within the feature space. To support this, we construct a comprehensive dataset comprising high-quality synthetic materials rendered with controlled shape variations and diverse lighting conditions, along with real-world materials processed and standardized using material transfer techniques. Extensive experiments demonstrate the superior performance, accuracy, and generalization capabil-

[*]Equal contribution.
[†]Corresponding author.

ities of MaRI across diverse and complex material retrieval tasks, outperforming existing methods.

1. Introduction

The creation of realistic appearances is a crucial aspect of 3D asset generation, with accurate material reconstruction, particularly through high-quality materials in UV texture space, being key to achieving photorealism [15, 18, 35, 36, 40, 41]. This is especially important in applications like augmented reality (AR), virtual reality (VR), digital content creation, and industrial design, where the seamless integration of virtual objects into real-world environments depends on faithfully reproducing material properties [11, 22, 28, 31]. A fundamental challenge in these domains is aligning the information from the material space with that from the image space. The goal is to project both into a shared embedding space, enabling accurate comparison and retrieval of materials. Achieving this alignment is critical, as it allows for high-quality material searches that can accurately match visual inputs with corresponding material representations, leading to more realistic and context-aware renderings.

Material retrieval can theoretically be viewed as an image search problem, which suggests that the task should be relatively straightforward given the vast array of image search techniques available today, including vision transformers [12], DINOv2 [25] and multimodal approaches like CLIP [27] and GPT-4V [13]. While the image space has been extensively explored and understood, applying image-based methods directly to material search often falls short. For example, some recent efforts have attempted to adapt image search techniques for material retrieval but the results have been suboptimal [14, 39]. We argue that the problem is caused by inherent differences between the material and image. As a result, material search and image search differ fundamentally since it requires capture a feature space specifically for materials properties including texture, reflectance, and surface roughness. Unfortunately, such a well-defined feature space for materials is unavailable due to the lack of comprehensive datasets and effective material descriptors. The absence of a meaningful material embedding makes it challenging to achieve accurate retrieval, highlighting the need for learning a shared space that align material and image information for more effective searches.

To handle these challenges, we introduce MaRI—a novel framework inspired by CLIP that learns a joint embedding space for both materials and images. MaRI employs dual encoders, jointly trained to align material properties with visual features in a shared space, facilitating direct and efficient comparisons. By leveraging pre-trained DINOv2 models as the backbone for both the material and image encoders, MaRI preserves generalizability while fine-tuning only the final Transformer block to capture domain-specific nuances effectively. We construct a large-scale dataset pairing images with materials, designed to capture both diversity and realism, for training a joint embedding. Since such a dataset is unavailable, we adopt both synthesis and generative approach to automatically construct the dataset. During synthesis, we construct each data pair by sampling a material from a material gallery, associating it to an object from an object dataset, and render it with Blender to obtain an image that pairs with the material. Although synthetic data alone yields satisfactory results for training, it still partially falls short in bridging the domain gap, as it cannot fully represent the diverse and nuanced appearances of real-world materials. To complement this, we introduce a generative approach that incorporate large-scale, unlabeled real-world image data and construct paired material with a material transfer technique (ZeST) [7]. As a result, the generative approach captures diverse material appearances under varying conditions through real images. It enables us to adopt a self-supervised learning framework, helping the model to learn robust material representations without being dependent on annotated datasets.

The effectiveness of MaRI is validated through a series of experiments. These evaluations focus on two distinct datasets: one emphasizing retrieval within a gallery of synthetic materials the model was trained on, and another assessing generalization to unseen materials. The results demonstrate MaRI's ability to bridge the domain gap between synthetic and real-world data, achieving significant improvements in both instance-level and class-level retrieval tasks. Our main contributions are summarized as:

- We propose MaRI, a framework that learns a joint embedding space for visual and material properties, providing a new approach to material retrieval by aligning visual features with material characteristics.
- We construct a diverse dataset that spans various material types and conditions, supporting both synthetic and real-world material retrieval.
- We show the ability of MaRI's retrieval pipeline to accurately scale across a wider variety of materials, achieving precise retrieval for complex and diverse material types.

2. Related Work

Datasets for Material Understanding. Over the years, various datasets have played a crucial role in advancing material understanding. Early works laid the foundation by capturing real-world material reflectance properties, providing basic 3D models, generating synthetic photorealistic images with ground truth annotations, and conducting similarity assessments for 3D reconstruction and material synthesis [2, 3, 5, 21, 26]. For instance, Flickr Material Database [23] contributed to material recognition by providing labeled images of real-world materials for classification tasks. Datasets like the Amazon-Berkeley Objects (ABO) [8] with high-resolution 3D models and PBR materials, and MatSynth [34] and OmniObject3D [38] with thousands of PBR mate-

rials and real-scanned objects, have significantly enhanced material diversity and realism.

Despite these advancements, challenges remain in achieving sufficient material diversity and integrating real-world and synthetic data. Existing datasets and methods have inspired us to construct a more diverse dataset, incorporating richer environmental contexts, complex shapes, and more detailed material information. AmbientCG [1] provides a rich library of high-quality, PBR materials, designed for use in physically-based rendering workflows, while Objaverse [10] offers an extensive collection of 3D models for material application and visual tasks across synthetic environments. Additionally, the ZeST [7] method's material transfer approach informs our efforts to capture diverse appearances under varying conditions.

