Hierarchical Material Recognition from Local Appearance

Matthew Beveridge and Shree K. Nayar
Columbia University
{beveridge,nayar}@cs.columbia.edu

Abstract

We introduce a taxonomy of materials for hierarchical recognition from local appearance. Our taxonomy is motivated by vision applications and is arranged according to the physical traits of materials. We contribute a diverse, in-the-wild dataset with images and depth maps of the taxonomy classes. Utilizing the taxonomy and dataset, we present a method for hierarchical material recognition based on graph attention networks. Our model leverages the taxonomic proximity between classes and achieves state-of-the-art performance. We demonstrate the model's potential to generalize to adverse, real-world imaging conditions, and that novel views rendered using the depth maps can enhance this capability. Finally, we show the model's capacity to rapidly learn new materials in a few-shot learning setting.

1. Importance of Material Recognition

Our ability as humans to recognize materials is critical to every action we take. Using vision alone, we can infer that a coffee cup will be hot to the touch if we see it is made of paper. If it is instead perceived as a ceramic cup, we can deduce it may be cooler and more rigid, but heavier to lift. The ability to perceive materials from a distance thus enables us to infer the consequences of an action before taking it. While this is second nature to us, humans, material perception by machines remains an active field of research.

Material is a fundamental visual unit of a scene. Objects, defined by their form and function, convey *what* is in a scene. In concert, materials inform *how* to interact with the scene. Consider the case where our coffee cup is toppled and spills its contents. An autonomous agent must first identify the spilled material as a liquid to then reason that a towel is the proper cleanup tool, and not a broom. Furthermore, knowing a material enables estimation of mechanical properties such as weight and elasticity. In the case of our example, the agent must know that the towel is deformable in order to handle it. In short, the ability to visually identify materials is key to the development of autonomous systems that can interact with the environment in more nuanced and intelligent ways than possible today.

The term "material" can carry different meanings depending on the context. In our example, if the coffee is spilled on a wooden table, it may suffice to know that coffee is a liquid. However, if it is spilled on a fabric, the cleaning method used may be based on the liquid type (*e.g.*, coffee v. water). In other words, material recognition may need to be done at different levels of detail based on the application [2].

Just as naturalists have taxonomized living organisms into the tree of life, materials too conform to a hierarchy (*e.g.*, rubber $\subset$ plastic $\subset$ polymer). With this in mind, we constructed a hierarchical taxonomy for materials (Sec. 3). Our taxonomy includes materials that are commonly encountered in a variety of science and engineering disciplines, and its hierarchical structure is based on the physical properties of the materials. We have collected an in-the-wild image dataset, *Matador*, to populate the taxonomy with images of materials taken at different scales (magnifications) and under different natural lighting conditions (Sec. 4).[1] *Matador* has $\sim$7,200 samples across 57 material classes, where each sample includes the material's local appearance (close-up texture), 3D structure (from lidar), and the surrounding context (object-level information). Having 3D structure allows us to render arbitrarily many additional novel views for each sample, corresponding to different magnifications, orientations, and camera settings.

We use the taxonomy to develop a model for material recognition by framing recognition as a hierarchical image classification task. This is done by utilizing graph representation learning to predict the full taxonomic classification of a material from its local appearance (Sec. 5). We show that structuring a graph neural network according to our taxonomy improves classification accuracy on existing benchmarks as well as our own *Matador* dataset, achieving state-of-the-art performance. The performance remains high even when the taxonomy is sparsely populated with images and when color information is excluded. A key advantage of our hierarchical approach is that, even when a material is misclassified at the finest level, it can be correctly recognized at a higher level, still enabling useful inferences regarding its mechanical properties. Finally, we demonstrate that our model generalizes to adverse, real-world imaging conditions, that rendering novel views for training data can enhance this capability, and that the model is effective at few-shot classification (Sec. 6).

[1] *Matador* webpage: https://cave.cs.columbia.edu/repository/Matador.

2. Related Work

Images can be interpreted at multiple levels of granularity. At the coarsest level, an image represents a scene. Within a scene, there are objects, and each object is composed of materials. Each material manifests as a texture in the image. Although texture and material are distinct concepts, they are closely related, and much of the prior work has utilized textures to recognize materials. We refer the interested reader to Liu et al. [46] and Dana [17] for comprehensive historical perspectives on this line of work.

Material Recognition from Texture. The earliest approaches to texture recognition are based on the use of filter banks [6, 31, 41, 52, 72] or statistical analysis [14, 53]. Significant progress was later made by representing a texture using textons [35] that are constructed from the frequency responses to a set of hand-crafted filters [15, 16, 37, 44, 51, 70, 73–75], including Gabor filters [24, 72] and Gaussian kernels [40, 51]. Textons are aggregated into a histogram and used for recognition, similar to bags of features for object recognition. Subsequent methods enhanced the filter bank to handle multi-scale image features [19, 29, 43], rotation invariance [64] and affine invariance [42], and the downstream recognition routines were improved with Fisher vectors [57] and contextual priors [45].

The field has since shifted from hand-crafted filter banks to learned models. Most works in this vein utilize convolutional architectures (CNNs) to either learn a filter bank [12, 13] or a dictionary [63, 71, 85, 86, 93]. Since forming histograms from textons is effectively spatial pooling, much attention has been given to feature aggregation in CNNs [88, 90, 91]. Recently, incorporating multi-scale geometry in feature pooling has shown promising results [11, 23, 55, 83]. Such methods aim to replace average or max pooling with an operator that captures the orderless, fine-grained structure of textures without losing the higher-level spatial order. In contrast to prior work, we use a hierarchical taxonomy of materials, which allows us to leverage the taxonomic proximity between materials for recognition.

Material Image Datasets. Texture-based recognition aims to classify materials based on the intensity fluctuations they produce in images. Such fluctuations can be described as having characteristic patterns such as "striped" or "checkered" [12, 65, 66]. In this approach, there is no direct measurement of the physical properties of a material or the way it interacts with incident light. Ideally, material recognition would be explicitly based on the optical properties of a material [17]. However, this would require at least a partial measurement of the bidirectional reflectance distribution function (BRDF) [92] or the bidirectional texture function (BTF) [7, 18, 62, 79], which can only be done in controlled settings using specialized setups.

