Learning Material-Aware Local Descriptors for 3D Shapes Supplementary Material

Hubert Lin[1] Melinos Averkiou[2] Evangelos Kalogerakis[3] Balazs Kovacs[4]
Siddhant Ranade[5] Vladimir G. Kim[6] Siddhartha Chaudhuri[6,7] Kavita Bala[1]
[1]Cornell Univ. [2]Univ. of Cyprus [3]UMass Amherst [4]Zoox [5]Univ. of Utah [6]Adobe [7]IIT Bombay

1. Introduction

This supplementary material is organized as follows. First we show the data collection interface and discuss additional statistics that may be of interest (section 2). Second, we discuss some additional training details (section 3). Third, we show statistics for our test set (section 4). Fourth, we discuss in detail the 2D classification baseline that we used in our evaluation (section 5). Fifth, we visualize embedding plots via t-SNE for our learned descriptor space (section 6). Sixth, we show confusion matrices for both Classification and Multitask networks (section 7), as well as for 3-view variants (section 8), and for network trained with only contrastive loss (section 9). Seventh, we show a sample of our dataset as well as a visual sample of our material prediction results (section 10).

2. Data collection

Our data collection interface is shown in Fig. 1. Four different rendered views covering the front, sides and back of the textured 3D shape were shown. At the foot of the page, a single shape component was highlighted while the rest of the 3D shape appeared faded. Each query highlighted a different component. Workers were asked to select a label from a set of materials for the highlighted component. In total, we collected 15923 labeled components in 3080 shapes. On average 76% of the surface area per mesh was labeled. For training, we kept only shapes with $> 50\%$ of components labeled (2134 shapes).

3. Training Points

To train the network, we sample 150 evenly-distributed surface points from each of our 3D training shapes. Points lacking a material label, or externally invisible, are discarded. Point visibility is determined via ray-mesh intersection tests. The remaining points are subsampled to 75 per shape. This subsampling is again performed so that selected points are approximately uniformly distributed along the shape surface. The choice to sample 75 points per shape is due to memory limitations (we store the dataset in the main memory to avoid slow I/O during training). The views corresponding to these points are preprocessed and saved as single channel, unsigned integer arrays which are read directly into memory at training time to prevent I/O bottlenecks. Note that sampling roughly 75 points per shape requires $\sim$ 60G memory. Preprocessing to store into memory rather than reading from disk offered us a $5-10$x speedup in training time.

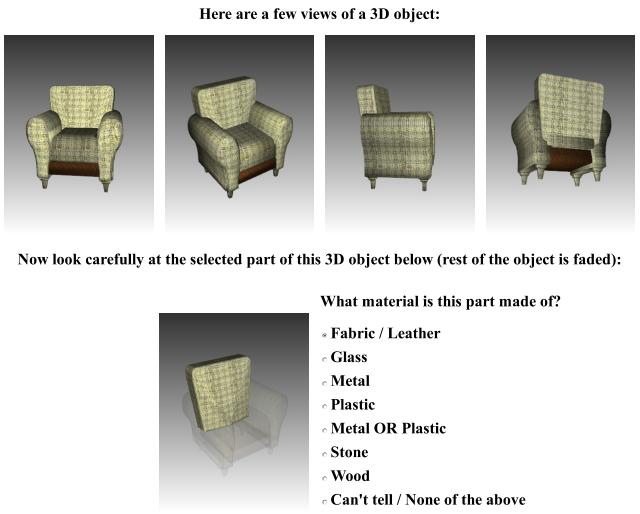


Figure 1: Our interface for collecting material labels for 3D shapes.

4. Benchmark Test Set Distributions

In Fig. 2, we show the distribution of labels across components in the benchmark shape dataset, as well as the distribution of labels across the points sampled from these shapes that form our evaluation test set. Notice that although there are a large number of metal and plastic components, relatively few metal or plastic points are sampled. This is because many metal or components are small thin structures (e.g. handles, table legs). Recall that we sample our test points uniformly across the surface of our shapes and thus the surface area of a component is proportional to the number of points sampled from that component.

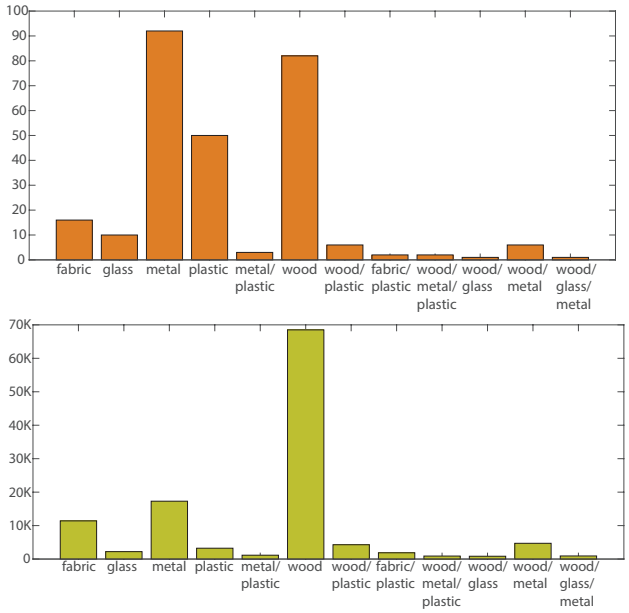


Figure 2: (top) Distribution of material labels assigned to components in our expert-labeled benchmark test shapes (bottom) Distribution of material labels assigned to the points sampled from the benchmark test shapes that form our test set.