Material Generation and Retrieval. Advances in material generation and retrieval have enhanced the realism of 3D content creation. Techniques like ControlMat [35] use diffusion models to generate high-resolution material maps, allowing precise control over surface properties such as roughness and normal maps. MatFuse [36] enables users to guide the material generation process through sketches, color palettes, or text prompts, providing greater flexibility in design. Similarly, Fantasia3D [6] employs pixel-level optimization for detailed material representations, though it faces challenges in maintaining stability with complex geometries. Make-it-3D [32] uses a two-stage diffusion process to generate high-fidelity 3D content from single images, emphasizing texture refinement for realistic outputs. Tools like Material Palette [24] and Matlas [4] further advance material manipulation by offering more control over varied environmental conditions and geometries.

Several studies [20, 30] have focused on perceptual similarity. However, none directly address the challenge of retrieving accurate materials from real-world images. MaPa [39] and Make-it-Real [14] integrate material retrieval within broader 3D generation workflows. MaPa utilizes GPT-4V [13] for initial material categorization and CLIP [27] for refining the material assignment process, retrieving material graphs from predefined libraries, though it often lacks precision in handling fine textures and complex surfaces. Make-it-Real leverages GPT-4V to assign materials to segmented 3D models using SVBRDF [16] mappings, but its reliance on pre-annotated datasets limits adaptability to novel materials and environments. These limitations underscore the need for more robust frameworks. Our work addresses this by bridging the gap between real-world and synthetic data, capturing both fine textures and complex material details.

3. Methodology

In this section, we introduce MaRI, a framework developed to address the domain gap in material retrieval between synthetic and real-world data. To clarify, we distinguish "image" as a 2D perspective view and "material" as a material ball representation (see Figure1 (b)), using these terms as abbreviations throughout. MaRI aligns image and material features within a shared feature space $\mathcal{F}$, allowing for direct comparisons across different domains. To achieve this alignment, we construct a comprehensive dataset $\mathcal{D} = \{(\mathcal{D}_{\text{synthetic}}, \mathcal{D}_{\text{real}})\}$ with pairs of image and material that combines controlled synthetic samples with diverse real-world materials. Inspired by CLIP, MaRI uses a dual-encoder architecture based on DINOv2, with separate encoders fine-tuned for image and material representations. The focus is on adapting the last Transformer block to retain general visual features while learning domain-specific variations. A contrastive loss guides the training, pulling matched pairs closer in $\mathcal{F}$ and pushing apart mismatched pairs, making MaRI effective in retrieving materials across both synthetic and real-world settings.

3.1. Problem Formulation

Material retrieval involves mapping visual representations and material properties into a shared feature space $\mathcal{F}$ to enable direct comparison and accurate retrieval. This is achieved through two encoders: E_I for image space $\mathcal{X}$ and E_M for material space $\mathcal{M}$. Given an image $x \in \mathcal{X}$ and a material $m \in \mathcal{M}$, the encoders map these inputs into $\mathcal{F}$ as $\mathbf{z}_I = E_I(x)$ and $\mathbf{z}_M = E_M(m)$, where $\mathbf{z}_I, \mathbf{z}_M \in \mathcal{F}$ represent the feature embeddings of the image and material, respectively. The mapping facilitates direct comparison of visual features and material attributes within the joint space, making it possible to measure their similarity.

Aligning similar images and materials in the shared space relies on a contrastive learning framework, trained on our constructed dataset $\mathcal{D} = \{(\mathcal{D}_{\text{synthetic}}, \mathcal{D}_{\text{real}})\}$. The dataset, comprising both synthetic and real-world material samples, provides the model with a diverse range of materials, supporting improved generalization. The objective is to maximize $\text{sim}(\mathbf{z}_I, \mathbf{z}_M)$ for positive pairs of image embeddings $\mathbf{z}_I$ and material embeddings $\mathbf{z}_M$, while minimizing $\text{sim}(\mathbf{z}_I, \mathbf{z}_{M'})$ for negative pairs $\mathbf{z}_{M'}$. The shared feature space $\mathcal{F}$ provides a structure for formulating the material retrieval task as a nearest-neighbor search. For a query image x_q, the objective is to find the material m^* that maximizes the similarity with the query's feature embedding:

$$m^* = \arg\max_{m \in \mathcal{M}} \text{sim}(\mathbf{z}_{I_q}, \mathbf{z}_M). \tag{1}$$

Through the training of the encoders E_I and E_M, the representations in $\mathcal{F}$ are aligned, supporting accurate and efficient material retrieval.

3.2. Dataset

Addressing the domain gap between synthetic and real-world materials requires a large-scale dataset that captures a broad range of material types and environmental conditions. To

