

# A Dense Material Segmentation Dataset for Indoor and Outdoor Scene Parsing

Paul Upchurch\* and Ransen Niu\*

Apple Inc.

**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].

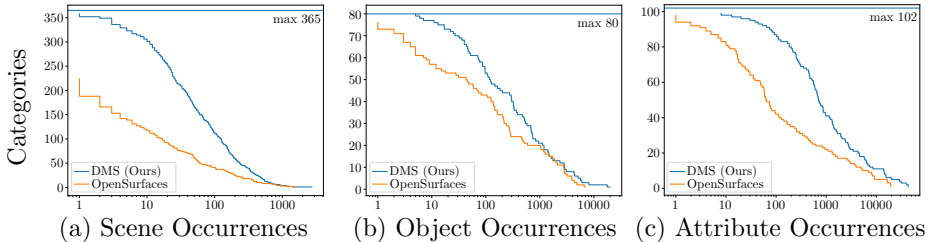
### 3 Data Collection

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].



**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.

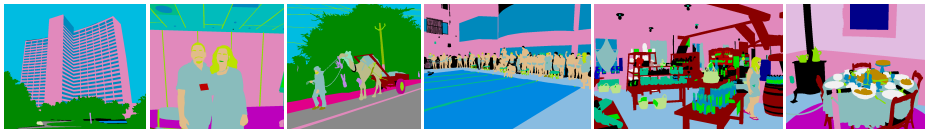
We defined a material instance as materials of the same type from the same manufacturing source. For example a wooden cabinet should be segmented separately from a wood floor but the planks making up a single-source floor would be one instance. DMS is the first large-scale densely segmented dataset to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

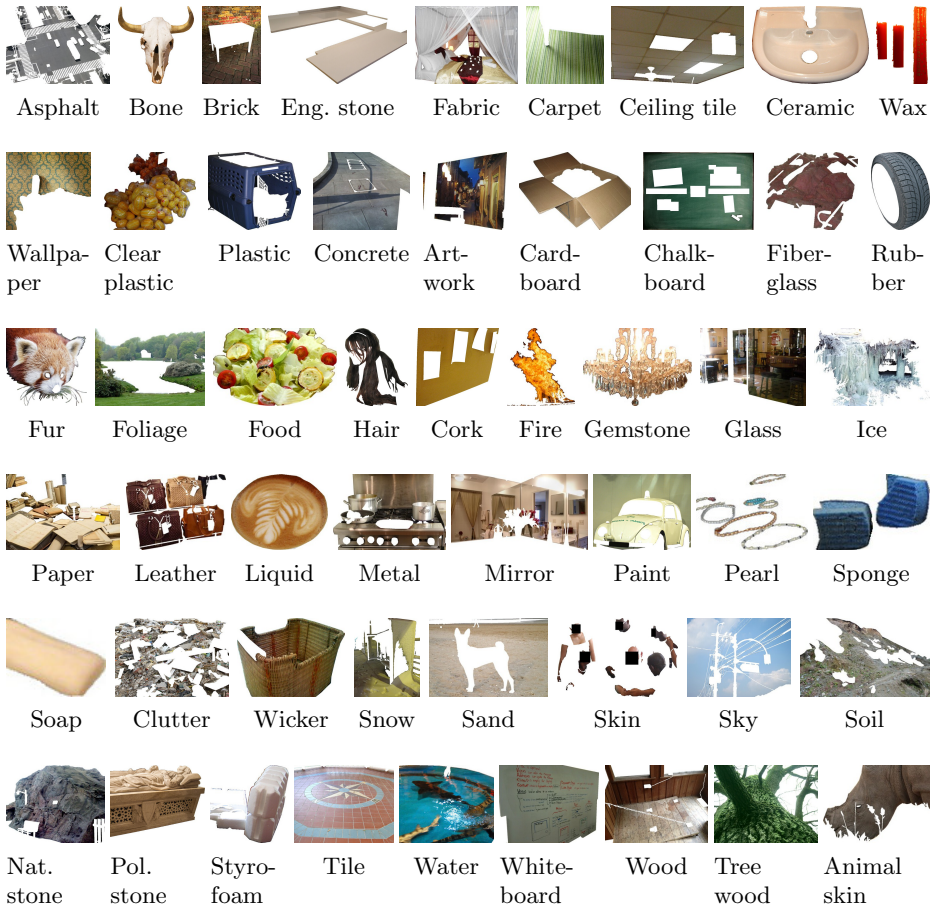
indicate a segment does not follow boundaries or is made of multiple material types. We collected 98,526 opinions across 44,560 images consisting of 8.2 million segment labels (counting only segments larger than 100 pixels).