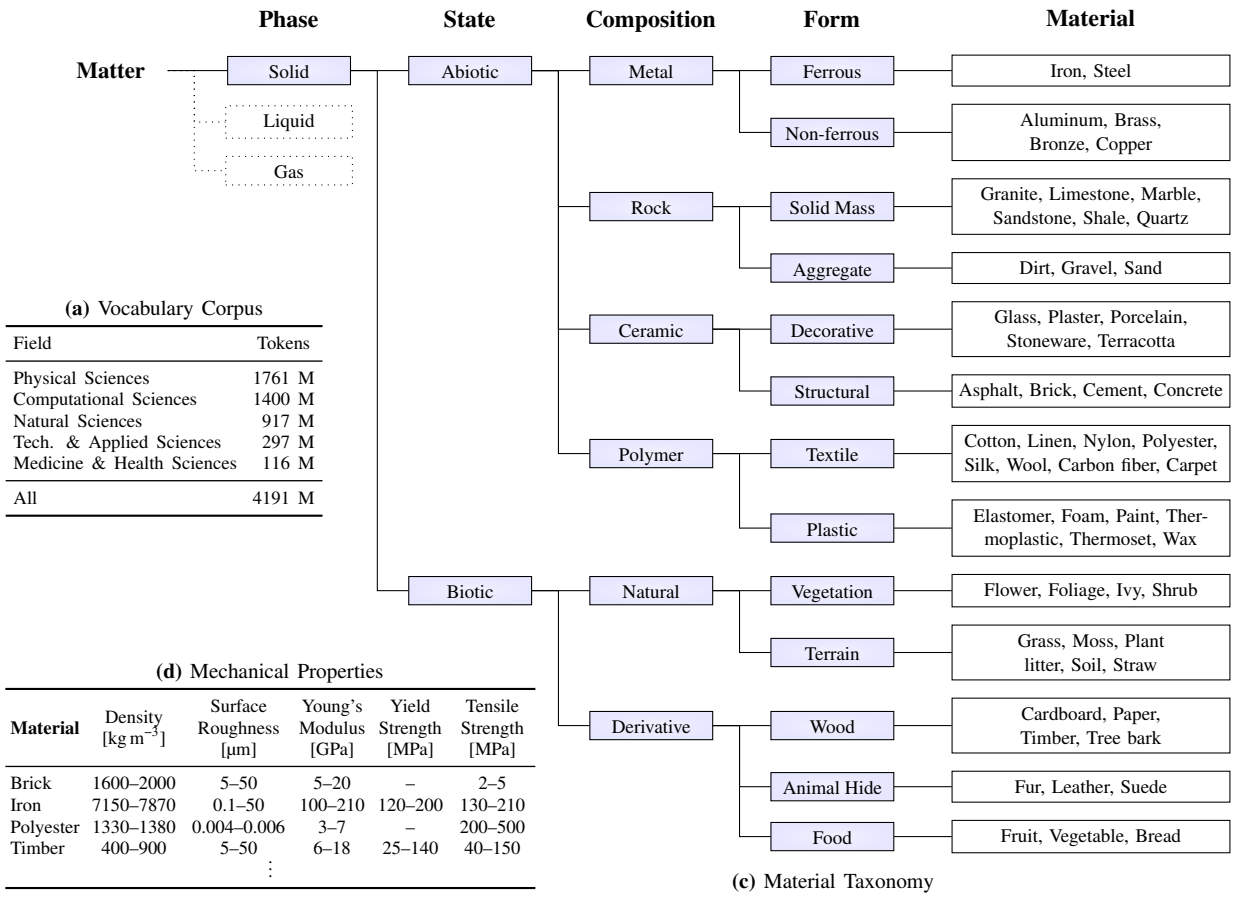
Learning-based approaches to material recognition instead lean on large image datasets to capture variability in camera pose and real-world lighting. Several datasets of this type exist and are either aggregated from online sources [3, 4, 69] or captured using a custom hardware platform [84, 85]. The former tend to also include some object-level information which can aid local material recognition [67]. In our work, we are chiefly interested in recognition based on local appearance. We present a novel dataset, *Matador*, of ∼7,200 in-the-wild samples spanning 57 material types. Using this dataset, we also render numerous novel views of each sample to augment model training by varying magnification, orientation, and camera settings [9].

Graph Representation Learning. The definition of a material is relative: while "brick" and "concrete" are distinct classes of materials, they are both instances of a "ceramic". This observation inspired our use of hierarchical learning to recognize materials. Hierarchical image classification [8, 10, 28, 48, 60, 87, 89, 94] has recently been used to take advantage of semantic relationships between classes, *e.g.*, in object recognition, where the hierarchy is defined by WordNet [54]. Several such methods utilize graph neural networks [81, 82] to explicitly represent hierarchical relationships, which have demonstrated impressive generalization capabilities [36, 56, 77, 78, 89]. Based on our taxonomy, we develop a graph neural network [39] to learn visual relationships between material classes. In doing so, should related taxonomy classes share visual traits (which is not necessarily guaranteed), we can exploit this for recognition.

3. A Visual Taxonomy of Materials

The names given to material categories are highly dependent on the context in which they are studied. While a roboticist may broadly characterize a cup as being made of plastic, a material scientist is likely to define it more finely, perhaps as polystyrene [34]. This is in part because different fields study materials at different scales. At one end, a material could be coarsely defined as a solid, and at the other end as a group of atoms. The granularity at which a material is perceived, therefore, depends on the level at which we wish to interact with it. We take inspiration from the tree of life in biology and define a taxonomy of materials, where the vocabulary is chosen to be pertinent to tasks requiring visual perception, and the structure is designed based on the physical traits of the materials.[2] The purpose of this taxonomy is to enable the creation of a framework that can (a) exploit visual similarities among related categories to improve classification, (b) relate unknown materials to known ones, and (c) make qualitative inferences about the mechanical properties of a material from its image.

[2]Modern phylogenetic trees of life are rooted in genome sequencing. Before this, most trees of life were based on external morphology (*i.e.*, visual appearance of organisms). In some ways, our approach represents a blend of both methods: our material taxonomy is organized by physical properties, but our recognition method utilizes passive visual imaging.



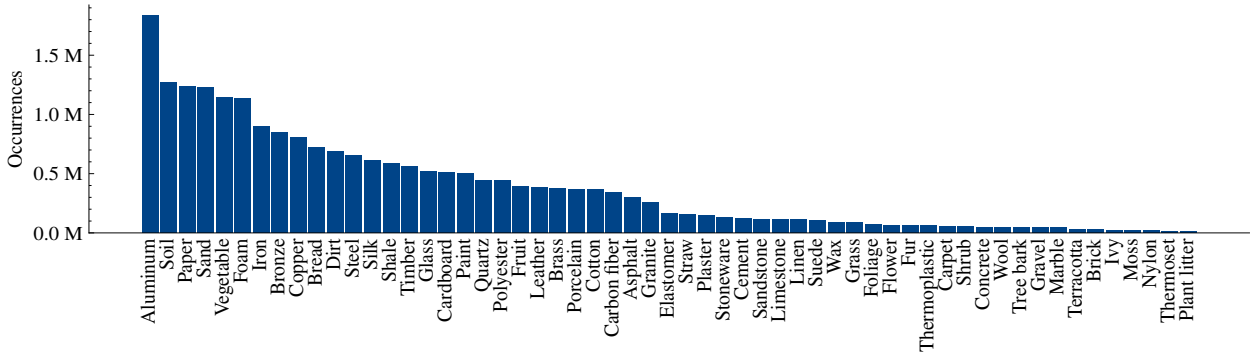
(a) Vocabulary Corpus

Field	Tokens
Physical Sciences	1761 M
Computational Sciences	1400 M
Natural Sciences	917 M
Tech. & Applied Sciences	297 M
Medicine & Health Sciences	116 M
All	4191 M