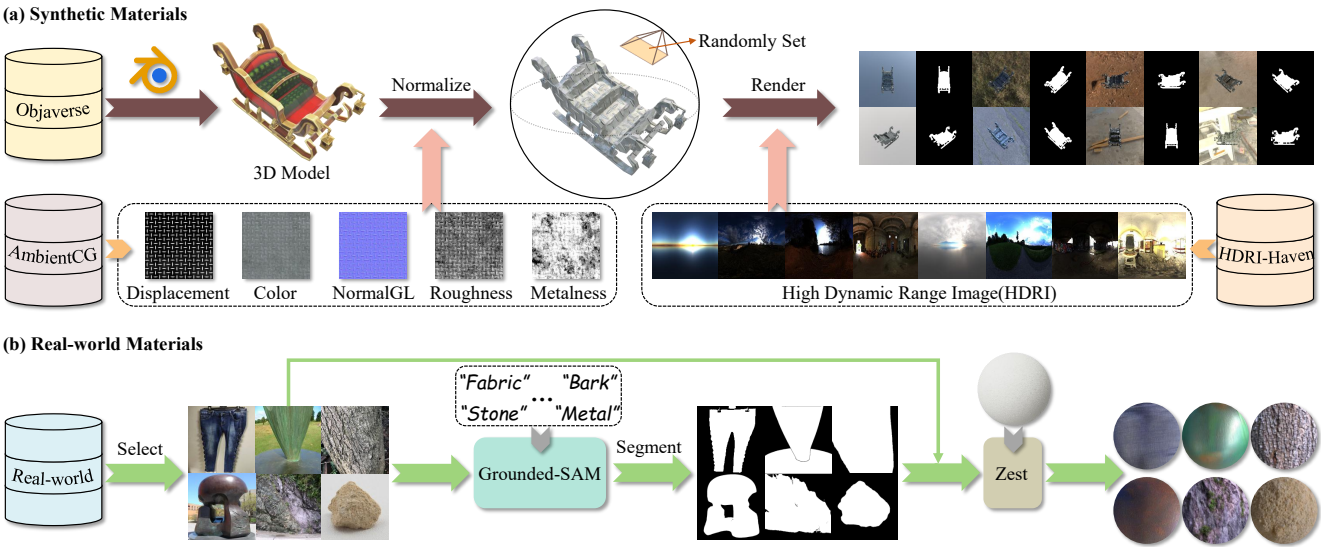

Figure 2. Overview of our dataset construction pipeline. (a) Synthetic materials are generated from 3D models obtained from Objaverse, combined with textures from AmbientCG, and rendered with HDR images. (b) Real-world materials are selected and segmented using Grounded-SAM and then transformed into material spheres via the ZeST method.

achieve this, we construct a dataset $\mathcal{D} = \{(\mathcal{D}_{\text{synthetic}}, \mathcal{D}_{\text{real}})\}$ to offer a rich training resource for material retrieval. The dataset is composed of two main components: $\mathcal{D}_{\text{synthetic}}$ and $\mathcal{D}_{\text{real}}$, each designed to cover different aspects of material appearance and variability, contributing to a more effective alignment between synthetic control and real-world complexity, as illustrated in Figure 2.

3.2.1. Synthetic Materials

We aim to create a synthetic dataset by rendering various objects associated with diverse materials in different lighting environments in Blender. The process utilizes 3D models O_i from Objaverse [10] and applies a systematic normalization process to ensure consistency across various rendering conditions. Let $\mathbf{B}_{\min} = (b_x^{\min}, b_y^{\min}, b_z^{\min})$ and $\mathbf{B}_{\max} = (b_x^{\max}, b_y^{\max}, b_z^{\max})$ denote the minimum and maximum coordinates of the model's axis-aligned bounding box, with $\mathbf{s} = \mathbf{B}_{\max} - \mathbf{B}_{\min}$ representing its spatial extent. We define the scaling factor α as $\alpha = 1/\max(s_x, s_y, s_z)$, where s_x, s_y, s_z are the dimensions of $\mathbf{s}$. To center the model at the origin, we compute the centroid $\mathbf{c} = (\mathbf{B}_{\max} + \mathbf{B}_{\min})/2$. The normalized model $\mathbf{O}_i'$ is then obtained as:

$$\mathbf{O}_i' = \frac{1}{\max(s_x, s_y, s_z)} \left(O_i - \frac{\mathbf{B}_{\max} + \mathbf{B}_{\min}}{2} \right). \quad (2)$$

The transformation scales the model to fit within a unit cube and centers it at the origin, providing uniformity across models and facilitating consistent rendering conditions.

Next, we apply materials $m_j \in \mathcal{M}$ to each model. These materials are drawn from a library of 1605 physically-based rendering (PBR) textures sourced from AmbientCG [1], covering 86 distinct material categories. The library is designed to encompass most material types. Each material includes Base Color, Normal Map, Roughness, Displacement maps, and Metalness(optional) which are integrated into a principled BSDF shader to simulate realistic surface interactions with light. The materials are represented as tuples of physical properties, capturing diverse physical properties crucial for accurate material representation.

Lighting conditions are varied using 712 HDRI files H_k sourced from HDRI Haven [17], simulating diverse real-world lighting scenarios. Cameras C_l are randomly positioned on a hemispherical surface of radius $r = 3$ units around each model, with latitude θ and longitude ϕ angles sampled as $\theta \in [5°, 75°]$ and $\phi \in [-180°, 180°]$. The upper hemisphere placement reduces shadows, while the random positions increase shape variance by capturing multiple perspectives. Each camera's position is calculated by:

$$(x, y, z) = r(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta). \quad (3)$$

For each model and material combination, we generate 8 different viewpoints using randomly positioned cameras, with lighting conditions also randomized through different HDRI environments. Each rendering generates an image x_i, its corresponding mask mask_i, and the applied material descriptor m_i, where the mask delineates the object's shape within the image. The complete synthetic dataset is defined as: $\mathcal{D}_{\text{synthetic}} = \{(x_i, \text{mask}_i, m_i)\}_{i=1}^{N_{\text{synthetic}}}$, where $N_{\text{synthetic}} = 394560$ represents the total number of samples. The dataset covers a wide range of shapes, textures, and lighting conditions, offering a diverse and controlled resource for the material retrieval task.

