A Dense Material Segmentation Dataset for Indoor and Outdoor Scene Parsing

Paul Upchurch* and Ransen Niu*

Apple Inc., One Apple Park Way, Cupertino CA, USA

Abstract. A key algorithm for understanding the world is material segmentation, which assigns a label (metal, glass, etc.) to each pixel. We find that a model trained on existing data underperforms in some settings and propose to address this with a large-scale dataset of 3.2 million dense segments on 44,560 indoor and outdoor images, which is 23x more segments than existing data. Our data covers a more diverse set of scenes, objects, viewpoints and materials, and contains a more fair distribution of skin types. We show that a model trained on our data outperforms a state-of-the-art model across datasets and viewpoints. We propose a large-scale scene parsing benchmark and baseline of 0.729 per-pixel accuracy, 0.585 mean class accuracy and 0.420 mean IoU across 46 materials.

1 Introduction

A goal of computer vision is to develop the cognitive ability to plan manipulation of something and predict how it will respond to stimuli. This is informed by the properties of what something is made of. Those properties can be discovered by segmenting a photograph into recognized materials. Material recognition can be understood through the science of material perception starting with Adelson's [1] proposal to divide the world into *things* (countable objects) and *stuff* (materials). Adelson argued stuff is important because of its ubiquity in everyday life. Ritchie *et al.* [25] describe material perception in two parts. The first part is categorical recognition of what something is made of. The second part is recognizing material properties (*e.g.*, glossy, flexible, sound absorbent, sticky) which tells us how something will feel or how it will interact with other objects. While Schwartz *et al.* [30] proposed to recognize properties from local image patches we follow Bell *et al.* [3] who segmented images by recognizing material classes.

Deep learning-based material recognition builds on some key developments. Sharan *et al.* [31] showed that people can recognize 10 kinds of materials in the wild [32] with 85% accuracy. Bell *et al.* [2], following [27], built an efficient annotation tool to create a large-scale material database from crowds and Internet photos. Next, Bell *et al.* [3] introduced large-scale training data and a deep learning approach leading to material segmentation as a building-block for haptics, material assignment, robotic navigation, acoustic simulation and context-aware mixed reality [11,23,29,43,4,8]. Xiao *et al.* [37] introduced a multi-task scene

* These authors contributed equally to this work.

Fig. 1. Densely annotated materials. Our annotations are full-scene, highly detailed and enable prediction of 46 kinds of materials.

parsing model which endows a photograph with a rich prediction of scene type, objects, object parts, materials and textures.

Despite widespread adoption of material segmentation, a lack of large-scale data means evaluation rests on the only large-scale segmentation dataset, OpenSurfaces [2]. We find there is room for improvement and propose the Dense Material Segmentation dataset (DMS) which has 3.2 million segments across 44k densely annotated images, and show empirically that our data leads to models which further close the gap between computer vision and human perception.

There are goals to consider for a material dataset. First, we need a general-purpose set of material labels. We want to mimic human perception so we choose distinguishable materials even if they are of the same type. For example, we separate clear from opaque plastic rather than have a single label for all plastics. We define fine-grained labels which have useful properties, physical or otherwise. For example, a painted whiteboard surface has utility not found in a *paint* label—it is appropriate for writing, cleaning and virtual content display. These functional properties come from how the material is applied rather than its physical structure. Ultimately we choose a set of 52 labels based on prior work and useful materials we found in photographs (details in Section 3.1).

Following [30], we also want indoor and outdoor scenes. Counter-intuitively, this could be unnecessary. Material is recognizable regardless of where it occurs in the world, and deep learning methods aim to create a model which generalizes to unseen cases. Thus, an indoor residential dataset [2] could be sufficient. We find this is not the case. In Section 4.1 we show that a model trained on [2] performs worse on outdoor scenes. This is a key finding which impacts all algorithms which use [2] for training. We also show that a model trained on our dataset is consistent across indoor and outdoor scenes.

We want our database to support many scene parsing tasks so we need broad coverage of objects and scene attributes (which include activities, *e.g.*, eating). In Section 3.2 we show that we achieve better coverage compared to [2].

We propose nine kinds of photographic types which distinguish different viewpoints and circumstances. Our motivation was to quantitatively evaluate cases where we had observed poor performance. This data can reveal new insights on how a model performs. We find that a state-of-the-art model underperforms in some settings whereas a model fit to our data performs well on all nine types.

Our final goal is to have diversity in skin types. Skin is associated with race and ethnicity so it is crucial to have fair representation across different types

Table 1. Large-scale datasets. We propose a dataset with 23x more segments, more classes and 2.3x more images as the largest segment-annotated dataset.

Dataset	Annotation	Classes	Images	Scenes
OpenSurfaces [2]	137k segments	37	19,447	Indoor residential
Materials in Context [3]	3M points	23	436,749	Home interior & exterior
Local Materials [30]	9.4k segments	16	5,845	Indoor & outdoor
DMS (Ours)	3.2M segments	52	44,560	Indoor & outdoor

of skin. We compare our skin type data to OpenSurfaces [2] in Section 3.2 and show our data has practical benefits for training in Section 4.2.

The paper is organized as follows. In Section 2 we review datasets. In Section 3 we describe how we collected data to achieve our goals. In Section 4 we compare our dataset to state-of-the-art data and a state-of-the-art model, study the impact of skin types on training, propose a material segmentation benchmark, and demonstrate material segmentation on real world photos.