(c) Material Taxonomy

(d) Mechanical Properties

Material	Density [kg m^{-3}]	Surface Roughness [µm]	Young's Modulus [GPa]	Yield Strength [MPa]	Tensile Strength [MPa]
Brick	1600–2000	5–50	5–20	–	2–5
Iron	7150–7870	0.1–50	100–210	120–200	130–210
Polyester	1330–1380	0.004–0.006	3–7	–	200–500
Timber	400–900	5–50	6–18	25–140	40–150
⋮					



(b) Overall Word Occurrences

Figure 1. Framework for the Taxonomic Classification of Materials. (a) We search through archives of text in the scientific literature to build a vocabulary for materials based on the frequency of word occurrences. (b) Shown here is the word occurrence plot for the 57 materials that are most popular in the fields listed in (a). Such a vocabulary can be generated for subfields as well (*e.g.*, robotics). (c) Using the vocabulary in (b), we create a taxonomy for the classification of materials. The structure of the taxonomy is organized based on the physical traits shared by the materials. The advantage of this taxonomy in the context of recognition is that, even when a material cannot be uniquely identified from its appearance, it still may be identified at a higher level in the taxonomy (form, composition, state, or phase). Note that the taxonomy focuses on solids, but could be extended to liquids and gases as well in the future. (d) In some applications, it may be useful to know how a material would behave when physically interacted with. To this end, we have listed the mechanical properties for the materials in our taxonomy (the full list is in Sec. S4 of the supplemental material).

A Vocabulary of Materials. Inspired by Bhushan et al. [5], we first set out to create a vocabulary that includes all types of materials an intelligent system might encounter. We began with all the names of materials in WordNet [54]. We then scanned for occurrences of these names in historical corpora of text, ranging in focus from popular science to niche conference and journal publications. The fields we focused on are listed in Fig. 1a. The complete corpus has billions of words and phrases sourced from M2D2 [59]. We then aggregated word occurrences of synonyms, and each synonym group is given a material name. We end up with a distilled set of 57 categories, the word occurrence distribution for which is shown in Fig. 1b. One could use all of these categories, or, for a given application domain, one could regenerate the distribution of word occurrences to find the categories that are most relevant to that domain.

Hierarchical Representation. Next, we arranged the above material categories into a taxonomy. Our taxonomy, inspired by Schwartz and Nishino [68], is shown in Fig. 1c. The higher-level structure (phase, state, composition, and form) was derived from existing material classifications used in the fields of material science, mechanical engineering, and chemical engineering. Note that we focus on solids in our work as they are most relevant to vision applications such as robotics and autonomous driving. Furthermore, the definitions of some of the materials have been stretched to accommodate corner cases. This is because most engineered materials are made of multiple constituent materials. In such cases, we deemed it more useful to categorize by the dominant component. This taxonomy is intended to serve as a starting point, and it can be expanded (to include liquids, for instance) with time.

Mechanical Properties. For an intelligent system to fully benefit from material recognition, it must know more than just the names of the materials in its field of view. It must also have an understanding of how each material will behave upon interaction. To this end, we have compiled a table of mechanical properties for each leaf (material) in our taxonomy. The relevant parameters include density, surface roughness, elasticity, and strength. The ranges for these parameters were obtained from existing literature in material science and mechanical engineering. An abbreviated version of this table is shown in Fig. 1d; the full table with all the materials is in the supplemental material (Sec. S4).

4. *Matador*: A Material Image Dataset

To populate our taxonomy, we collected a dataset, named *Matador*, that consists of over 7,200 samples across the 57 different materials in our taxonomy. Fig. 2 shows examples of samples in the dataset. For each sample, we capture a close-up, high-resolution image of the local appearance (color in Fig. 2a, grayscale in Fig. 2b) and a registered lidar scan of the material's 3D surface structure (Fig. 2c). The 3D

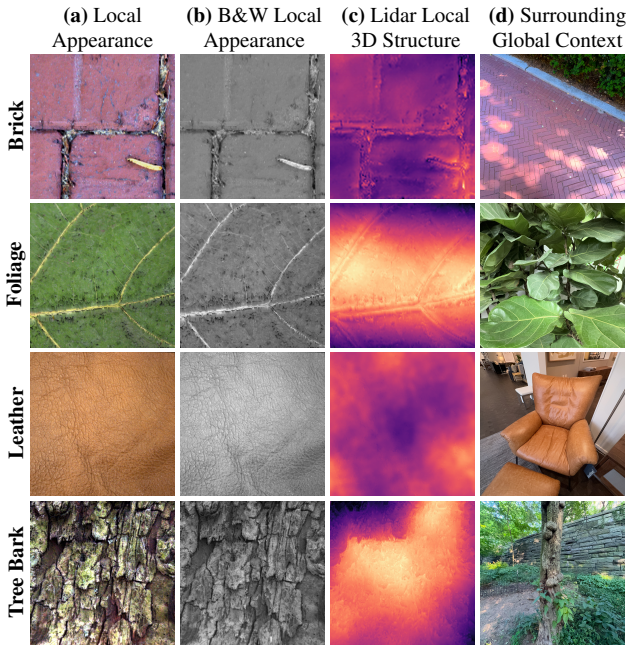


Figure 2. The *Matador* Dataset. (a-b) Each sample includes a real-world image of a material taken at a high resolution and (c) its 3D structure (depth map). (d) The surrounding context is also captured. The dataset comprises ∼7,200 samples across the 57 material categories of the proposed taxonomy (Sec. 3).

structure was captured to enable us to generate additional views of the material, corresponding to different magnifications, orientations, and camera settings. A wide-angle image of the sample in its surrounding context is also captured (Fig. 2d). *Matador* is unique in three respects: (a) the diversity of materials it includes, (b) the hierarchical labels assigned to samples, and (c) the use of 3D structure to generate novel views. It is available on the *Matador* webpage.

