Supplemental: 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

7 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 occurrence. 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

* These authors contributed equally to this work.

Table 9. continued from previous page

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
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.

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

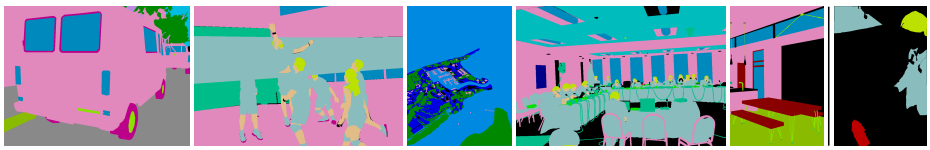

Fig. 6. Fused material labels. *Left to right:* van, sports, aerial photo, conference and dining area. The 5th image has a label density close to the mean density of DMS. The rightmost image is a fused label map from OpenSurfaces with a label density close to the mean density of OpenSurfaces. See Table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

9 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.

9.1 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.

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

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

Color Jitter. We apply color jitter with probability 0.9, using torchvision[1] 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.

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.

[1] https://pytorch.org/vision/

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

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

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

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

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.

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

9.3 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][2], under a BSD 3-Clause License.

[2] 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		

9.4 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, 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.

9.5 More Real-World Examples

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

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	**Sponge**	0.051	0.050
Fabric	0.787	0.690	Asphalt	0.560	0.386	**Liquid**	0.048	0.044
Brickwork	0.785	0.514	Metal	0.548	0.305	**Fiberglass**	0.034	0.034
Whiteboard	0.771	0.508	Sand	0.548	0.407	**Styrofoam**	0.003	0.003
Tile	0.752	0.564	Snow	0.495	0.414	**Pearl**	0.000	0.000
Chalkboard	0.747	0.616	Clear plastic	0.441	0.254	**Soap**	0.000	0.000
Ceramic	0.746	0.482	Mirror	0.423	0.297			
Paint	0.707	0.640	Artwork	0.407	0.271			

10 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[3] 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 Button, samuelrodgers752, Ron Cogswell, David Costa, Janet McKnight, Jennifer, Adam Bartlett, 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, tapetenpics, DLSimaging, Andy / Andrew Fogg, Scott, Justin Ruckman, espring4224, objectivised, Li-Ji, Bruno Kussler Marques, and BurnAway.

[3] https://creativecommons.org/licenses/by/2.0/

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	**Sponge**	0.066	0.060
Whiteboard	0.801	0.515	Non-clear plastic	0.555	0.319	**Fiberglass**	0.029	0.029
Brickwork	0.789	0.496	Metal	0.548	0.289	**Liquid**	0.009	0.009
Ceramic	0.772	0.471	Animal skin	0.517	0.272	**Pearl**	0.000	0.000
Tile	0.745	0.576	Pol. stone	0.489	0.254	**Soap**	0.000	0.000
Chalkboard	0.744	0.593	Sand	0.463	0.389	**Styrofoam**	0.000	0.000
Paper	0.718	0.509	Artwork	0.445	0.294			
Glass	0.696	0.502	Mirror	0.434	0.308			

References

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)

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	**Liquid**	0.029	0.022
Whiteboard	0.752	0.539	Asphalt	0.536	0.369	**Fiberglass**	0.017	0.016
Fire	0.739	0.654	Cardboard	0.529	0.367	**Sponge**	0.003	0.003
Ceramic	0.737	0.499	Sand	0.498	0.407	**Pearl**	0.000	0.000
Brickwork	0.734	0.501	Pol. stone	0.459	0.238	**Soap**	0.000	0.000
Chalkboard	0.733	0.634	Artwork	0.438	0.276	**Styrofoam**	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.