In summary, our contributions are:

- We introduce DMS, a large-scale densely-annotated material segmentation dataset and show it is diverse with extensive analysis (Section 3).
- We advance fairness toward skin types in material datasets (Section 3.2).
- We introduce photographic types which reveal new insights on prior work and show that a model fit to our data performs better across datasets and viewpoints compared to the state-of-the-art (Section 4.1).
- We propose a new large-scale indoor and outdoor material segmentation benchmark of 46 materials and present a baseline result (Section 4.3).

2 Related Work

Material Segmentation Datasets. The largest dataset is OpenSurfaces [2] which collected richly annotated polygons of residential indoor surfaces on 19k images, including 37 kinds of materials. The largest material recognition dataset is the Materials in Context Database [3] which is 3M point annotations of 23 kinds of materials across 437k images. This data enables material segmentation by CNN and a dense CRF tuned on OpenSurfaces segments. The Local Materials Database [30] collected segmentations, with the goal of studying materials using only local patches, of 16 kinds of materials across 5,845 images sourced from existing datasets. The Light-Field Material Dataset [35] is 1,200 4D indoor and outdoor images of 12 kinds of materials. The Multi-Illumination dataset [21] captured 1,016 indoor scenes under 25 lighting conditions and annotated the images with 35 kinds of materials. Table 1 lists the largest datasets.

Materials have appeared in purpose-built datasets. The Ground Terrain in Outdoor Scenes (GTOS) database [39] and GTOS-mobile [38] are 30k images of hundreds of instances of 40 kinds of ground materials and 81 videos of 31 kinds of

ground materials, respectively. The Materials in Paintings dataset [34] is bounding box annotations and extracted segmentations on 19k paintings of 15 kinds of materials depicted by artists, partly distinguished into 50 fine-grained categories. COCO-Stuff [6] is segmentations of 91 kinds of stuff on 164k COCO [18] images. While this is a source of material data, it is not a general-purpose material dataset because important surfaces (*e.g.*, objects labeled in COCO) are not assigned material labels. ClearGrasp [28] is a dataset of 50k synthetic and 286 real RGB-D images of glass objects built for robotic manipulation of transparent objects. The Glass Detection Dataset [20] is 3,916 indoor and outdoor images of segmented glass surfaces. The Mirror Segmentation Dataset [41] is 4,018 images with segmented mirror surfaces across indoor and outdoor scenes. Fashionpedia [15] is a database of segmented clothing images of which 10k are annotated with fashion attributes which include fine-grained clothing materials. Figaro [33] is 840 images of people with segmented hair distinguished into 7 kinds of hairstyles.

Categorical Material Names. Bell *et al.* [2] created a set of names by asking annotators to enter free-form labels which were merged into a list of material names. This approach is based on the appearance of surfaces as perceived by the annotators. Schwartz *et al.* [30] created a three-level hierarchy of material names where materials are organized by their physical properties. Some categories were added for materials which could not be placed in the hierarchy. In practice, both approaches resulted in a similar set of entry-level [22] names which also closely agree with prior studies of categorical materials in Internet images [32,14].

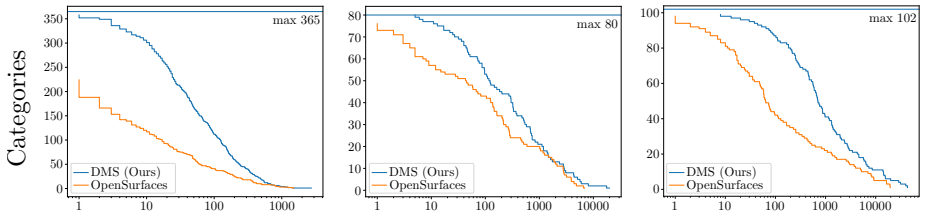
3 Data Collection

DMS is a set of dense polygon annotations of 52 material classes across 44,560 images, which are a subset of OpenImages [17]. We followed a four step process. First, a set of labels was defined. Next, a large set of images was studied by people and algorithms to select images for annotation. Next, the selected images were fully segmented and labeled by a human annotator. Finally, each segmented image was relabeled by multiple people and a final label map was created by fusing all labels. The last three steps were followed multiple times.

3.1 Material Labels

We choose to predefine a label set which is the approach of COCO-Stuff [6]. This encourages annotators to create consistent labels suitable for machine learning. We instructed annotators to assign *not on list* to recognized materials which do not fit in any category and *I cannot tell* to unknown and unrecognizable surfaces (*e.g.*, watermarks and under-/over-saturated pixels).

We defined a label set based on appearance, which is the approach of OpenSurfaces [2]. A label can represent a solid substance (*e.g.*, wood), a distinctive arrangement of substances (*e.g.*, brickwork), a liquid (*e.g.*, water) or a useful non-material (*e.g.*, sky). We used 35 labels from OpenSurfaces and *asphalt* from [30].



(a) Scene Occurrences (b) Object Occurrences (c) Attribute Occurrences

Fig. 2. Image diversity. We plot number of categories (y-axis) vs. occurrence in images (log-scale x-axis) of Places365 scene type (a), COCO objects (b), and SUN attributes (c). Our dataset (*blue*) is larger, more diverse and more balanced across categories (higher slope) compared to the largest segmentation dataset (*orange*).