Data Collection. To collect the dataset, we developed an iOS application. A smartphone allows for increased mobility when capturing in-the-wild samples, meaning the user can image the sample at the camera's minimum focus distance to maximize spatial resolution. We chose the iPhone 15 Pro Max as our platform for the quality of its cameras and lidar.[3] For each sample, the local appearance and 3D structure are captured using the wide-angle camera (12MP, 74° FOV) and lidar (100 points/degree^2, 74° FOV). Subsequently, an additional context image is taken from a more distant viewpoint using the ultrawide-angle camera (12MP, 104° FOV). Both images (appearance and context) are 12-bit Bayer raw. For cases where one may wish to register the local and context images, we also record the phone's

[3]Among tested smartphones, the iPhone 15 Pro Max has the only lidar capable of capturing depth maps at close distances (around the camera's minimum focus distance of ∼15 cm). We use lidar since stereo matching can be unreliable for weakly textured materials. Additionally, the iPhone 15 Pro Max has superior optical performance with a measured MTF50 of 0.240 cycles/pixel compared to, *e.g.*, the Pixel 8 Pro's 0.158 cycles/pixel.

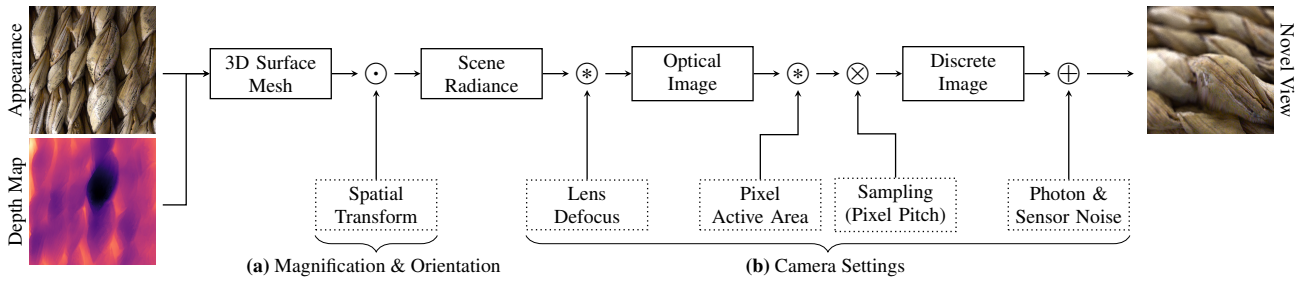


Figure 3. Rendering Novel Views from a Real-World Sample. From a captured material sample, we simulate its appearance under different magnifications, orientations, and camera settings. (a) We first create a 3D mesh and texture map it with the appearance image. We then apply spatial transformations to the mesh to change its pose. (b) The optical image of a novel view (including depth of field effects) is obtained by raytracing. It is then blurred to account for pixel area, and the result is sampled to produce the discrete image. Finally, noise is added, resulting in the novel view. By varying the parameters in this process, we render numerous novel views for each real-world sample.

motion during the time between the two captures using the phone's inertial measurement unit. Since the data collection is done in the wild, lighting and camera viewpoint vary across samples. We have publicly released the iOS application, enabling *Matador* to grow in size and scope over time.

Rendering Novel Views. Using the 3D structure of the captured sample, we can simulate images under different magnifications, orientations, and camera settings. This process is outlined in Fig. 3. We begin by creating a 3D mesh from the depth map and texture-mapping it with the surface radiance values measured in the local appearance image. Spatial transformations are then applied to the mesh to change the pose of the original sample. For the purpose of view generation, we assume the sample to be Lambertian and do not alter illumination (each *Matador* class already captures a wide variety of point and extended sources). A high-resolution (optical) image for a novel viewpoint is then generated by raytracing the mesh, using a thin-lens model to include depth-of-field (defocus) effects [32]. Raytracing implicitly accounts for occlusions and perspective effects, which are both important in the case of significant surface depth variations. To produce the discrete image, we first blur the optical image with a box filter whose support is equal to the pixel active area. The result is then sampled by a pulse train with period equal to the pixel pitch. Finally, we add photon and sensor noise to produce the novel view.

By altering the pose of the sample, defocus of the lens, pixel active area, pixel pitch, and image noise level, we can simulate how an arbitrary camera would image the sample at any magnification and orientation. Many novel views are rendered for each sample (see Fig. 4). This process is applied to all ∼7,200 raw samples in the *Matador* dataset to obtain a larger and more diverse set of material images to supplement model training, which we show improves generalization to real-world settings (Sec. 5).

Generalization Test Set. To evaluate the impact novel views have on recognition performance in the wild, we construct an additional out-of-distribution (OOD) test set. Given a material sample in *Matador*, we begin with the region of interest (ROI) defined in the appearance image. Using the depth map and motion between appearance and context captures, we map the ROI onto the context image and extract a patch of similar scale. Accelerometer noise means this patch does not exactly align with the ROI seen in the appearance image, and instead captures a different area on that instance of material. In addition, as the context image has a different viewpoint, image sensor, and field of view, the resultant patches comprise an OOD test set. See Sec. S2.4 of the supplemental material for further details on the construction of this test set, as well as example images.

Figure 4. Rendered Novel Views for Gravel. Examples of the many novel views rendered from a single real-world sample of gravel (left) using the process described in Sec. 4.

5. Recognizing Materials

We now describe a method that leverages our proposed taxonomy to recognize materials from their local appearance. Consider the materials "iron" and "brass" in Fig. 1c; they are distinct classes but share a common ancestor, "metal", which is characterized by a range of mechanical properties. They are arranged in the taxonomy according to their mechanical properties, and we speculate that materials with close taxonomic proximity may share visual traits. While not guaranteed, should such similarities exist, our aim is to exploit them for recognition. In addition, in cases where we are unable to recognize a local appearance as "brass", it would still be useful for a downstream application to know that it is a metal. For this reason, we design our recognition method to account for the structure of our taxonomy. To this end, we use a graph neural network (GNN) to encode the structure of the taxonomy and thereby constrain how materials share features in latent space (based on their taxo-

nomic proximity). While training this model, our objective is to predict an image's full taxonomic classification, *i.e.*, the class at each level of the taxonomy. We refer to this approach as *hierarchical material recognition*.

To achieve hierarchical recognition, the proposed taxonomy is first translated into a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes in the taxonomy and $\mathcal{E}$ is the set of edges that represent parent-child relationships. We then treat recognition as a node classification task. Given a set of images $\mathcal{T}$, the label for each image x is the hierarchical path of the material in the taxonomy $\ell \subseteq \mathcal{V}$. For example, the label of an image of steel would be {solid, abiotic, metal, ferrous, steel}. The set of all images whose label set ℓ contains the node v_i is then $\mathcal{T}_i = \{x_j \in \mathcal{T} \,|\, v_i \in \ell_j\}$. We initialize a learnable vector for node v_i in our graph model with the average image features over $\mathcal{T}_i$, produced by an encoder ϕ such as a convolutional neural network. Explicitly, the initial feature vector of v_i is $\mathbf{h}_i^0 = \frac{1}{||\mathcal{T}_i||} \sum_{x_j \in \mathcal{T}_i} \phi(x_j)$. To classify an image x, a global node $\mathbf{h}_g = \phi(x)$ is inserted into the graph with outgoing edges to every other node.