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  $\Delta$ , the absolute difference in a metric across DMS-Val and OS-Val. A low  $\Delta$  indicates a model is more consistent across datasets. We find that fitting a model to DMS training data leads to higher performance and lower  $\Delta$  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.  $\Delta$  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 $\uparrow$	$\Delta$ $\downarrow$	DMS-Val $\uparrow$	OS-Val $\uparrow$
DMS (Ours)	DMS-25	Acc	<b>0.777</b>	<b>0.047</b>	0.753	0.800
		mAcc	<b>0.689</b>	<b>0.006</b>	0.686	0.692
		mIoU	<b>0.500</b>	<b>0.014</b>	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	<i>Chalkboard</i>	0.712	0.548	<i>Artwork</i>	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	<i>Sand</i>	0.444	0.340
Skin	0.886	0.640	Natural stone	0.665	0.436	<i>Ice</i>	0.440	0.362
Hair	0.881	0.673	Glass	0.653	0.483	<i>Tree wood</i>	0.428	0.261
Food	0.868	0.668	Asphalt	0.628	0.442	Pol. stone	0.379	0.236
<i>Ceiling tile</i>	0.867	0.611	Leather	0.615	0.373	<i>Clear plastic</i>	0.360	0.222
Water	0.866	0.712	<i>Snow</i>	0.610	0.465	Rubber	0.255	0.163
Carpet/rug	0.849	0.592	Concrete	0.603	0.304	<i>Clutter</i>	0.182	0.152
<i>Whiteboard</i>	0.838	0.506	Metal	0.575	0.303	<i>Fire</i>	0.176	0.147
Fabric/cloth	0.801	0.692	<i>Wax</i>	0.573	0.371	<i>Gemstone</i>	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	<i>Cork</i>	0.082	0.066
Brickwork	0.746	0.491	<i>Non-clear plastic</i>	0.519	0.321	<i>Bone/horn</i>	0.074	0.070
Paper	0.729	0.508	Soil/mud	0.511	0.332			
Tile	0.722	0.550	<i>Animal skin</i>	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.

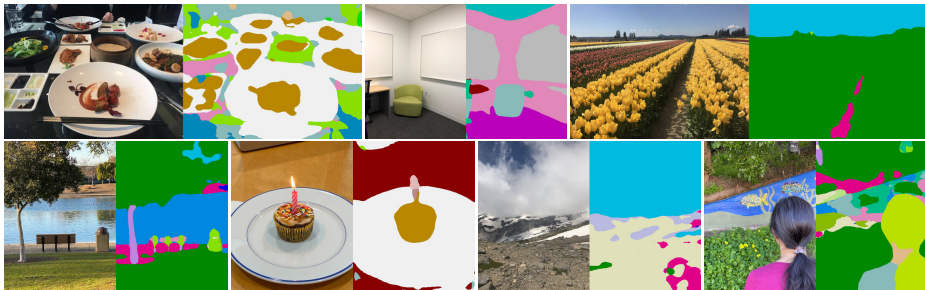
#### 4.4 Real-World Examples

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.<sup>1</sup> 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.