We added 2 fine-grained people and animal categories (*bone* and *animal skin*). We introduced 3 labels for workplaces (*ceiling tile*, *whiteboard* and *fiberglass wool*), 6 for indoor scenes (*artwork*, *clutter*, *non-water liquid*, *soap*, *pearl* and *gemstone*) and 4 for outdoors (*sand*, *snow*, *ice* and *tree wood*). *Artwork* identifies an imitative surface which is photographic or fine art—affording further analysis by Materials In Paintings [34]. *Clutter* is a region of visually indistinguishable manufactured stuff (typically a mixture of metal, plastic and paper) which occurs in trash piles. Lastly, we defined a label called *engineered stone* for artificial surfaces which imitate stone, which includes untextured and laminated solid surfaces. See Figure 4 for an example of each label.

3.2 Image Selection

Bell *et al.* [3] found that a balanced set of material labels can achieve nearly the same performance as a 9x larger imbalanced set. Since we collect dense annotations we cannot directly balance classes. Instead, we searched 191k images for rare materials and assumed common materials would co-occur. Furthermore, we ran Detectron [12] to detect COCO [18] objects, and Places365 [45] to classify scenes and recognize SUN [24] attributes. EXIF information was used to infer country. These detections were used to select images of underrepresented scenes, objects and countries. Figure 2 compares the diversity of the 45k images in DMS to the 19k images in OpenSurfaces by a plot of the number of categories, y, which have at least x occurrences. Occurrences of scene type, object and SUN attribute are plotted. Note that the x-axis is logarithmic scale. We find our dataset is more diverse having more classes present in greater amounts (more than can be explained by the 2.24x difference in image count).

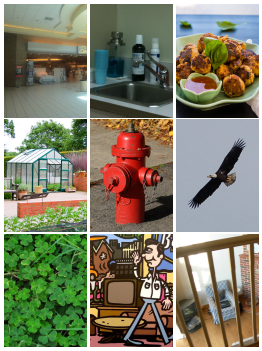
We balance the distribution of skin appearance in DMS so that algorithms trained with our data perform well on all kinds of skin [5]. We use Fitzpatrick [10] skin type to categorize skin into 3 groups, inspired by an approach used by [40]. We ran the DLIB [16] face detector and labeled a subset of the faces. Our 157 manual annotations were used to calibrate a preexisting face attribute predictor

Table 2. Skin types. We report estimated occurrences. Our dataset has 12x more occurrences of the smallest group and 4.8x more fair representation by ratio.

	OpenSurfaces	DMS (Ours)
Type I-II (light)	2,332	4,535
Type III-IV (medium)	3,889	9,776
Type V-VI (dark)	375	5,899
Ratio of largest to smallest group	10.37 : 1	2.16 : 1

Table 3. Photographic types. Our data contains indoor views (*top*), outdoor views (*middle*), and close-up and unusual views (*bottom*).

Photographic Type	Images
An area with visible enclosure	16,013
A collection of indoor things	6,064
A tightly cropped indoor thing	2,634
A ground-level view of reachable outdoor things	3,127
A tightly cropped outdoor thing	1,196
Distant unreachable outdoor things	971
A real surface without context	847
Not a real photo	805
An obstructed or distorted view	204



(trained on a different dataset) which was then used to predict skin types for the rest of DMS. We found that the ratio of the largest group to the smallest was 9.4. Next, we selected images which would increase the most underrepresented skin type group and found this reduced the ratio to 2.2. We calibrated the same detector for OpenSurfaces faces and measured its ratio as 10.4. According to the findings of [5], we expect skin classifiers trained on OpenSurfaces would underperform on dark skin. Table 2 shows the distribution of skin types.

We used Places365 scene type detection to select outdoor images but we found this did not lead to outdoor materials. We took two steps to address this. First, we annotated our images with one of nine *photographic types* which distinguish outdoor from indoor from unreal images. Table 3 shows the annotated types. Next, we used these labels to select outdoor scenes and underrepresented viewpoints. This was effective—growing the dataset by 17% more than doubled 9 kinds of outdoor materials: *ice* (3x), *sand* (4.4x), *sky* (8x), *snow* (9.5x), *soil* (3x), *natural stone* (2.4x), *water* (2.5x), *tree wood* (2.3x) and *asphalt* (9.2x).

3.3 Segmentation and Instances

Images were given to annotators for polygon segmentation of the entire image. We instructed annotators to segment parts larger than a fingertip, ignore gaps smaller than a finger, and to follow material boundaries tightly while ignoring

Table 4. Annotator agreement rates. High rates indicate consistent label assignment. Low rates indicate disagreement, confusion or unstructured error.

Hair	0.95	Glass	0.80	Wood	0.67	Non-clear plastic	0.60
Skin	0.93	Paper	0.76	Tree wood	0.66	Leather	0.53
Foliage	0.86	Carpet/rug	0.73	Tile	0.66	Cardboard	0.53
Sky	0.86	Nat. stone	0.72	Metal	0.65	Artwork	0.51
Food	0.84	Ceramic	0.70	Paint/plaster	0.62	Clear plastic	0.50
Fabric/cloth	0.82	Mirror	0.68	Rubber	0.61	Concrete	0.45

geometry and shadow boundaries. Following [2], annotators were instructed to segment glass and mirror surfaces rather than the covered or reflected surfaces. Unreal elements such as borders and watermarks were segmented separately. Images with objectionable content (*e.g.*, violence) were not annotated.