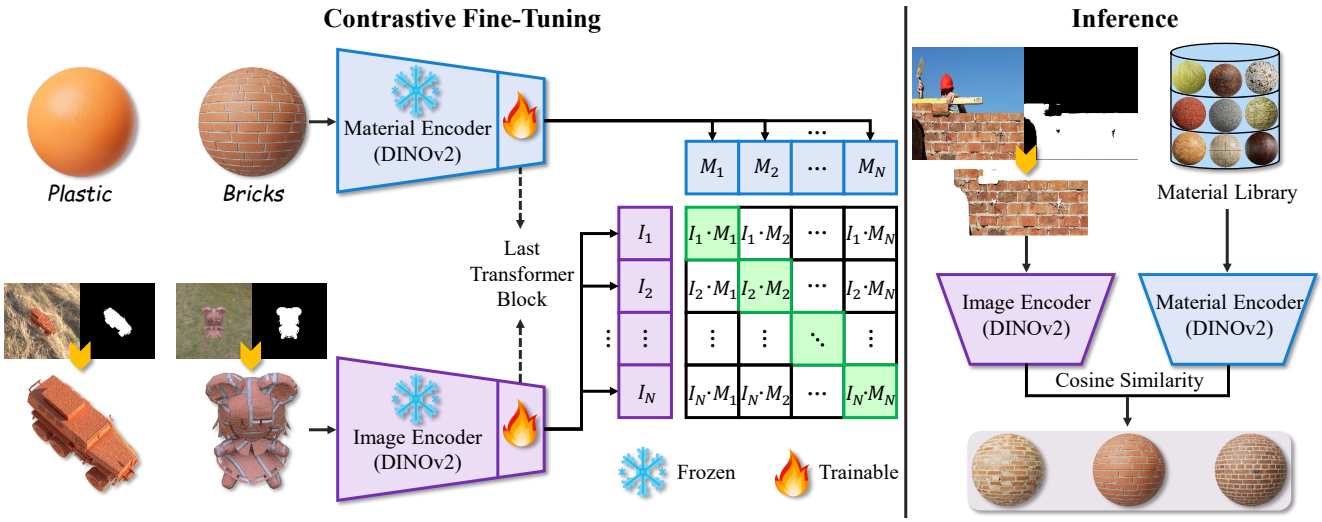

Figure 3. The architecture of the MaRI framework for contrastive fine-tuning in material retrieval. MaRI uses DINOv2-based encoders for both image and material feature extraction, fine-tuning only the last Transformer block, while keeping the rest of the model frozen. During inference, cosine similarity between image and material embeddings is used to retrieve the most relevant materials from the library.

3.2.2. Real-world Materials

Blender-based synthetic rendering produces high-quality, diverse material samples, yet occasionally encounters a domain gap when applied to real-world material retrieval tasks. Additionally, even though the synthetic data covers a vast majority of material types, the sheer diversity of materials in the real world means that many are still underrepresented.

Inspired by the ZeST [7] method's ability to transfer material appearances from real-world images onto neutral material spheres, we expanded our dataset to include a wider variety of real-world materials. We first curated a dataset comprising thousands of real-world images, collected from online sources and various datasets [9, 19, 23, 37]. Priority was then given to images with clearly identifiable foreground objects, which were segmented using the Grounded SAM model [29] with material-specific prompts to produce accurate object masks. Each image, along with its segmentation mask, was processed through the ZeST pipeline to generate a standardized material representation on a neutral sphere.

The real-world materials dataset also stores three components for each sample: the original image x_i, the segmentation mask mask_i, and the rendered material sphere m_i. This results in a structured dataset: $\mathcal{D}_{\text{real}} = \{(x_i, \text{mask}_i, m_i)\}_{i=1}^{N_{\text{real}}}$, where $N_{\text{real}} = 30,000$. The dataset mainly covers 8 material categories, such as metals, fabrics, woods, and ceramics. By integrating these real-world samples, our dataset can effectively reduce the domain gap between synthetic and real-world materials.

3.3. Domain-Adaptive Contrastive Learning

Building on the diverse range of material data in $\mathcal{D}$, our proposed MaRI framework is inspired by the contrastive learning approach of CLIP, but adapts it to bridge the domain gap between synthetic and real-world visual representations—a key challenge in material retrieval tasks. Rather than aligning different modalities, MaRI focuses on aligning varied visual features within a single modality but across different domains. It utilizes two DINOv2-based encoders, E_I for masked image representations and E_M for material properties, to project masked image inputs x_i and material spheres m_i into a shared feature space $\mathcal{F}$:

$$\mathbf{z}_I^i = E_I(x_i \odot \text{mask}_i), \quad \mathbf{z}_M^i = E_M(m_i). \tag{4}$$

Here, $\mathbf{z}_I^i, \mathbf{z}_M^i \in \mathbb{R}^d$ are the embeddings of the masked rendered image and the material sphere image, and $\odot$ denotes element-wise multiplication to apply the mask.

We fine-tune only the last Transformer block of each encoder, allowing the model to capture domain-specific variations in materials while retaining the general pre-trained features of DINOv2. As shown in Figure 3, the similarity between the image and material embeddings is computed using a scaled dot-product function:

$$\text{sim}(\mathbf{z}_I^i, \mathbf{z}_M^j) = \frac{\mathbf{z}_I^i \cdot \mathbf{z}_M^j}{\sqrt{d}}, \tag{5}$$

where d is the dimensionality of the feature space. We then use the InfoNCE loss with a temperature parameter $\tau = 0.07$, which controls the sharpness of the resulting distribution, to pull the representations of matching pairs closer while pushing non-matching pairs apart in the shared feature space:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\big(\text{sim}(\mathbf{z}_I^i, \mathbf{z}_M^i)/\tau\big)}{\sum_{j=1}^{N} \exp\big(\text{sim}(\mathbf{z}_I^i, \mathbf{z}_M^j)/\tau\big)}. \tag{6}$$

The loss encourages positive pairs $(\mathbf{z}_I^i, \mathbf{z}_M^i)$ to have higher similarity scores than any other pair $(\mathbf{z}_I^i, \mathbf{z}_M^j)$ with $j \neq i$, thus aligning the features in a domain-agnostic manner. MaRI effectively creates a shared space that supports robust material retrieval across varying data sources.