<sup>1</sup> 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., Mallocci, 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.: Clear-Grasp: 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)
48. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: ENet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147* (2016)

## A Dataset Details

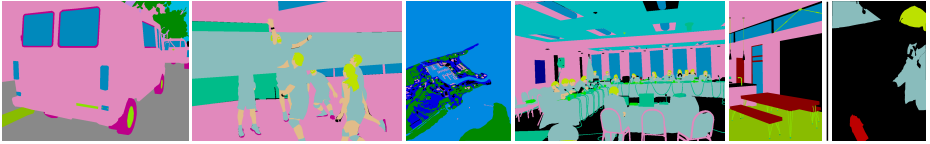
In this section we supplement Section 3 of the main paper.

In Table 9 we list names used in annotation tools. For brevity, names in the main paper are shortened and “Photograph/painting” is called *artwork*. We also report the number of images in which a material occurs and total area, the sum over all images of the fraction of pixels covered by a material.

In Table 10 we show the number of annotated pixels for each class. This count is according to the resized images which are smaller than the original images.

Table 9: **Material occurence.** We report the number of images and total area (in units of image proportion, rounded).

	Image Count				Total Area			
	All	Train	Val	Test	All	Train	Val	Test
Animal skin	1,007	479	260	268	34	14	8	11
Bone/teeth/horn	3,751	2,084	858	809	4	2	1	2
Brickwork	1,654	862	388	404	204	113	46	44
Cardboard	3,150	1,773	681	696	133	73	30	30
Carpet/rug	9,516	5,470	2,073	1,973	985	567	208	209
Ceiling tile	2,524	1,460	529	535	299	173	65	61
Ceramic	8,314	4,608	1,854	1,852	260	135	69	56
Chalkboard/blackboard	668	332	166	170	68	34	16	19
Clutter	128	41	43	44	12	3	5	5
Concrete	2,853	1,381	731	741	400	186	109	105
Cork/corkboard	273	122	78	73	9	4	2	3
Engineered stone	299	134	81	84	18	8	5	5
Fabric/cloth	31,489	17,727	6,875	6,887	4,799	2,732	1,038	1,030
Fiberglass wool	33	12	9	12	3	1	1	1
Fire	412	184	110	118	12	5	4	3
Foliage	11,384	5,902	2,714	2,768	1,377	640	372	364
Food	2,908	1,553	687	668	287	126	82	79
Fur	1,567	761	398	408	206	95	55	55
Gemstone/quartz	369	165	99	105	10	5	2	3
Glass	28,934	16,142	6,378	6,414	2,159	1,192	488	479
Hair	17,766	10,076	3,823	3,867	336	190	74	72
Ice	96	31	32	33	27	10	8	8
Leather	7,354	4,146	1,609	1,599	210	118	50	42
Liquid, non-water	294	129	83	82	9	2	4	3
Metal	30,504	16,917	6,801	6,786	805	427	187	190
Mirror	3,242	1,871	684	687	315	176	67	72
Paint/plaster/enamel	39,323	21,765	8,773	8,785	10,965	6,073	2,434	2,458
Paper	20,763	11,692	4,592	4,479	883	485	200	199
Pearl	282	129	77	76	0	0	0	0
Photograph/painting	4,344	2,435	976	933	174	90	41	43



**Fig. 6. Fused material labels.** *Left to right:* van, sports, aerial photo, conference and dining area. The 5th image has a label density close to the mean density of DMS. The rightmost image is a fused label map from OpenSurfaces with a label density close to the mean density of OpenSurfaces. See Table 5 for color legend.

Table 9: continued from previous page

Plastic, clear	6,431	3,583	1,425	1,423	129	69	28	31
Plastic, non-clear	30,506	17,154	6,662	6,690	1,278	708	282	288
Rubber/latex	7,811	4,244	1,788	1,779	65	32	17	16
Sand	272	110	76	86	70	24	20	26
Skin/lips	18,524	10,444	4,014	4,066	509	287	113	108
Sky	3,306	1,447	911	948	1,020	435	286	298
Snow	191	70	60	61	57	19	20	18
Soap	154	58	50	46	0	0	0	0
Soil/mud	1,855	860	495	500	165	73	42	51
Sponge	326	149	89	88	1	1	0	0
Stone, natural	2,076	962	569	545	355	156	102	98
Stone, polished	1,831	993	435	403	187	97	46	44
Styrofoam	88	33	27	28	2	1	0	1
Tile	10,173	5,722	2,206	2,245	1,490	845	321	323
Wallpaper	1,076	577	252	247	233	127	56	49
Water	2,063	959	552	552	564	260	156	149
Wax	1,107	578	260	269	7	3	2	2
Whiteboard	1,171	642	265	264	111	60	24	27
Wicker	1,895	1,031	438	426	75	35	22	18
Wood	24,248	13,496	5,309	5,443	3,608	2,006	802	800
Wood, tree	2,026	929	561	536	72	30	19	22
Asphalt	474	211	132	131	73	35	17	22