Annotators segmented resized images, with median longest edge of 1024 pixels, creating over 3.2 million segments (counting only those larger than 100 pixels) with a mean of 72 segments per image. The created segments are detailed—wires, jewelry, teeth, eyebrows, shoe soles, wheel rims, door hinges, clasps, buttons and latches are some of the small and thin materials segmented separately. See Figure 1 and Figure 3 for examples of detailed segmentations.

We defined a material instance as materials of the same type from the same manufacturing source. For example a wooden cabinet should be segmented separately from a wood floor but the planks making up a single-source floor would be one instance. DMS is the first large-scale densely segmented dataset to have detailed material instances.

3.4 Labeling

The annotator who segmented an image also assigned labels based on their judgment and our instruction. We found that surfaces coated with another material or colored by absorbing ink required clarification. Appearance-changing coatings were labeled *paint* while clear or appearance-enhancing coatings (*e.g.*, varnish, cosmetics, sheer hosiery) were labeled as the underlying material. Small amounts of ink (*e.g.*, printed text) are disregarded. Some surfaces imitate the appearance of other materials (*e.g.*, laminate). High-quality imitations were labeled as the imitated material and low-quality imitations as the real material.

Our instructions were refined in each iteration and incorrect labels from early iterations were corrected. Some cases needed special instruction. We instructed annotators to label electronic displays as *glass* and vinyl projection screens as *not on list*. Uncovered artwork or photographs were to be labeled *artwork* while glass-covered art should be labeled *glass*. In ambiguous cases, we assume framed artwork has a glass cover. *Sky* includes day sky, night sky and aerial phenomenon (*e.g.*, clouds, stars, moon, and sun).

We collected more opinions by presenting a segmentation, after removing labels, to a different annotator who relabeled the segments. The relabeling annotator could fix bad segments by adjusting polygons or assign special labels to

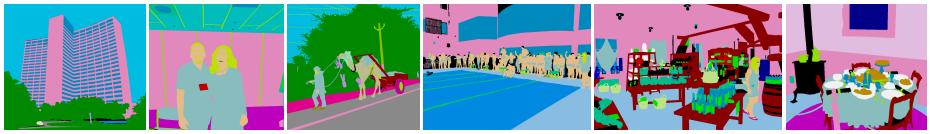


Fig. 3. Fused labels. We show segmentation quality and variety of scenes, activities and materials (*left to right:* building exterior, workplace, road, swimming pool, shop, dining room). See Table 5 for color legend. Black pixels are unlabeled (no consensus).

Table 5. Material occurrence in images. We report the number of images in which a label occurs. The colors are used for visualizations.

Paint/plaster	39,323	Sky	3,306	Chalkboard	668
Fabric/cloth	31,489	Mirror	3,242	Asphalt	474
Non-clear plas	30,506	Cardboard	3,150	Fire	412
Metal	30,504	Food	2,908	Gemstone	369
Glass	28,934	Concrete	2,853	Sponge	326
Wood	24,248	Ceiling tile	2,524	Eng. stone	299
Paper	20,763	Natural stone	2,076	Liquid	294
Skin	18,524	Water	2,063	Pearl	282
Hair	17,766	Tree wood	2,026	Cork	273
Foliage	11,384	Wicker	1,895	Sand	272
Tile	10,173	Soil/mud	1,855	Snow	191
Carpet/rug	9,516	Pol. stone	1,831	Soap	154
Ceramic	8,314	Brickwork	1,654	Clutter	128
Rubber	7,811	Fur	1,567	Ice	96
Leather	7,354	Whiteboard	1,171	Styrofoam	88
Clear plastic	6,431	Wax	1,107	Fiberglass wool	33
Artwork	4,344	Wallpaper	1,076		
Bone/horn	3,751	Animal skin	1,007		

indicate a segment does not follow boundaries or is made of multiple material types. We collected 98,526 opinions across 44,560 images consisting of 8.2 million segment labels (counting only segments larger than 100 pixels).

We studied label agreement by counting occurrences of a segment label and matching pixel-wise dominant label by a different annotator. We found an agreement rate of 0.675. In cases of agreement, 8.9% were unrecognizable (*I cannot tell*) and 0.6% were *not on list*. Table 4 shows the agreement rate for classes larger than the median number of segments per class. Among the largest classes the most agreed-upon labels are *hair*, *skin*, *foliage*, *sky*, and *food*. We only analyze the largest classes since unstructured error (*e.g.*, misclicks) can overwhelm the statistics of small classes, which are up to 2,720 times smaller.