For our construction, we use a graph attention network (GAT) [76]. The message passing update for a node v_i at level k of the model is therefore defined by the relation:

$$\mathbf{h}_i^{k+1} = \psi_a \left(\mathbf{h}_i^k, \oplus_{j \in \mathcal{N}_i} \left(\alpha_{ij}^k \, \psi_b \left(\mathbf{h}_i^k, \mathbf{h}_j^k \right) \right) \right), \qquad (1)$$

where ψ_a and ψ_b are multi-layer perceptrons. The local neighborhood $\mathcal{N}_i$ of node v_i is fixed by the adjacency matrix of $\mathcal{G}$, and $\oplus$ is a permutation-invariant aggregation function (summation in our case). The edge attention mechanism α_{ij}^k denotes the connective strength between node v_i and its neighbor v_j, and this is used to weigh the sharing of visual traits amongst taxonomy nodes (*e.g.*, from individual child nodes to their parent). Since Eq. (1) only operates on the local neighborhood of each node, we cascade up to D layers to match the diameter of the taxonomy and propagate updates throughout the model. Residual connections are also added between layers to mitigate oversmoothing gradients.

We jointly train ϕ and the GAT classifier (α, ψ_a, ψ_b) end-to-end, where ϕ is a ResNet50 [30] initialized by pretraining on the IG-1B [50] and ImageNet1k [20] datasets. As our goal is multi-label hierarchical classification, we use binary cross-entropy (BCE) as our loss on the label set $\ell = \left[\ell^0, \ell^1, \ldots, \ell^{D-1}\right]$. We also use cross-entropy (CE) as our loss for each hierarchy level. The full loss is a greedy combination of both: $\max\left(\text{BCE}(\hat{\ell}, \ell), \frac{1}{D}\sum_{d=0}^{D-1} \text{CE}\left(\hat{\ell}^d, \ell^d\right)\right)$, where the first term encourages learning complete hierarchical paths, and the second term encourages correct predictions within each level of the hierarchy.

5.1. Flat Classification

To validate our model, we first compare its performance with recent methods on standard material recognition datasets. These local appearance datasets include KTH-TIPS2-b [7], FMD [69], GTOS [84], and GTOS-Mobile [85]. KTH-TIPS2-b extends CUReT [18] and is a BTF dataset where materials are imaged in a controlled setting with different magnifications, orientations, and illuminations. FMD includes a similar number of materials, but the images are instead aggregated from online sources. GTOS is larger in scope, comprising radiometric images of ground terrain captured using specialized hardware. GTOS-Mobile extends GTOS by adding smartphone images for a subset of the original GTOS materials. Further dataset details can be found in the supplemental material (Tab. S1).

Table 1. Efficacy of Graph Representation Learning. We compare our hierarchical model with recent material recognition methods. In all cases, including ours, a ResNet50 feature extractor was used. The best overall results are highlighted in bold, and the second best are underlined. Leveraging the structure of our taxonomy through graph representation learning yields significant improvements in recognition performance.

	Top-1 Accuracy ↑			
Method	KTH-2-b	FMD	GTOS	GTOS-M
DeepTEN [93]	82.0±3.3	80.2±0.9	84.5±2.9	—†
MAPNet [90]	84.5±1.3	85.7±0.7	84.7±2.2	86.6
DSRNet [91]	85.9±1.3	86.0±0.8	85.3±2.0	<u>87.0</u>
FENet [83]	86.6±0.1	82.3±0.3	83.1±0.2	85.1
CLASSNet [11]	87.7±1.3	86.2±0.9	<u>85.6±2.2</u>	85.7
DFAEN [88]	86.3	86.9	—†	86.9
RADAM [63]	88.5±3.2	85.3±0.4	81.8±1.1	81.0
FRP [23]	<u>90.7</u>	<u>88.8</u>	—†	—†
Ours	**93.5**±4.0	**96.1**±0.6	**87.9**±2.1	**92.2**

† Value not reported in the original work.

To evaluate the performance of our model, categories from the above datasets are mapped into our taxonomy, renaming and inserting leaf nodes as necessary. This process does not alter the high-level structure of the taxonomy. We measure test accuracy on multi-class flat classification after training our model with multi-label hierarchical learning. The top-1 accuracy is reported in Tab. 1, where all models use the same ResNet50 backbone. We can see our hierarchical, graph-based classifier results in a significant performance increase with respect to existing methods.

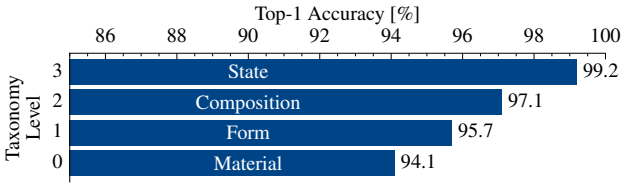
Next, we evaluate the performance of our model on our *Matador* dataset. For this evaluation, we emphasize performance on the materials in *Matador* that we deem to have sufficient texture to be identifiable using local appearance alone. We create a subset of *Matador* where materials such as glass, plastics, and paint are omitted as they are flat in appearance. In addition, some material classes are consolidated into a single class. For instance, all metals have similar appearances (finishes) except for changes in hue, and hence are combined into a single category. With these omissions and consolidations, we arrived at a dataset containing 37 material classes with ∼6,600 samples. We refer to this slice of *Matador* as *Matador-C1*, and details of its construction are given in the supplemental material (Sec. S2).

(a) Top-1 accuracy, ablating novel views during training of our models.

Method	Params	Top-1 Accuracy ↑ Matador	Matador-C1	Out-of-Dist.†
Vision Foundation Models (Zero-Shot)				
CLIP [58]	151 M	24.8	40.0	32.3
GPT-4.1 [1]	1.76 T‡	51.4	65.9	53.4
Material Recognition Models (ResNet50 Backbone, Finetuned)				
DeepTEN [93]	24 M	79.2	88.8	61.5
DEPNet [85]	25 M	82.7	87.6	76.1
FRP [23]	28 M	74.8	89.4	71.0
MSLac [55]	24 M	82.6	88.5	75.4
Modern Image Recognition Models (Finetuned)				
ConvNext-V2 [80]	28 M	83.1	89.7	81.9
EVA-02 [21]	22 M	85.8	90.1	82.7
Hierarchical Material Recognition Models (Various Backbones)				
Ours (ResNet50)	28 M	85.8	94.1	82.9 (+2.8)
Ours (ConvNext-V2)	31 M	85.9	94.2	86.8 (+2.4)
Ours (EVA-02)	24 M	**88.3**	**94.7**	**87.0** (+4.9)

Gain from Novel Views

† Train on appearance images, evaluate on patches of context images.
‡ Unreported for GPT-4.1, value reported is for its predecessor GPT-4.



(b) "Ours (ResNet50)" top-1 accuracy on *Matador-C1* by taxonomy level.