We found that asking annotators to label all surfaces required extensive instruction. Our training document grew to include clarifications for rare and uncommon cases. In Table 11 we summarize how we choose to resolve cases.

In Table 12 we report the number of images in which an object class is detected by [12], and the number of images which are predicted by [45] to have scene elements for an activity. There are 80 object classes and 30 functional scene attributes. For brevity, we report only the largest classes.

For most images we collected two unique opinions for labels. In Table 13 we report the number of images with a given number of opinions.

**Table 10. Material occurrence in pixels.** We report the number of pixels covered by each label according to the resized images used by annotation tools.

Animal skin	22,995,883	Paint/plaster/enamel	7,796,144,397
Bone/teeth/horn	3,050,548	Paper	628,009,751
Brickwork	145,410,237	Pearl	411,455
Cardboard	93,881,191	Photograph/painting	123,296,052
Carpet/rug	707,147,207	Plastic, clear	93,002,805
Ceiling tile	216,289,692	Plastic, non-clear	906,618,216
Ceramic	185,191,692	Rubber/latex	45,644,757
Chalkboard/blackboard	48,346,203	Sand	47,860,125
Clutter	8,845,550	Skin/lips	359,727,474
Concrete	283,303,562	Sky	702,864,398
Cork/corkboard	6,468,131	Snow	40,936,881
Engineered stone	13,140,139	Soap	265,782
Fabric/cloth	3,408,488,743	Soil/mud	114,322,155
Fiberglass wool	1,874,005	Sponge	1,075,671
Fire	7,965,989	Stone, natural	253,271,347
Foliage	961,103,715	Stone, polished	134,425,626
Food	192,755,372	Styrofoam	1,552,343
Fur	145,359,760	Tile	1,068,909,615
Gemstone/quartz	7,273,649	Wallpaper	168,289,772
Glass	1,535,538,311	Water	390,040,955
Hair	238,600,730	Wax	4,791,692
Ice	18,308,742	Whiteboard	80,692,711
Leather	149,122,712	Wicker	50,066,493
Liquid, non-water	5,861,652	Wood	2,584,799,129
Metal	573,827,793	Wood, tree	50,922,547
Mirror	224,631,105	Asphalt	51,218,822

In Figure 6 we expand on Figure 3 by showing more fused label maps and we show a fused label map from DMS and OpenSurfaces which are representative of the mean density of the respective datasets.

## B Skin Type Experiment

In Section 4.2, we compared skin accuracies for three skin groups, Type I-II, Type III-IV, and Type V-VI. In order to compute accuracy we have to assign ground truth pixels to a group. We do this for images which contain detections of only one skin group. However, there are images where multiple skin groups co-occur and where no skin groups were detected. We do not evaluate on these two scenarios to avoid assigning groups incorrectly.

**Table 11. Case resolution.** For some cases we provided additional instruction, which we summarize here.

Case	Resolution
Skin with sparse hair	<i>Skin</i> for people; <i>animal skin</i> for animals.
Coat of hair (e.g., horse)	<i>Fur</i> .
Smoothed stone	<i>Polished stone</i> .
Laminated paper	<i>Clear plastic</i> .
Sauces	<i>Food</i> on food; <i>non-water liquid</i> during preparation.
Chandelier prisms	<i>Gemstone</i> or <i>glass</i> based on appearance.
Seasoned or blued metal	<i>Metal</i> .
Metal patina	<i>Metal</i> .
Printed text	The underlying material.
Mirror-like finishes	<i>Mirror</i> if sole purpose is to reflect; the material otherwise.
Wrapped items	The material of the wrap.
Electronic display	<i>Glass</i> .
Glass-top surface	<i>Glass</i> .
Thatch	<i>Wicker</i> .
Stained wood	<i>Wood</i> .
Projection screen	<i>Not on list</i> .
Vinyl	The closest of <i>non-clear plastic</i> , <i>rubber</i> or <i>leather</i> .