3.5 Label Fusion

Each annotator's segments are rendered to create a label map. Label maps were inspected for correctness and we fixed incorrect labels in 1,803 images. Next,

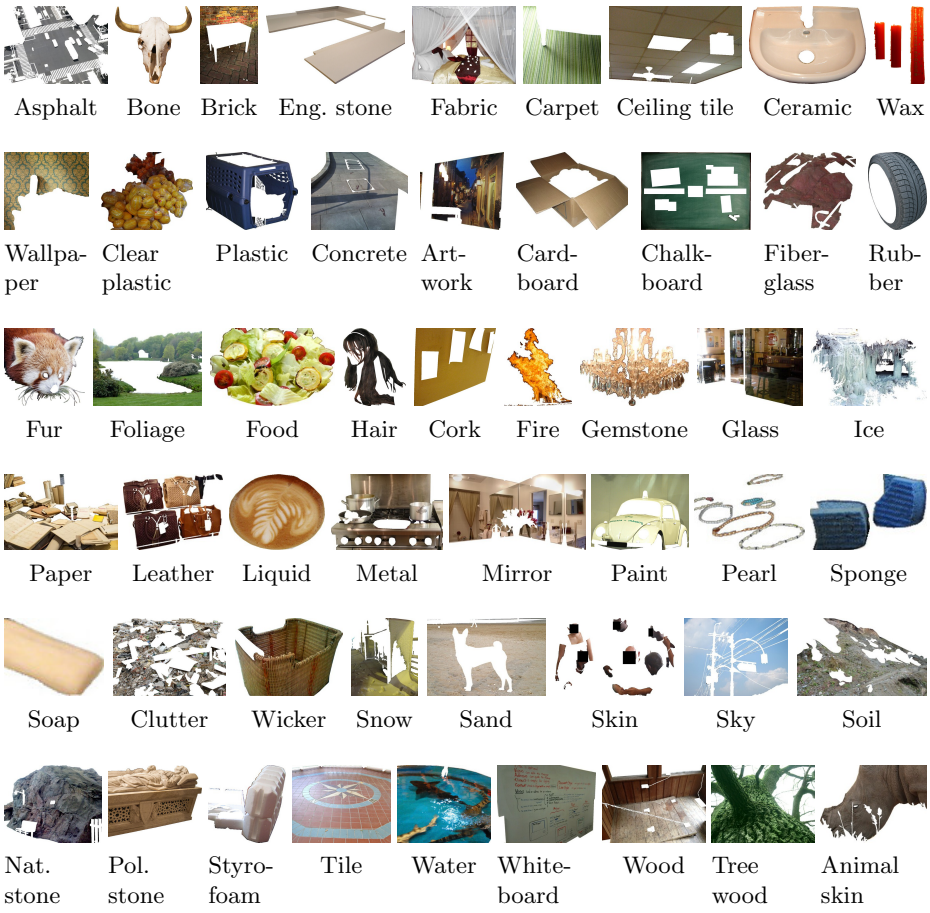


Fig. 4. Material labels. For each label we show a cut-out example.

we create a single *fused label map* for each image. First, we combined label maps pixel-wise by taking the strict majority label. Next, we overlaid manual corrections and reassigned non-semantic labels (*e.g.*, *I cannot tell*) to *no label*. The fused maps have a mean labeled area fraction of 0.784. For comparison, we created fused label maps for OpenSurfaces and found its density is 0.210. DMS is 2.3x larger and 3.7x denser, which is 8.4x more labeled area. Compared to the 3M points in MINC [3], DMS has 3.2M fused segments which carry more information about shape, boundary and co-occurrences. While MINC annotations span 10x more images, point annotations cannot evaluate segmentation boundaries for scene parsing tasks. Example fused maps and class occurrences are shown in Figure 3 and Table 5. The smallest class appears in 33 images whereas the largest class, *paint*, appears in 39,323 images, which is 88% of the images.

4 Experiments

First, we investigate the impact of our data on training deep learning models with a cross-dataset comparison (Section 4.1). Then, we compare the impact of skin type distributions on fairness of skin recognition (Section 4.2). Next, we establish a material segmentation benchmark for 46 kinds of materials (Section 4.3). Finally, we show predictions on real world images (Section 4.4).

Splits. We created train, validation and test splits for our data by assigning images according to material occurrence. The smallest classes are assigned a ratio of $1:1:1$, which increases to $2.5:1:1$ for the largest. An image assignment impacts the ratio of multiple classes so small classes are assigned first. There are 24,255 training images, 10,139 validation images and 10,166 test images.

4.1 Cross-Dataset Comparison

Does training with our data lead to a better model? This experiment compares a model fit to our data against two baselines fit to OpenSurfaces data—the strongest published model [37] and a model with the same architecture as ours. There are two sources of data. The first is OpenSurfaces data with the splits and 25 labels proposed by [37]. The second is comparable DMS training and validation data ([37] does not define a test split) created by translating our labels to match [37]. The evaluation set, which we call Avg-Val, is made of both parts—the validation sets of OpenSurfaces and DMS, called OS-Val and DMS-Val, respectively—weighted equally. For evaluation of our data we fit models to DMS training data and choose the model that performs best on DMS-Val. This model, which we call DMS-25, is a ResNet-50 architecture [13] with dilated convolutions [7,42] as the encoder, and Pyramid Pooling Module from PSPNet [44] as the decoder. The first baseline (Table 6, row 2) is UPerNet [37], a multitask scene parsing model which uses cross-domain knowledge to boost material segmentation performance. The second baseline (Table 6, row 3), called OS-25, has the same architecture as DMS-25 but is fit to OpenSurfaces training data. Table 6 shows the results. We report per-pixel accuracy (Acc), mean class accuracy (mAcc), mean intersection-over-union (mIoU) and Δ, the absolute difference in a metric across DMS-Val and OS-Val. A low Δ indicates a model is more consistent across datasets. We find that fitting a model to DMS training data leads to higher performance and lower Δ on all metrics. We also report the metrics on each validation set and find that both baselines underperform on DMS-Val. We find that DMS-25 performs 0.01 lower on OS-Val mAcc compared to a model trained on OpenSurfaces data. This may be due to differences in annotation and image variety. We use our photographic type labels to investigate the larger performance gaps on DMS-Val.