4. Experiments

In this section, we conduct a comprehensive evaluation of the MaRI framework for material retrieval tasks. Section 4.1 introduces the test datasets and evaluation metrics employed in our experiments, providing context for the subsequent analyses. Section 4.2 presents a comparative analysis, where MaRI's performance is benchmarked against existing models commonly used for material search. We also explore, in Section 4.3, how key design elements like model architecture, training strategies, and data composition influence MaRI's results. Finally, Section 4.4 provides additional qualitative results, demonstrating MaRI's capacity to retrieve accurate matches from the Unseen Materials dataset.

4.1. Test Datasets and Metrics

To evaluate the effectiveness of MaRI, we design two distinct test datasets and corresponding evaluation protocols to assess material retrieval performance.

The first test dataset, marked as "Trained" in all experiments, evaluates material retrieval performance using novel test images from a material gallery derived from AmbientCG. Specifically, we selected approximately 200 materials from the synthetic dataset, spanning mainly eight categories: wood, metal, plastic, leather, fabric, stone, ceramic, and rubber. For each selected material, corresponding real-world images were collected online, and material-specific regions were annotated with segmentation masks. Retrieval was performed using these real-world images as queries, with the target being the original 200 materials in the gallery of 1605 synthetic materials. We report the following metrics: (1) Top-1 and Top-5 instance-level accuracy(T1I and T5I), which measure the ability of the model to precisely retrieve the correct material from the gallery; (2) Top-1 class-level accuracy(T1C), assessing how accurately the model classifies materials into predefined categories; and (3) Top-3 Intersection-over-Union (T3IoU), which quantifies the overlap between the predicted material categories and the ground truth, providing a measure of category-level alignment. The use of Top-1 and Top-3 metrics is driven by the inherent scarcity of certain materials in the gallery, as higher Top-k metrics may include irrelevant matches due to the limited number of materials closely resembling the target.

The second test dataset is marked as "Unseen" in all experiments, which evaluates generalization to novel test materials by utilizing a new gallery of approximately 200 materials curated from the Textures website [33], which were unseen during the training process. Similar to the previous setup, for each material, real-world images and their corresponding material segmentation masks were collected, and retrieval was conducted within this newly constructed gallery. This scenario evaluates MaRI's ability to retrieve accurate matches for real-world queries from a gallery of previously unseen materials. Due to the diverse and unstructured nature of the material categories in this dataset, the evaluation emphasizes Top-1 and Top-5 instance-level accuracy, showcasing the framework's capability to identify the most relevant matches.

4.2. Comparative Analysis of Material Retrieval

Given the novelty of the material retrieval task, there are currently no directly comparable methods designed specifically for this purpose. Our comparisons therefore draw on related approaches, including general-purpose image search models like ViT, CLIP, and DINOv2, which serve as baselines for instance- and class-level retrieval evaluations. Additionally, we compare MaRI against existing material retrieval approaches used in other works. Make-it-Real leverages GPT-4V to perform hierarchical material searches, starting with high-level category identification followed by detailed matching within a structured material library. MaPa integrates GPT-4V with CLIP by first performing coarse material categorization using GPT-4V and then refining the search within the predicted category using CLIP for detailed material retrieval.

Table 1. Material Retrieval Performance on Trained and Unseen Datasets. Best values are highlighted in **blue**.

Method	Trained				Unseen	
	T1I	T5I	T1C	T3IoU	T1I	T5I
ViT [12]	3.5%	12.0%	16.0%	0.41	16.5%	56.0%
DINOv2 [25]	7.5%	28.0%	69.0%	0.67	31.0%	62.5%
CLIP [27]	2.0%	11.0%	36.5%	0.47	14.0%	29.5%
Make-it-Real [14]	8.5%	16.0%	76.5%	0.60	42.5%	75.0%
MaPa [39]	2.5%	17.5%	80.0%	**0.80**	19.5%	69.0%
MaRI	**26.0%**	**90.0%**	**81.5%**	0.77	**54.0%**	**89.0%**

Table 1 highlights the material retrieval performance across both known and unseen galleries. In the Trained Materials, MaRI achieves Top-1 instance accuracy of 26.0%, Top-5 instance accuracy of 90.0%, and Top-1 class accuracy of 81.5%, outperforming all other methods. Although MaRI's Top-3 IoU score of 0.77 is slightly lower than MaPa's 0.80, this difference arises from the distinct retrieval processes employed by the two methods. MaPa utilizes GPT-4V for coarse-grained classification into material categories before conducting a fine-grained search within the same category. As a result, its IoU scores remain consistent across Top-k predictions and directly align with its Top-1 class accuracy. In contrast, MaRI performs retrieval directly within a unified embedding space, enabling the discovery of visually similar materials that may belong to different categories. Although