## C Benchmark Experiment Details

In this section we include more details on training our material segmentation benchmark model, DMS-46, from Section 4.3 of the main paper. All the models are trained on NVIDIA Tesla V100 GPUs with 32 GB of memory.

### C.1 Data Augmentation

In this section we show details on how we apply different data augmentation in training. We apply the following data transformation in order:

**Scale.** We first scale the input image so that the shortest dimension is 512 given that the training image size has height 512 and width 512. Then we randomly scale the input dimension with a ratio in [1, 2, 3, 4] uniformly.

**Horizontal Flip.** We apply random horizontal flip with probability 0.5.

**Vertical Flip.** We apply random vertical flip with probability 0.5.

**Color Jitter.** We apply color jitter with probability 0.9, using torchvision<sup>2</sup> ColorJitter with brightness 0.4, contrast 0.4, saturation 0.4, and hue 0.1.

**Gaussian Blur or Gaussian Noise.** We apply this transformation with probability 0.5. Gaussian blur or Gaussian noise is selected with equal chance. We use a kernel size of 3 for Gaussian blur with uniformly chosen standard deviation in [0.1, 2.0]. Gaussian noise has mean of 0 and standard deviation 3 across all the pixels.

<sup>2</sup> <https://pytorch.org/vision/>

**Table 12. Objects and functional spaces.** We report the number of images for the largest classes of detected objects (*top*) and estimated scene functions (*bottom*).

	All	Train	Val	Test		All	Train	Val	Test
person	19,966	11,219	4,303	4,426	tie	1,398	802	280	314
chair	17,617	9,987	3,826	3,780	bench	1,196	671	244	277
dining table	8,086	4,511	1,765	1,806	keyboard	1,192	648	272	272
bottle	5,964	3,320	1,313	1,325	cell phone	1,121	629	269	222
cup	5,656	3,136	1,248	1,265	mouse	939	516	199	224
potted plant	5,078	2,762	1,122	1,191	refrigerator	834	504	161	168
book	4,384	2,465	976	939	backpack	739	420	154	165
tv	4,303	2,411	947	942	oven	737	399	173	165
laptop	3,076	1,737	664	675	remote	718	403	166	148
bowl	2,900	1,579	636	682	dog	692	369	162	160
couch	2,846	1,614	628	602	cat	685	344	162	178
vase	2,790	1,551	626	609	toilet	677	383	144	149
bed	2,357	1,348	524	482	knife	579	335	123	120
sink	1,747	949	395	402	car	542	292	128	121
handbag	1,617	906	366	345	boat	524	227	136	161
wine glass	1,473	797	332	343	suitcase	510	310	94	106
clock	1,452	814	294	343	spoon	477	258	106	112
working	14,343	8,032	3,124	3,166	swimming	868	397	240	230
reading	14,039	7,931	3,118	2,970	sports	824	442	181	198
socializing	8,545	4,869	1,794	1,873	using tools	686	369	149	167
congregating	7,317	4,129	1,559	1,620	praying	649	363	144	138
eating	5,862	3,217	1,294	1,345	touring	626	283	159	180
shopping	2,419	1,325	563	526	waiting in line	593	362	118	113
studying	2,070	1,147	459	463	exercise	574	329	106	137
competing	1,960	1,085	410	458	diving	556	275	163	117
spectating	1,489	845	305	335	bathing	524	288	120	115
training	1,335	744	295	295	research	451	251	92	108
transporting	1,153	587	268	297	cleaning	445	247	94	104
boating	876	371	235	267	driving	404	199	92	113

**Rotation.** We apply random rotation in  $[-45, 45]$  degrees with probability 0.5. We fill 0 for the area outside the rotated color image and an ignore value for the rotated segmentation map. The loss calculation ignores those pixels.