Why do models trained with OpenSurfaces underperform on our validation images? In Table 7 we report per-pixel accuracy of DMS-25, UPerNet, and OS-25 across nine categories. We find that DMS-25 performs consistently across categories with the lowest performing category (unreal images) 0.071 below the highest performing category (images of enclosed areas). UPerNet shows lower

Table 6. Training data evaluation. We compare segmentation of 25 materials with our training data (*row 1*) to OpenSurfaces data with two kinds of models (*rows 2 and 3*). Avg-Val is the equally-weighted validation sets of each dataset, DMS-Val and OS-Val. Δ is the difference in a metric across datasets. A convnet fit to our data achieves higher performance and is more consistent across datasets.

Training data	Model	Metric	Avg-Val ↑	Δ ↓	DMS-Val ↑	OS-Val ↑
DMS (Ours)	DMS-25	Acc	**0.777**	**0.047**	0.753	0.800
		mAcc	**0.689**	**0.006**	0.686	0.692
		mIoU	**0.500**	**0.014**	0.507	0.493
OpenSurfaces [2]	UPerNet [37]	Acc	0.682	0.310	0.527	0.837
		mAcc	0.486	0.274	0.349	0.623
		mIoU	0.379	0.298	0.230	0.528
OpenSurfaces [2]	OS-25	Acc	0.705	0.231	0.589	0.820
		mAcc	0.606	0.193	0.509	0.702
		mIoU	0.416	0.199	0.316	0.515

performance across all categories with a drop of 0.426 from images of enclosed areas to images of distant outdoor things. And OS-25 shows similar performance with a drop of 0.407. We observe that both UPerNet and OS-25 have low performance on outdoor images and images without any context. This study shows that photographic types can improve our understanding of how material segmentation models perform in different settings. And, these results justify our decision to collect outdoor images and images of different photographic types.

4.2 Recognition of Different Skin Types

Models trained on face datasets composed of unbalanced skin types exhibit classification disparities [5]. Does this impact skin recognition? Without any corrections for skin type imbalance we find that DMS-25 has a 3% accuracy gap among different skin types on DMS-val (Type I-II: 0.933, Type III-IV: 0.924, Type V-VI: 0.903) while OS-25 has a larger gap of 13.3% (Type I-II: 0.627, Type III-IV: 0.571, Type V-VI: 0.494). This confirms that skin type imbalance impacts skin recognition. Our contribution lies in providing more data for all skin types (Table 2), which makes it easier for practitioners to create fair models.

4.3 A Material Segmentation Benchmark

It is common practice to select large categories and combine smaller ones (our smallest occurs in only 12 training images) for a benchmark. Yet, we cannot know *a priori* how much training data is sufficient to learn a category. We choose to be guided by the validation data. We fit many models to all 52 categories then inspect the results to determine which categories can be reliably learned. We select ResNet50 [13] with dilated convolutions [7,42] as the encoder, and Pyramid

Table 7. Performance analysis with photographic types. A model fit to our data, DMS-25 (*Table 6, row 1*), performs well on all photographic types whereas two models fit to OpenSurfaces, UPerNet and OS-25 (*Table 6, rows 2-3*) have low performance outdoors (*middle*) and on surfaces without any context (*row 7*).

Photographic Type	Per-Pixel Accuracy		
	DMS-25 (Ours)	UPerNet [37]	OS-25
An area with visible enclosure	0.756	0.615	0.632
A collection of indoor things	0.752	0.546	0.622
A tightly cropped indoor thing	0.710	0.441	0.561
A view of reachable outdoor things	0.750	0.265	0.388
A tightly cropped outdoor thing	0.731	0.221	0.359
Distant unreachable outdoor things	0.736	0.189	0.225
A real surface without context	0.691	0.222	0.348
Not a real photo	0.685	0.528	0.551
An obstructed or distorted view	0.729	0.370	0.496

Pooling Module from PSPNet [44] as the decoder. We choose this architecture because it has been shown to be effective for scene parsing [44,47]. Our best model, which we call DMS-52, predicts 52 materials with per-pixel accuracy 0.735, mean class accuracy 0.535 and mIoU 0.392 on DMS-val.

We inspected a few strongest DMS-52 fitted models and found that 6 categories consistently stood out as underperforming—having 0 accuracy in some cases and, at best, not much higher than chance. Those categories are *non-water liquid*, *fiberglass*, *sponge*, *pearl*, *soap* and *styrofoam*, which occur in 129, 12, 149, 129, 58 and 33 training images, respectively. Guided by this discovery we select the other 46 material labels for a benchmark.

We train a model, called DMS-46, to predict the selected categories, with the same architecture as DMS-52. We use a batch size of 64 and stochastic gradient descent optimizer with 1e-3 base learning rate and 1e-4 weight decay. We use ImageNet pretraining [46,47] to initialize the encoder weights, and scale the learning rate for the encoder by 0.25. We update the learning rate with a cosine annealing schedule with warm restart [19] every 30 epochs for 60 epochs. Because the classes are imbalanced we use weighted symmetric cross entropy [36], computed across DMS training images, as the loss function, which gives more weight to classes with fewer ground truth pixels. We apply stochastic transformations for data augmentation (scale, horizontal and vertical flips, color jitter, Gaussian noise, Gaussian blur, rotation and crop), scale inputs into [0, 1], and normalize with mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225] from ImageNet [9]. The training tensor has height and width of 512.