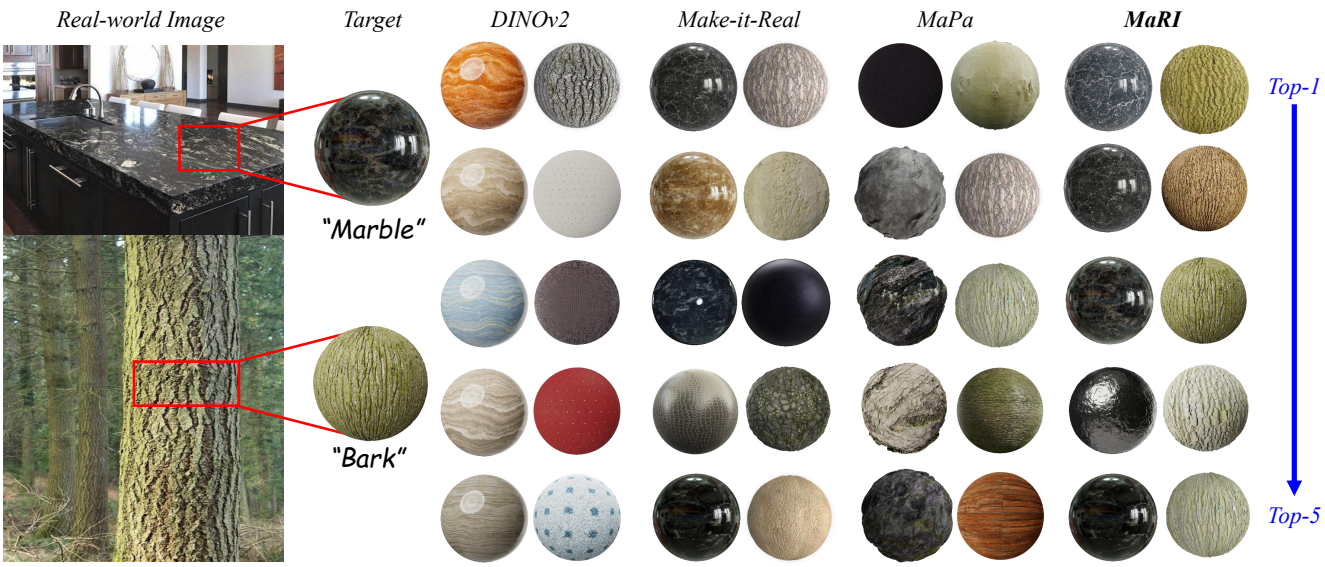

Figure 4. Qualitative comparison of material retrieval results using the Trained Materials dataset as the gallery.

the expanded search scope can occasionally result in category mismatches among the Top-k results, causing a minor decrease in Intersection over Union (IoU) compared to MaPa, it offers substantial benefits when it comes to retrieving materials at the instance level, outperforming all other methods in this regard. In the unseen materials gallery, MaRI also achieves significant improvements, with Top-1 and Top-5 instance accuracies of 54.0% and 89.0%, respectively, far surpassing Make-it-Real (42.5% and 75.0%). These results showcase MaRI's superior performance across both known and unseen material retrieval tasks.

The qualitative results in Figure 4 illustrate that general image search models, such as the original DINOv2, struggle to capture the intricate relationships between material textures and their corresponding images. As previously demonstrated through quantitative evaluations, methods like CLIP and ViT exhibit similarly poor performance and are therefore omitted from this figure. In contrast, MaRI also outperforms GPT-4V-based material retrieval methods, including MaPa and Make-it-Real, by more effectively capturing fine-grained material characteristics. For instance, in the "Bark" case shown in Figure 4, MaRI consistently retrieves materials in its Top-5 predictions that closely resemble the target in both texture and color.

4.3. Ablation Studies

Impact of Synthetic Data Scale. The complexity of texture and material information necessitates a large dataset for contrastive learning to capture material features effectively and enable accurate retrieval within the embedding space. We conducted an ablation study to evaluate the impact of synthetic data scale on MaRI's performance, as detailed in Table 2. The findings in Table 2 illustrate a strong correlation between the scale of synthetic data and the improvement in instance-level retrieval accuracy. For the Trained Materials dataset, Top-1 instance accuracy increases from 19.5% with 25% of the data to 26.0% with the full dataset, and Top-5 instance accuracy sees a substantial rise from 55.5% to 90.0%. Similarly, in the Unseen Materials dataset, Top-1 instance accuracy improves from 44.5% to 54.0%, showcasing MaRI's enhanced capability to generalize to previously unseen materials. Interestingly, while instance-level retrieval improves consistently with data scale, the Top-1 class-level accuracy exhibits relatively smaller gains, peaking at 82.0% for 50% of the dataset and slightly decreasing to 81.5% with the full dataset. The plateau suggests that class-level classification may benefit less from additional synthetic data due to the saturation of categorical information in the dataset. The observations highlight how a larger and more diverse dataset contributes to overall performance improvements, especially in instance-level retrieval.

Table 2. Ablation study evaluating the impact of synthetic data usage. Best values are highlighted in **blue**.

Data %	Trained				Unseen	
	T1I	T5I	T1C	T3IoU	T1I	T5I
25%	19.5%	55.5%	77.5%	0.76	44.5%	83.5%
50%	20.0%	63.5%	**82.0%**	**0.79**	46.0%	85.5%
75%	22.0%	79.5%	80.5%	0.78	48.5%	80.0%
100%	**26.0%**	**90.0%**	81.5%	0.77	**54.0%**	**89.0%**

Model Architecture and Data Composition. Building on the findings regarding the significance of synthetic data scale, we further analyze the contributions of key architectural components and data composition in optimizing MaRI's performance. As demonstrated in Table 3, the combination of dual encoders with both synthetic and real-world data

Table 3. Ablation study on model architecture and data composition. ✓ indicates the component is enabled, while ✗ indicates it is disabled. Best values are highlighted in **blue**. Abbreviations: DE = Dual Encoder, RD = Real Data, SD = Synthetic Data.