Figure 5. Performance on *Matador*. (a) Comparison with existing methods. The best overall results are highlighted in bold, and the second best are underlined. Our model demonstrates state-of-the-art recognition, and supplementing training with rendered novel views enhances performance on out-of-distribution (OOD) data (see Sec. 4). (b) Hierarchical classification accuracy. Note that accuracy increases with the taxonomy level, suggesting that exploited visual similarities at lower levels improve recognition at higher levels, and that an image may be correctly classified at a higher level (*e.g.*, form) even when its material class is uncertain.

The performance of our method and existing methods on the full *Matador* dataset and the consolidated *Matador-C1* dataset is summarized in Fig. 5a. All finetuned models are trained with the same optimizers (Muon [33] and AdamW [49]). Compared to previous methods, we again achieve state-of-the-art performance. We additionally show how supplementing the training data with rendered novel views considerably improves generalization to OOD imaging conditions (up to 4.9%). We refer the reader to Sec. S3 of the supplemental material for class-specific accuracies of our model, its performance on grayscale images, and competing material recognition model performance when only finetuning the classifier head instead of the entire model.

5.2. Hierarchical Classification

Though we use hierarchical inference during training of our model, to this point, we have only reported performance for flat classification of the taxonomy leaves. We now examine the capability of our model for multi-label hierarchical classification. Standard methods for hierarchical classification use a deterministic [70] or probabilistic [26] walk down the levels of the tree. Using these techniques, there is a risk of outputting invalid label combinations unless care is taken to enforce the hierarchy structure (*e.g.*, through a specialized loss function or inference routine). In our case, we encode the hierarchy structure in the GNN and also use a hierarchical loss function. For this evaluation, however, we do not enforce hierarchical consistency in the inference routine (which we do later in Sec. 6). In Fig. 5b, we present classification accuracy on *Matador-C1* for each level of the taxonomy. As expected, we see higher performance at higher levels of the taxonomy where the number of classes is fewer. This result suggests that using a GNN to encode the taxonomy structure effectively guides hierarchical learning by exploiting any shared visual traits amongst taxonomy nodes. It also implies that when we are unable to correctly recognize an image at the finest level (material), we may still be able to recognize it at a coarser one (*e.g.*, form).

6. Implications and Extensions

We now show how our proposed model may be used in practice and discuss potential extensions of our work.

Probing Materials in a Scene. In many vision tasks, such as navigation, it is desirable to know the material at an arbitrary point in a scene. That is, we would like to query a pixel in an image to determine its material. As noted in Sec. 4, materials may appear very different at different scales. Therefore, what is a good strategy for selecting an appropriate region around a probed image point to use for classification? If depth is available, one could fix the window size using metric units. Here, as in Sec. 5, we focus on the more common case where depth is not available during inference. Given a probed image pixel, we use windows of increasing size from 64×64 px^2 to 1024×1024 px^2. These windows are passed through our model using Monte Carlo dropout [25, 38] to build a distribution of the predictions. This distribution and its entropy then guide a best-first search of the taxonomy tree to obtain a hierarchically consistent classification for the chosen image point.

Fig. 6 shows the above probing approach applied to several points in an imaged scene. In each case, the window that produced the highest confidence classification is displayed. Note that all probed points in this scene are at significantly greater distances (lower magnifications) than the training samples in *Matador*. Yet, we are able to correctly classify most of the points due to the rendered novel views used during training. In cases where we are unable to correctly identify the low-level material, we can still recognize it at a higher level of the taxonomy. Two examples of this are the wool blanket on the couch that was classified as car-

Figure 6. Probing Materials in a Scene. Given an image, we can probe the material at a point to determine its taxonomic classification. For each probed point, we automatically find the window size that maximizes classification confidence. Each classification is hierarchical, and the finest level that exceeds a confidence threshold is shown in bold. We can associate the predicted material with its known mechanical properties to provide information that could be useful, *e.g.*, for robot manipulation. These properties (*e.g.*, "Strong, Rigid, Light") are listed below the finest classification achieved by our model. Strong v. Fragile is determined by a tensile strength threshold, Rigid v. Deformable by a Young's modulus threshold, and Heavy v. Light using a density threshold.[4] Note that when we are unable to identify the exact material (*e.g.*, the wool blanket misclassified as carpet), we still achieve correct classification ("textile") at a coarser level in the taxonomy.

pet and the soil in the planter that was classified as moss. A natural extension of this method is to apply probing to all pixels in the image and then use neighboring pixels to inhibit or reinforce the classification of each pixel. The result would be a material-based segmentation of the entire scene.

What about Unseen Materials? We now discuss an important extension to material recognition: few-shot learning [22, 27, 47, 61] of novel categories. Such a capability is desirable for an intelligent system to infer the mechanical properties of previously unseen materials. We evaluate our hierarchical model in this setting by training copies on modified versions of the *Matador-C1* dataset, each having a single material category held out. In Fig. 7, the blue plot shows the average accuracy of our model on the held-out class as a function of the number of held-out samples reintroduced into training. Note that novel material classes are learned rapidly. The red plot shows the average path distance, *i.e.*, the number of edges (hops) between the predicted node and the correct node in the taxonomy. We observe that, with as few as 16 samples of a novel class, we achieve an average accuracy of ∼90% and an average path distance of less than 2 hops. The latter suggests that our model, using a small number of samples, is able to learn the hierarchical label of a material in our taxonomy that has not been seen before. Even when the model is unable to correctly recognize the material, it is able to identify it correctly at one level higher in the taxonomy. This potential for rapid adaptation could enable intelligent systems to learn novel materials added to our taxonomy (*e.g.*, liquids and gases) with little expense.

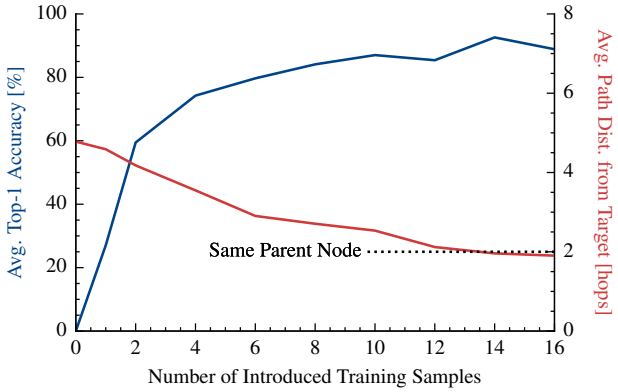


Figure 7. Few-Shot Learning of New Materials. (Blue Plot) Average performance for classes not seen during pretraining, as a function of the number of training samples reintroduced during finetuning. The plot shows that novel classes can be learned with a small number of samples, achieving ∼90% accuracy with just 16 samples. (Red Plot) With the same number of samples, misclassifications are, on average, localized to siblings at the lowest level of the taxonomy (*i.e.*, "Materials" with the same "Form").