**Crop.** Finally, we randomly crop a subregion, height 512 and width 512, to feed into the neural network.

## C.2 Loss Function

We use weighted symmetric cross entropy [36] as the loss function for DMS-46. The weight  $W_i$  for each class is calculated as a function of frequency of pixel

**Table 13. Judgments.** We report the number of unique opinions (*i.e.*, label maps) collected for images.

Label Map Count	Images
1	1,245
2	35,039
3	7,459
4	122
5	867

count,  $F_i$ , for each material class  $i \in N$  [48], in Equation 1.

$$W_i = \frac{1}{\log \left( 1.02 + \frac{F_i}{\sum_{i=1}^N F_i} \right)} \quad (1)$$

The number 1.02 is introduced in [48] to restrict the class weights in [1, 50] as the probability approaches 0. The weights we are using for DMS-46 are presented in Table 14.

Symmetric cross entropy (SCE) [36] is composed of a regular cross entropy (CE) and a reverse cross entropy (RCE) to avoid overfitting to noisy labels. Given the target distribution  $P$  and the predicted distribution  $Q$ , Equation 2 shows each part of the loss function for SCE. We choose  $\alpha = 1$  and  $\beta = 0.5$  for the weighting coefficients.

$$L_{SCE} = \alpha L_{CE} + \beta L_{RCE} = \alpha(-\sum P \log Q) + \beta(-\sum Q \log P) \quad (2)$$

### C.3 Model Architecture Implementation

We select ResNet50 [13] with dilated convolutions [7,42] as the encoder, and Pyramid 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]. We use a publicly-available implementation of ResNet50dilated architecture with pre-trained weights (on an ImageNet task) from [46,47]<sup>3</sup>, under a BSD 3-Clause License.

### C.4 Material Class Selection For Benchmark

In Section 4.3 we reported empirically finding that six material categories (*non-water liquid, fiberglass, sponge, pearl, soap* and *styrofoam*) fail consistently across models. We present the three top candidates of DMS-52 which led us to this conclusion. Each one is the best fitted model, according to DMS-val, from a comprehensive hyper-parameter search on learning rate, learning rate scheduler,

<sup>3</sup> <https://github.com/CSAILVision/semantic-segmentation-pytorch>

**Table 14. Class weights.** We show the class weights we applied in the loss function for DMS-46.

Label	Weight	Label	Weight	Label	Weight
Bone	50.259	Whiteboard	43.585	Hair	33.870
Wax	50.140	Clear plastic	42.709	Water	30.402
Clutter	50.136	Soil	42.585	Skin	29.049
Cork	49.995	Cardboard	42.482	Sky	24.133
Fire	49.945	Artwork	40.905	Metal	23.981
Gemstone	49.826	Fur	40.427	Paper	22.447
Engineered stone	49.459	Pol. stone	40.226	Carpet	20.422
Ice	49.163	Brickwork	38.979	Foliage	19.325
Animal skin	48.646	Leather	38.715	Non-clear plastic	17.986
Snow	47.972	Food	38.368	Tile	15.895
Sand	47.603	Wallpaper	37.854	Glass	12.555
Tree wood	46.759	Ceramic	37.201	Wood	8.388
Rubber	46.672	Nat. stone	35.919	Fabric	6.596
Wicker	46.465	Mirror	34.651	Paint	3.415
Chalkboard	46.462	Ceiling tile	34.617		
Asphalt	46.447	Concrete	34.095		

and optimizer. The first model, called DMS-52, is the best model across all models, is introduced in the main paper, and we report the per-class performance in Table 15. The second model, called DMS-52 variant A, has the same architecture as DMS-52 and uses all of OpenSurfaces data as additional training data. We report the per-class performance of DMS-52A in Table 16. The third model, called DMS-52 variant B, has a ResNet101 architecture and uses OpenSurfaces data as additional training data. We report the per-class performance of DMS-52B in Table 17. Across DMS-52, DMS-52A and DMS-52B the same six material classes are the worst-performing categories. Based on these findings we selected the other 46 categories for a benchmark and leave these six to future work.