Configuration			Trained				Unseen	
DE	RD	SD	T1I	T5I	T1C	T3IoU	T1I	T5I
✓	✗	✓	20.5%	62.0%	75.5%	0.76	44.0%	78.0%
✓	✓	✗	9.0%	27.5%	45.0%	0.49	35.0%	63.5%
✗	✓	✓	20.5%	61.0%	77.5%	0.74	49.5%	85.5%
✓	✓	✓	**26.0%**	**90.0%**	**81.5%**	**0.77**	**54.0%**	**89.0%**

achieves the highest retrieval accuracies. The results validate the theoretical premise that the material and image spaces represent distinct domains, and employing dual encoders effectively reduces the domain gap by learning separate representations for each space while aligning them in the shared embedding space. Training with both synthetic and real data outperforms using either dataset alone. For instance, in the Trained Materials dataset, the combination of synthetic and real data achieves the highest Top-1 instance accuracy of 26.0%. Removing synthetic or real data significantly reduces instance-level accuracy, with Top-5 instance accuracy dropping from 90.0% to 62.0% or 27.5%, respectively. Additionally, the Unseen Materials dataset further underscores the importance of real data in enhancing generalization. Training with both datasets yields a Top-1 instance accuracy of 54.0%, compared to 44.0% or 35.0% when excluding synthetic or real data. Leveraging both datasets allows MaRI to establish stronger relationships between material and image spaces, resulting in robust performance across diverse and previously unseen materials.

Fine-Tuning Configurations. An analysis of Table 4 reveals that fine-tuning only the final Transformer block of DINOv2, while freezing other parameters, consistently yields better results across both InfoNCE and Triplet loss configurations. This is because the early layers of the pre-trained DINOv2 model capture generalizable low-level and mid-level features critical for material and image representations. Freezing these layers prevents overfitting to the training dataset and retains the model's ability to generalize across diverse material distributions. Under the same configurations, InfoNCE loss demonstrates superior performance over Triplet loss due to its ability to optimize a batch-wise similarity matrix, which evaluates all material-image pairs simultaneously. It allows the model to capture more nuanced relationships within the embedding space, effectively aligning material and image features. In contrast, Triplet loss focuses on individual anchor-positive-negative triplets, which limits its capacity to fully explore the complex associations.

4.4. More Qualitative Results

The retrieval results in Figure 5 highlight MaRI's effectiveness in identifying visually similar materials from the Unseen Materials dataset. On the left side, each example

Table 4. Ablation study on fine-tuning configurations. ✓ indicates the component is enabled, while ✗ indicates it is disabled. Best values are highlighted in **blue**. Abbreviations: LBO = Last Block Only, TL = Triplet Loss, IL = InfoNCE Loss.

Configuration			Trained				Unseen	
LBO	TL	IL	T1I	T5I	T1C	T3IoU	T1I	T5I
✗	✗	✓	13.0%	42.5%	69.0%	0.65	31.5%	67.0%
✗	✓	✗	7.5%	21.0%	53.0%	0.49	15.5%	52.5%
✓	✓	✗	5.5%	31.5%	73.0%	0.71	38.5%	71.5%
✓	✗	✓	**26.0%**	**90.0%**	**81.5%**	**0.77**	**54.0%**	**89.0%**

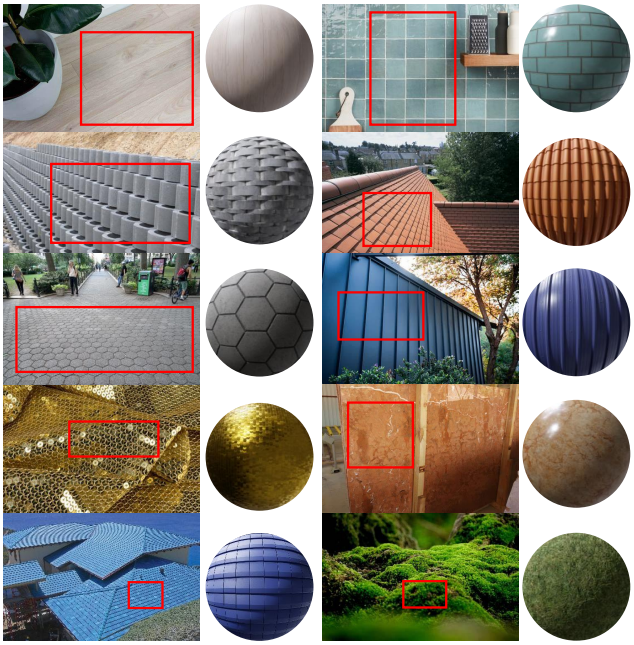

Figure 5. Examples of Top-1 material retrieval results using the Unseen Materials gallery as the search space.

presents a real-world image, with the corresponding material sphere retrieved from the unseen gallery shown on the right. MaRI successfully captures fine-grained details, such as the textured surface of tiles, the ruggedness of brick patterns, and the organic structure of moss. These examples demonstrate MaRI's strong generalization across diverse material.