7. Summary

In this work, we have introduced a taxonomy of materials suited for visual perception, a diverse dataset (*Matador*) of material images, a rendering-based training augmentation that improves OOD model generalization, and a graph representation method that leverages the taxonomy for visual recognition. On existing datasets and *Matador*, we have shown the efficacy of hierarchical learning using graph attention. We demonstrated the ability of this model to estimate the taxonomic class of a material, and to quickly adapt to unseen materials through few-shot learning. We believe that hierarchical material recognition can help intelligent systems perform more sophisticated tasks and, in addition, achieve higher levels of safety and reliability.

[4] Thresholds used are 1 MPa, 12 GPa and 1,600 kg m^{-3}, respectively.

Acknowledgments

This work was supported by the Office of Naval Research (ONR) under award number UWIS-0000003591. The authors are grateful to Behzad Kamgar-Parsi for his support and encouragement. The authors thank Mohit Gupta and Jeremy Klotz for their technical feedback, and Aubrey Toland for discussions on engineered materials. The authors are also grateful to Hannah Fox, Sidharth Sharma, Joel Salzman, and Pranav Sukumar for help with collecting the Matador dataset, and Jessica Zhang and Lulu Wang for assistance refining the Matador website.

References

[1] Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/.

[2] Edward H. Adelson. On seeing stuff: The perception of materials by humans and machines. In *Human Vision and Electronic Imaging VI*, pages 1–12. SPIE, 2001.

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

[4] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material Recognition in the Wild With the Materials in Context Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3479–3487, 2015.

[5] Nalini Bhushan, A. Ravishankar Rao, and Gerald L. Lohse. The Texture Lexicon: Understanding the Categorization of Visual Texture Terms and Their Relationship to Texture Images. *Cognitive Science*, 21(2):219–246, 1997.

[6] A.C. Bovik, M. Clark, and W.S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):55–73, 1990.

[7] B. Caputo, E. Hayman, and P. Mallikarjuna. Class-specific material categorisation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1597–1604 Vol. 2, Beijing, China, 2005. IEEE.

[8] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your "Flamingo" is My "Bird": Fine-Grained, or Not. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11471–11480, Nashville, TN, USA, 2021. IEEE.

[9] Jacob R. Cheeseman, Roland W. Fleming, and Filipp Schmidt. Scale ambiguities in material recognition. *iScience*, 25(3): 103970, 2022.

[10] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-Grained Representation Learning and Recognition by Exploiting Hierarchical Semantic Embedding. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 2023–2031, Seoul Republic of Korea, 2018. ACM.

[11] Zhile Chen, Feng Li, Yuhui Quan, Yong Xu, and Hui Ji. Deep Texture Recognition via Exploiting Cross-Layer Statistical Self-Similarity. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5227–5236, Nashville, TN, USA, 2021. IEEE.

[12] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, Columbus, OH, USA, 2014. IEEE.

[13] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3828–3836, Boston, MA, USA, 2015. IEEE.

[14] George R. Cross and Anil K. Jain. Markov Random Field Texture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(1):25–39, 1983.

[15] O.G. Cula and K.J. Dana. Compact representation of bidirectional texture functions. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–I, 2001.

[16] Oana G. Cula and Kristin J. Dana. 3D Texture Recognition Using Bidirectional Feature Histograms. *International Journal of Computer Vision*, 59(1):33–60, 2004.

[17] Kristin J. Dana. *Computational Texture and Patterns: From Textons to Deep Learning*. Springer International Publishing, Cham, 2018.

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

[19] J.S. De Bonet and P. Viola. Texture recognition using a non-parametric multi-scale statistical model. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, pages 641–647, 1998.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[21] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

[22] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[23] Joao B. Florindo. Fractal pooling: A new strategy for texture recognition using convolutional neural networks. *Expert Systems with Applications*, 243:122978, 2024.

[24] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biological Cybernetics*, 61(2):103–113, 1989.

[25] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, 2016.

[26] Dehong Gao, Wenjing Yang, Huiling Zhou, Yi Wei, Yi Hu, and Hao Wang. Deep Hierarchical Classification for Category Prediction in E-commerce System, 2020.

[27] Michelle Guo, Edward Chou, De-An Huang, Shuran Song, Serena Yeung, and Li Fei-Fei. Neural Graph Matching Networks for Fewshot 3D Action Recognition. In *Proceedings of*

the European Conference on Computer Vision (ECCV), pages 653–669, 2018.

[28] Yanming Guo, Yu Liu, Erwin M. Bakker, Yuanhao Guo, and Michael S. Lew. CNN-RNN: A large-scale hierarchical image classification framework. *Multimedia Tools and Applications*, 77(8):10251–10271, 2018.

[29] E. Hadjidemetriou, M.D. Grossberg, and S.K. Nayar. Multiresolution histograms and their use for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):831–847, 2004.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[31] Anil K. Jain and Farshid Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24 (12):1167–1186, 1991.

[32] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, and Delio Vicini. DR.JIT: A just-in-time compiler for differentiable rendering. *ACM Transactions on Graphics*, 41(4):1–19, 2022.

[33] Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2025.

[34] William D. Callister Jr and David G. Rethwisch. *Callister's Materials Science and Engineering*. John Wiley & Sons, 2020.

[35] Bela Julesz. Experiments in the Visual Perception of Texture. *Scientific American*, 232(4):34–43, 1975.

[36] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. Rethinking Knowledge Graph Propagation for Zero-Shot Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11479–11488, Long Beach, CA, USA, 2019. IEEE.

[37] L.M. Kaplan. Extended fractal analysis for texture classification and segmentation. *IEEE Transactions on Image Processing*, 8(11):1572–1585, 1999.

[38] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, 2017.

[39] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, 2017.

[40] Jan J. Koenderink. The structure of images. *Biological Cybernetics*, 50(5):363–370, 1984.

[41] Kenneth I. Laws. Rapid Texture Identification. In *Image Processing for Missile Guidance*, pages 376–381. SPIE, 1980.

[42] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

[43] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2169–2178, 2006.

[44] Thomas Leung and Jitendra Malik. Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[45] Ce Liu, Lavanya Sharan, Edward H. Adelson, and Ruth Rosenholtz. Exploring features in a Bayesian framework for material recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 239–246, 2010.

[46] Li Liu, Jie Chen, Paul Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietikäinen. From BoW to CNN: Two Decades of Texture Representation for Texture Classification. *International Journal of Computer Vision*, 127(1):74–109, 2019.

[47] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. Prototype Propagation Networks (PPN) for Weakly-supervised Few-shot Learning on Category Graph, 2019.

[48] Yang Liu, Lei Zhou, Pengcheng Zhang, Xiao Bai, Lin Gu, Xiaohan Yu, Jun Zhou, and Edwin R. Hancock. Where to Focus: Investigating Hierarchical Attention Relationship for Fine-Grained Visual Classification. In *Computer Vision – ECCV 2022*, pages 57–73, Cham, 2022. Springer Nature Switzerland.