## C.5 More Real-World Examples

We show more DMS-46 predictions on real world images in Figure 7.

## D Image Credits

Photos in the paper and supplemental are used with permission. We thank the following Flickr users for sharing their photos with a CC-BY-2.0<sup>4</sup> license. Some photos in the main paper were changed to remove logos or faces, scale, mask, or crop.

Image credits: Random Retail, Ross Harmes, Amazing Almonds, Jonathan Hetzel, Patrick Lentz, Colleen Benelli, Jannes Pockele, FaceMePLS, Michael

<sup>4</sup> <https://creativecommons.org/licenses/by/2.0/>



**Table 15. DMS-Val results for DMS-52.** Results are sorted by accuracy.

	Acc	IoU		Acc	IoU		Acc	IoU
Sky	0.937	0.891	Glass	0.703	0.489	Animal skin	0.396	0.268
Fur	0.913	0.694	Paper	0.686	0.496	Rubber	0.345	0.240
Foliage	0.897	0.769	Leather	0.676	0.397	Pol. stone	0.332	0.236
Ceiling tile	0.890	0.679	Nat. stone	0.634	0.447	Tree wood	0.327	0.224
Hair	0.885	0.673	Wax	0.626	0.430	Ice	0.320	0.284
Food	0.882	0.689	Wicker	0.622	0.432	Bone	0.213	0.178
Water	0.881	0.695	Wallpaper	0.603	0.397	Clutter	0.209	0.186
Skin	0.876	0.647	Concrete	0.579	0.333	Gemstone	0.127	0.077
Carpet	0.855	0.582	Soil	0.578	0.376	Cork	0.115	0.102
Fire	0.821	0.621	Cardboard	0.571	0.340	Eng. stone	0.096	0.069
Wood	0.801	0.657	Non-clear plastic	0.562	0.322	<b>Sponge</b>	0.051	0.050
Fabric	0.787	0.690	Asphalt	0.560	0.386	<b>Liquid</b>	0.048	0.044
Brickwork	0.785	0.514	Metal	0.548	0.305	<b>Fiberglass</b>	0.034	0.034
Whiteboard	0.771	0.508	Sand	0.548	0.407	<b>Styrofoam</b>	0.003	0.003
Tile	0.752	0.564	Snow	0.495	0.414	<b>Pearl</b>	0.000	0.000
Chalkboard	0.747	0.616	Clear plastic	0.441	0.254	<b>Soap</b>	0.000	0.000
Ceramic	0.746	0.482	Mirror	0.423	0.297			
Paint	0.707	0.640	Artwork	0.407	0.271			

Button, samuelrogers752, Ron Cogswell, David Costa, Janet McKnight, Jennifer, Adam Bartlett, [www.toprq.com/iphone](http://www.toprq.com/iphone), Seth Goodman, Municipalidad Antofagasta, Tom Hughes-Croucher, Travis Grathwell, Associated Fabrication, Tjeerd Wiersma, mike.benedetti, Frédéric BISSON, Wendy Cutler, with wind, Barry Badcock, Joel Kramer, Gwydion M. Williams, Andreas Kontokanis, Jim Winstead, Mike Mozart, Keith Cooper, Kurman Communications, Inc., Paragon Apartments, Pedro Ribeiro Simões, jojo nicdao, Gobierno Cholula, David Becker, Emmanuel DYAN, Ewen Roberts, Supermac1961, fugzu, Erik (HASH) Hersman, Eugene Kim, Bernt Rostad, andrechinn, Geología Valdivia, peapod labs, Alex Indigo, Turol Jones, un artista de cojones, Blake Patterson, cavenderamy, tape-tenpics, DLSimaging, Andy / Andrew Fogg, Scott, Justin Ruckman, espring4224, objectivised, Li-Ji, Bruno Kussler Marques, and BurnAway.