DMS-46 predicts 46 materials with per-pixel accuracy 0.731/0.729, mean class accuracy 0.598/0.585 and mIoU 0.435/0.420 on DMS-val/DMS-test respectively. We report the test set per-class accuracy and IoU in Table 8. We find that *sky*, *fur*, *foliage*, *skin* and *hair* have the highest recognition rates, similar to

Table 8. Test set results. We report metrics for our model, DMS-46. 17 materials, in italics, are new—not predicted by prior general-purpose models [3,37,30].

Category	Acc	IoU	Category	Acc	IoU	Category	Acc	IoU
Sky	0.962	0.892	*Chalkboard*	0.712	0.548	*Artwork*	0.454	0.301
Fur	0.910	0.707	Paint/plaster	0.694	0.632	Mirror	0.452	0.278
Foliage	0.902	0.761	Wicker	0.674	0.460	*Sand*	0.444	0.340
Skin	0.886	0.640	Natural stone	0.665	0.436	*Ice*	0.440	0.362
Hair	0.881	0.673	Glass	0.653	0.483	*Tree wood*	0.428	0.261
Food	0.868	0.668	Asphalt	0.628	0.442	Pol. stone	0.379	0.236
Ceiling tile	0.867	0.611	Leather	0.615	0.373	*Clear plastic*	0.360	0.222
Water	0.866	0.712	*Snow*	0.610	0.465	Rubber	0.255	0.163
Carpet/rug	0.849	0.592	Concrete	0.603	0.304	*Clutter*	0.182	0.152
Whiteboard	0.838	0.506	Metal	0.575	0.303	*Fire*	0.176	0.147
Fabric/cloth	0.801	0.692	*Wax*	0.573	0.371	*Gemstone*	0.116	0.096
Wood	0.797	0.635	Cardboard	0.570	0.363	Eng. stone	0.088	0.071
Ceramic	0.757	0.427	Wallpaper	0.544	0.329	*Cork*	0.082	0.066
Brickwork	0.746	0.491	*Non-clear plastic*	0.519	0.321	*Bone/horn*	0.074	0.070
Paper	0.729	0.508	Soil/mud	0.511	0.332			
Tile	0.722	0.550	*Animal skin*	0.472	0.308			

the findings of [3]. 17 materials do not appear in any prior large-scale material benchmarks. Among these new materials we report high recognition rates for *ceiling tile*, *whiteboard* and *chalkboard*. To our knowledge, DMS-46 is the first material segmentation model evaluated on large-scale dense segmentations and predicts more classes than any general-purpose model.

4.4 Real-World Examples

In Figure 5 we demonstrate DMS-46 on indoor and outdoor photos from daily life. Our model recognizes and localizes *food* on *ceramic* plates, workplace materials (*whiteboard* and *ceiling tile*), ground cover materials (*soil*, *stone*, *foliage* and *snow*), unprocessed *tree wood*, and *fire* on a *wax* candle.

A Failure Case. The last image is a failure case where our model is confused by decorative tile artwork. We also see opportunities for further improving boundaries and localizing small surfaces.

5 Discussion and Conclusion

Dense Annotation. Prior works [2,3,30] instruct annotators to locate and segment regions made of a given material. Our approach is different. We instruct annotators to segment and label the entire image. This approach collects different data because annotators address all surfaces—not just those which are readily recognized. We hypothesize this creates a more difficult dataset, and propose this approach is necessary for evaluation of scene parsing, which predicts all pixels.

Fig. 5. Real-world examples. Our model, DMS-46, predicts 46 kinds of indoor and outdoor materials. See Table 5 for color legend.

Real vs. Synthetic. Synthetic data has achieved high levels of realism (*e.g.*, Hypersim [26]) and may be a valuable generator of training data. We opted to label real photos because models trained on synthetic data need a real evaluation dataset to confirm the domain gap from synthetic to real has been bridged.

Privacy. Material predictions can be personal. Knowing a limb is not made of skin reveals a prosthetic. The amount of body hair reveals one aspect of appearance. Precious materials in a home reveals socio-economic status. Clothing material indicates degree of nakedness. Care is needed if material segmentation is tied to identity. Limiting predicted materials to only those needed by an application or separating personal materials from identity are two ways, among many possible ways, to strengthen privacy and protect personal information.

6 Conclusion

We present the first large-scale densely-annotated material segmentation dataset which can train or evaluate indoor and outdoor scene parsing models. [1] We propose a benchmark on 46 kinds of materials. Our data can be a foundation for algorithms which utilize material type, make use of physical properties for simulations or functional properties for planning and human-computer interactions. We look forward to expanding the number of materials, finding new methods to reach even better full-scene material segmentation, and combining the point-wise annotations of MINC [3] with our data in future work.

Acknowledgements. We thank Allison Vanderby, Hillary Strickland, Laura Snarr, Mya Exum, Subhash Sudan, Sneha Deshpande, and Doris Guo for their help with acquiring data; Richard Gass, Daniel Kurz and Selim Ben Himane for their support.

[1] Our data is available at https://github.com/apple/ml-dms-dataset.