5. Conclusion

We introduce MaRI, a novel framework designed specifically to address the challenges of material retrieval by aligning image and material features in a shared embedding space. A key component of MaRI is the construction of a comprehensive dataset, integrating both synthetic and real-world materials to effectively bridge the domain gap. MaRI successfully captures essential material properties and achieves strong generalization to unseen materials. Unlike prior methods, MaRI provides an effective framework for tackling the unique challenges of material retrieval, achieving strong performance in diverse scenarios.

References

[1] AmbientCG. AmbientCG, 2024. https://www.ambientcg.com/. 3, 4

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

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *CVPR*, pages 3479–3487, 2015. 2

[4] Duygu Ceylan, Valentin Deschaintre, Thibault Groueix, Rosalie Martin, Chun-Hao Huang, Romain Rouffet, Vladimir Kim, and Gaëtan Lassagne. Matatlas: Text-driven consistent geometry texturing and material assignment, 2024. 3

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015. 2

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation, 2023. 3

[7] Ta-Ying Cheng, Prafull Sharma, Andrew Markham, Niki Trigoni, and Varun Jampani. Zest: Zero-shot material transfer from a single image, 2024. 2, 3, 5

[8] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding, 2022. 2

[9] CV Mart. Cv mart datasets, 2024. Accessed: 2024. 5

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects, 2022. 3, 4

[11] Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3d shape and svbrdf acquisition, 2021. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2, 6

[13] OpenAI et al. Gpt-4 technical report, 2024. 2, 3

[14] Ye Fang, Zeyi Sun, Tong Wu, Jiaqi Wang, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. Make-it-real: Unleashing large multimodal model for painting 3d objects with realistic materials, 2024. 2, 3, 6

[15] Paul Guerrero, Miloš Hašan, Kalyan Sunkavalli, Radomír Měch, Tamy Boubekeur, and Niloy J Mitra. Matformer: A generative model for procedural materials. *arXiv preprint arXiv:2207.01044*, 2022. 2

[16] Michal Haindl, Jiří Filip, Michal Haindl, and Jiří Filip. Spatially varying bidirectional reflectance distribution functions. *Visual Texture: Accurate Material Appearance Measurement, Representation and Modeling*, pages 119–145, 2013. 3

[17] HDRI Haven. HDRI Haven - Free High-Quality HDRIs. https://hdri-haven.com/, 2024. Accessed: 2024. 4

[18] Yiwei Hu, Paul Guerrero, Milos Hasan, Holly Rushmeier, and Valentin Deschaintre. Generating procedural materials from text or image prompts. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[19] Kaggle. Kaggle datasets, 2024. Accessed: 2024. 5

[20] Manuel Lagunas, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. A similarity measure for material appearance. *ACM Trans. Graph.*, 38(4), 2019. 3

[21] Andreas Ley, Ronny Hänsch, and O. Hellwich. Syb3r: A realistic synthetic benchmark for 3d reconstruction from images. In *European Conference on Computer Vision*, 2016. 2

[22] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. In *SIGGRAPH Asia 2018 Technical Papers*, page 269. ACM, 2018. 2

[23] Ce Liu, Lavanya Sharan, Edward H. Adelson, and Ruth Rosenholtz. Exploring features in a bayesian framework for material recognition. In *CVPR*, pages 239–246, 2010. 2, 5

[24] Ivan Lopes, Fabio Pizzati, and Raoul de Charette. Material palette: Extraction of materials from a single image, 2023. 3

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2, 6

[26] Keunhong Park, Konstantinos Rematas, Ali Farhadi, and Steven M. Seitz. Photoshape: photorealistic materials for large-scale shape collections. *ACM Transactions on Graphics*, 37(6):1–12, 2018. 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3, 6

[28] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation, 2023. 2

[29] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 5

[30] Ana Serrano, Bin Chen, Chao Wang, Michal Piovarči, Hans-Peter Seidel, Piotr Didyk, and Karol Myszkowski. The effect of shape and illumination on material perception: model and applications. *ACM Trans. Graph.*, 40(4), 2021. 3

[31] Prafull Sharma, Julien Philip, Michaël Gharbi, William T. Freeman, Fredo Durand, and Valentin Deschaintre. Materialistic: Selecting similar materials in images, 2023. 2

[32] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior, 2023. 3

[33] Textures.com. Textures.com - Free Textures, Photos, and Background Images, 2024. [Accessed: 2024]. 6

[34] Giuseppe Vecchio and Valentin Deschaintre. Matsynth: A modern pbr materials dataset, 2024. 2

[35] Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. Controlmat: A controlled generative approach to material capture. *ACM Transactions on Graphics*, 43(5): 1–17, 2024. 2, 3

[36] Giuseppe Vecchio, Renato Sortino, Simone Palazzo, and Concetto Spampinato. Matfuse: controllable material generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4429–4438, 2024. 2, 3

[37] Visual Geometry Group, University of Oxford. Sculptures6k dataset, 2024. Accessed: 2024. 5

[38] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation, 2023. 2

[39] Shangzhan Zhang, Sida Peng, Tao Xu, Yuanbo Yang, Tianrun Chen, Nan Xue, Yujun Shen, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. Mapa: Text-driven photorealistic material painting for 3d shapes, 2024. 2, 3, 6

[40] Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. Tilegen: Tileable, controllable material generation and capture. In *SIGGRAPH Asia 2022 conference papers*, pages 1–9, 2022. 2

[41] Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Nima Khademi Kalantari. Photomat: A material generator learned from single flash photos. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2