**Table 16. DMS-Val results for DMS-52A.** Results are sorted by accuracy.

	Acc	IoU		Acc	IoU		Acc	IoU
Sky	0.946	0.889	Leather	0.695	0.407	Clear plastic	0.405	0.255
Fur	0.921	0.692	Paint	0.680	0.625	Rubber	0.367	0.240
Foliage	0.912	0.768	Wicker	0.670	0.436	Tree wood	0.358	0.221
Ceiling tile	0.886	0.686	Concrete	0.646	0.347	Wax	0.327	0.246
Hair	0.883	0.677	Soil	0.635	0.385	Ice	0.230	0.228
Water	0.883	0.707	Fire	0.626	0.570	Eng. stone	0.207	0.108
Skin	0.877	0.636	Nat. stone	0.620	0.439	Clutter	0.204	0.185
Food	0.875	0.688	Wallpaper	0.600	0.417	Bone	0.167	0.139
Carpet	0.830	0.614	Asphalt	0.599	0.401	Cork	0.126	0.112
Wood	0.821	0.654	Cardboard	0.586	0.362	Gemstone	0.087	0.057
Fabric	0.801	0.700	Snow	0.584	0.484	<b>Sponge</b>	0.066	0.060
Whiteboard	0.801	0.515	Non-clear plastic	0.555	0.319	<b>Fiberglass</b>	0.029	0.029
Brickwork	0.789	0.496	Metal	0.548	0.289	<b>Liquid</b>	0.009	0.009
Ceramic	0.772	0.471	Animal skin	0.517	0.272	<b>Pearl</b>	0.000	0.000
Tile	0.745	0.576	Pol. stone	0.489	0.254	<b>Soap</b>	0.000	0.000
Chalkboard	0.744	0.593	Sand	0.463	0.389	<b>Styrofoam</b>	0.000	0.000
Paper	0.718	0.509	Artwork	0.445	0.294			
Glass	0.696	0.502	Mirror	0.434	0.308			

**Table 17. DMS-Val results for DMS-52B.** Results are sorted by accuracy.

	Acc	IoU		Acc	IoU		Acc	IoU
Sky	0.943	0.865	Glass	0.690	0.488	Tree wood	0.352	0.257
Foliage	0.905	0.776	Nat. stone	0.685	0.402	Rubber	0.310	0.265
Hair	0.891	0.687	Wicker	0.684	0.454	Animal skin	0.301	0.254
Water	0.889	0.655	Paper	0.681	0.510	Ice	0.239	0.232
Food	0.862	0.687	Wallpaper	0.651	0.384	Bone	0.206	0.177
Skin	0.861	0.675	Leather	0.603	0.431	Wax	0.202	0.166
Ceiling tile	0.858	0.673	Snow	0.593	0.507	Eng. stone	0.198	0.106
Carpet	0.847	0.566	Concrete	0.587	0.316	Cork	0.192	0.134
Fur	0.829	0.720	Metal	0.553	0.300	Clutter	0.131	0.113
Wood	0.820	0.642	Soil	0.542	0.337	Gemstone	0.095	0.082
Fabric	0.789	0.701	Non-clear plastic	0.540	0.344	<b>Liquid</b>	0.029	0.022
Whiteboard	0.752	0.539	Asphalt	0.536	0.369	<b>Fiberglass</b>	0.017	0.016
Fire	0.739	0.654	Cardboard	0.529	0.367	<b>Sponge</b>	0.003	0.003
Ceramic	0.737	0.499	Sand	0.498	0.407	<b>Pearl</b>	0.000	0.000
Brickwork	0.734	0.501	Pol. stone	0.459	0.238	<b>Soap</b>	0.000	0.000
Chalkboard	0.733	0.634	Artwork	0.438	0.276	<b>Styrofoam</b>	0.000	0.000
Paint	0.705	0.633	Clear plastic	0.392	0.251			
Tile	0.704	0.535	Mirror	0.358	0.265			



**Fig. 7. Real-world examples.** Our model, DMS-46, predicts 46 kinds of indoor and outdoor materials. See Table 5 for color legend.