References

1. Adelson, E.H.: On seeing stuff: The perception of materials by humans and machines. In: Human vision and electronic imaging VI. vol. 4299, pp. 1–12. SPIE (2001)
2. Bell, S., Upchurch, P., Snavely, N., Bala, K.: OpenSurfaces: A richly annotated catalog of surface appearance. ACM Transactions on graphics (TOG) **32**(4), 1–17 (2013)
3. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the Materials in Context database. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3479–3487 (2015)
4. Brandao, M., Shiguematsu, Y.M., Hashimoto, K., Takanishi, A.: Material recognition CNNs and hierarchical planning for biped robot locomotion on slippery terrain. In: 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). pp. 81–88. IEEE (2016)
5. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018)
6. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
8. Chen, L., Tang, W., John, N.W., Wan, T.R., Zhang, J.J.: Context-aware mixed reality: A learning-based framework for semantic-level interaction. In: Computer Graphics Forum. vol. 39, pp. 484–496. Wiley Online Library (2020)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types I through VI. Archives of dermatology **124**(6), 869–871 (1988)
11. Gao, Y., Hendricks, L.A., Kuchenbecker, K.J., Darrell, T.: Deep learning for tactile understanding from visual and haptic data. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 536–543. IEEE (2016)
12. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. `https://github.com/facebookresearch/detectron` (2018)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Hu, D., Bo, L., Ren, X.: Toward robust material recognition for everyday objects. In: BMVC. vol. 2, p. 6. Citeseer (2011)
15. Jia, M., Shi, M., Sirotenko, M., Cui, Y., Cardie, C., Hariharan, B., Adam, H., Belongie, S.: Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In: European conference on computer vision. pp. 316–332. Springer (2020)
16. King, D.E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research **10**, 1755–1758 (2009)
17. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J.,

Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: OpenImages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html (2017)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
19. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)
20. Mei, H., Yang, X., Wang, Y., Liu, Y., He, S., Zhang, Q., Wei, X., Lau, R.W.: Don't hit me! glass detection in real-world scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3687–3696 (2020)
21. Murmann, L., Gharbi, M., Aittala, M., Durand, F.: A dataset of multi-illumination images in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4080–4089 (2019)
22. Ordonez, V., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: From large scale image categorization to entry-level categories. In: Proceedings of the ieee international conference on computer vision. pp. 2768–2775 (2013)
23. Park, K., Rematas, K., Farhadi, A., Seitz, S.M.: PhotoShape: Photorealistic materials for large-scale shape collections. ACM Trans. Graph. **37**(6) (Nov 2018)
24. Patterson, G., Hays, J.: SUN attribute database: Discovering, annotating, and recognizing scene attributes. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2751–2758. IEEE (2012)
25. Ritchie, J.B., Paulun, V.C., Storrs, K.R., Fleming, R.W.: Material perception for philosophers. Philosophy Compass **16**(10), e12777 (2021)
26. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10912–10922 (2021)
27. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: A database and web-based tool for image annotation. International journal of computer vision **77**(1), 157–173 (2008)
28. Sajjan, S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., Song, S.: ClearGrasp: 3D shape estimation of transparent objects for manipulation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 3634–3642. IEEE (2020)
29. Schissler, C., Loftin, C., Manocha, D.: Acoustic classification and optimization for multi-modal rendering of real-world scenes. IEEE transactions on visualization and computer graphics **24**(3), 1246–1259 (2017)
30. Schwartz, G., Nishino, K.: Recognizing material properties from images. IEEE transactions on pattern analysis and machine intelligence **42**(8), 1981–1995 (2019)
31. Sharan, L., Liu, C., Rosenholtz, R., Adelson, E.H.: Recognizing materials using perceptually inspired features. International journal of computer vision **103**(3), 348–371 (2013)
32. Sharan, L., Rosenholtz, R., Adelson, E.H.: Accuracy and speed of material categorization in real-world images. Journal of vision **14**(9), 12–12 (2014)
33. Svanera, M., Muhammad, U.R., Leonardi, R., Benini, S.: Figaro, hair detection and segmentation in the wild. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 933–937. IEEE (2016)

34. Van Zuijlen, M.J., Lin, H., Bala, K., Pont, S.C., Wijntjes, M.W.: Materials in Paintings (MIP): An interdisciplinary dataset for perception, art history, and computer vision. Plos one **16**(8), e0255109 (2021)
35. Wang, T.C., Zhu, J.Y., Hiroaki, E., Chandraker, M., Efros, A.A., Ramamoorthi, R.: A 4D light-field dataset and CNN architectures for material recognition. In: European conference on computer vision. pp. 121–138. Springer (2016)
36. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 322–330 (2019)
37. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)
38. Xue, J., Zhang, H., Dana, K.: Deep texture manifold for ground terrain recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 558–567 (2018)
39. Xue, J., Zhang, H., Dana, K., Nishino, K.: Differential angular imaging for material recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 764–773 (2017)
40. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 547–558 (2020)
41. Yang, X., Mei, H., Xu, K., Wei, X., Yin, B., Lau, R.W.: Where is my mirror? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8809–8818 (2019)
42. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations (2016)
43. Zhao, C., Sun, L., Stolkin, R.: A fully end-to-end deep learning approach for real-time simultaneous 3D reconstruction and material recognition. In: 2017 18th International Conference on Advanced Robotics (ICAR). pp. 75–82. IEEE (2017)
44. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
45. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
46. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
47. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)