[49] Ilya Loshchilov and Frank Hutter. Fixing Weight Decay Regularization in Adam, 2018.

[50] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the Limits of Weakly Supervised Pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

[51] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and Texture Analysis for Image Segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

[52] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[53] Jianchang Mao and Anil K. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 25(2):173–188, 1992.

[54] George A. Miller. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41, 1995.

[55] Akshatha Mohan and Joshua Peeples. Lacunarity Pooling Layers for Plant Image Classification using Texture Analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5384–5392, Seattle, WA, USA, 2024. IEEE.

[56] Nihal V. Nayak and Stephen H. Bach. Zero-Shot Learning with Common Sense Knowledge Graphs, 2022.

[57] Florent Perronnin and Christopher Dance. Fisher Kernels on Visual Vocabularies for Image Categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[58] 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.

[59] Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. M2D2: A Massively Multi-domain Language Modeling Dataset, 2022.

[60] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR 2011*, pages 1481–1488, 2011.

[61] Victor Garcia Satorras and Joan Bruna Estrach. Few-Shot Learning with Graph Neural Networks. In *International Conference on Learning Representations*, 2018.

[62] Mirko Sattler, Ralf Sarlette, and Reinhard Klein. Efficient and Realistic Visualization of Cloth.

[63] Leonardo Scabini, Kallil M. Zielinski, Lucas C. Ribas, Wesley N. Gonçalves, Bernard De Baets, and Odemir M. Bruno. RADAM: Texture Recognition through Randomized Aggregated Encoding of Deep Activation Maps, 2023.

[64] C. Schmid. Constructing models for content-based image retrieval. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages II–II, 2001.

[65] Gabriel Schwartz and Ko Nishino. Visual Material Traits: Recognizing Per-Pixel Material Context. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 883–890, 2013.

[66] Gabriel Schwartz and Ko Nishino. Automatically Discovering Local Visual Material Attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3565–3573, 2015.

[67] Gabriel Schwartz and Ko Nishino. Material Recognition from Local Appearance in Global Context, 2017.

[68] Gabriel Schwartz and Ko Nishino. Recognizing Material Properties from Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1981–1995, 2020.

[69] Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H. Adelson. Recognizing Materials Using Perceptually Inspired Features. *International Journal of Computer Vision*, 103(3): 348–371, 2013.

[70] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[71] Yang Song, Fan Zhang, Qing Li, Heng Huang, Lauren J. O'Donnell, and Weidong Cai. Locally-Transferred Fisher Vectors for Texture Classification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4922–4930, Venice, 2017. IEEE.

[72] M. R. Turner. Texture discrimination by Gabor functions. *Biological Cybernetics*, 55(2):71–82, 1986.

[73] Manik Varma and Rahul Garg. Locally Invariant Fractal Features for Statistical Texture Classification. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[74] Manik Varma and Andrew Zisserman. A Statistical Approach to Texture Classification from Single Images. *International Journal of Computer Vision*, 62(1):61–81, 2005.

[75] Manik Varma and Andrew Zisserman. A Statistical Approach to Material Classification Using Image Patch Exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2032–2047, 2009.

[76] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks, 2018.

[77] Jin Wang and Bo Jiang. Zero-Shot Learning via Contrastive Learning on Dual Knowledge Graphs. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 885–892, Montreal, BC, Canada, 2021. IEEE.

[78] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6857–6866, Salt Lake City, UT, 2018. IEEE.

[79] Michael Weinmann, Juergen Gall, and Reinhard Klein. Material Classification Based on Training Data Synthesized Using a BTF Database. In *Computer Vision – ECCV 2014*, pages 156–171. Springer International Publishing, Cham, 2014.

[80] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023.

[81] Peng Xia, Xingtong Yu, Ming Hu, Lie Ju, Zhiyong Wang, Peibo Duan, and Zongyuan Ge. HGCLIP: Exploring Vision-Language Models with Graph Representations for Hierarchical Understanding, 2024.

[82] Yifan Xing, Tong He, Tianjun Xiao, Yongxin Wang, Yuanjun Xiong, Wei Xia, David Wipf, Zheng Zhang, and Stefano Soatto. Learning Hierarchical Graph Neural Networks for Image Clustering, 2021.

[83] Yong Xu, Feng Li, Zhile Chen, Jinxiu Liang, and Yuhui Quan. Encoding Spatial Distribution of Convolutional Features for Texture Representation. In *Advances in Neural Information Processing Systems*, pages 22732–22744. Curran Associates, Inc., 2021.

[84] Jia Xue, Hang Zhang, Kristin Dana, and Ko Nishino. Differential Angular Imaging for Material Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–773, 2017.

[85] Jia Xue, Hang Zhang, and Kristin Dana. Deep Texture Manifold for Ground Terrain Recognition, 2018.

[86] Jia Xue, Hang Zhang, Ko Nishino, and Kristin J. Dana. Differential Viewpoints for Ground Terrain Material Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1205–1218, 2022.

[87] Tomoaki Yamazaki, Seiya Ito, and Kouzou Ohara. Hierarchical Image Classification with Conceptual Hierarchies Generated via Lexical Databases.

[88] Zhijing Yang, Shujian Lai, Xiaobin Hong, Yukai Shi, Yongqiang Cheng, and Chunmei Qing. DFAEN: Double-order knowledge fusion and attentional encoding network for texture recognition. *Expert Systems with Applications*, 209: 118223, 2022.

[89] Kai Yi, Xiaoqian Shen, Yunhao Gou, and Mohamed Elhoseiny. Exploring Hierarchical Graph Representation for Large-Scale Zero-Shot Image Classification. In *Computer Vision – ECCV 2022*, pages 116–132. Springer Nature Switzerland, Cham, 2022.

[90] Wei Zhai, Yang Cao, Jing Zhang, and Zheng-Jun Zha. Deep Multiple-Attribute-Perceived Network for Real-World Texture Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3612–3621, Seoul, Korea (South), 2019. IEEE.

[91] Wei Zhai, Yang Cao, Zheng-Jun Zha, HaiYong Xie, and Feng Wu. Deep Structure-Revealed Network for Texture Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11010–11019, 2020.

[92] Hang Zhang, Kristin Dana, and Ko Nishino. Reflectance Hashing for Material Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3071–3080, 2015.

[93] Hang Zhang, Jia Xue, and Kristin Dana. Deep TEN: Texture Encoding Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2896–2905, Honolulu, HI, 2017. IEEE.

[94] Yifan Zhao, Ke Yan, Feiyue Huang, and Jia Li. Graph-based High-Order Relation Discovery for Fine-grained Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15074–15083, Nashville, TN, USA, 2021. IEEE.