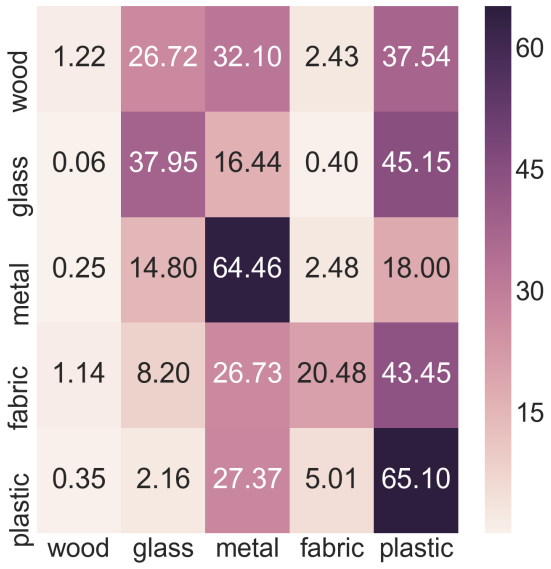


Figure 3: Confusion matrix for top-1 classification predictions for MINC-bw tested on 3D shapes.

5. 2D Classification Network Baseline

To evaluate the baseline of using a 2D material classification network, we use MINC. The network is based on GoogLeNet. We take their pretrained network and finetune on their dataset. The classification layer is finetuned to only classify the five materials we consider in this paper. Furthermore, we choose to finetune with greyscale images. The reason for this is that our our texture-less 3D renderings do not offer any color cues; therefore, we train the 2D network under similar conditions. This network is trained until validation losses converge with batchsize 24 with stochastic gradient descent with momentum. The initial learning rate is set to 0.001 and momentum is set to 0.4. The learning rate policy is polynomial decay with power 0.5. L_2 weight decay is set to 0.0002. We call this finetuned network MINC-bw. The confusion matrix for the network on our test 3D renderings is in Fig. 3. The poor performance suggests that it is non-trivial to adapt 2D photos train a network to learn material descriptors for 3D shapes.

6. Embedding Visualization

We visualize the learned material-aware descriptor embedding with t-SNE in Fig. 4. In both the Classification and Multitask variations, we see a tendency of our network to cluster datapoints.

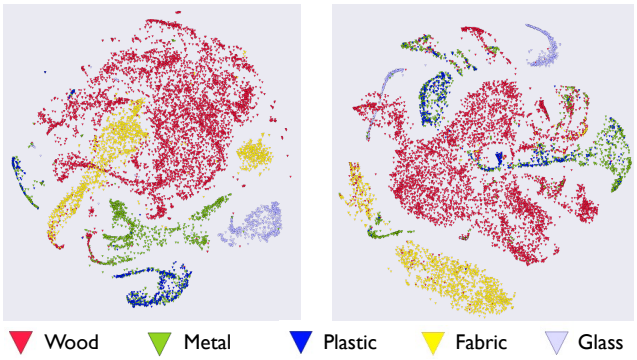
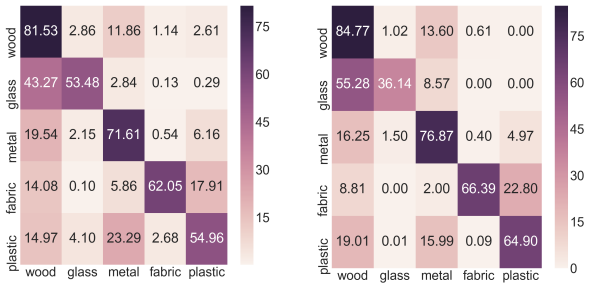


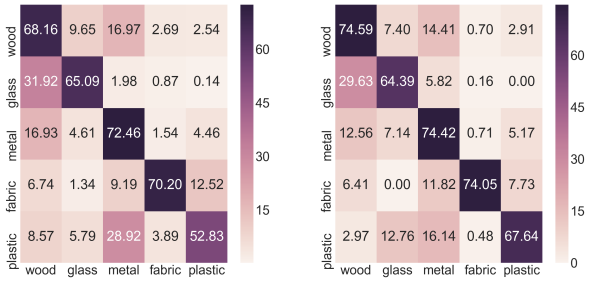
Figure 4: t-SNE embeddings for training points. *Left:* Classification loss, *Right:* Multitask loss. Points with multiple ground truth labels are shown with one label randomly selected.

7. Classification vs Multitask Confusion Matrices

We show the confusion matrices for Classification (as well as Multitask, for reference) in Fig. 5. Note that Classification predictions are more biased towards wood. As a result, its glass performance drops after CRF since many glass points tend to lie on surfaces that resemble wood surfaces (e.g. flat table tops, flat cabinet doors) – if many local predictions are wood rather than glass, it is likely that the CRF will smooth the predictions to wood.

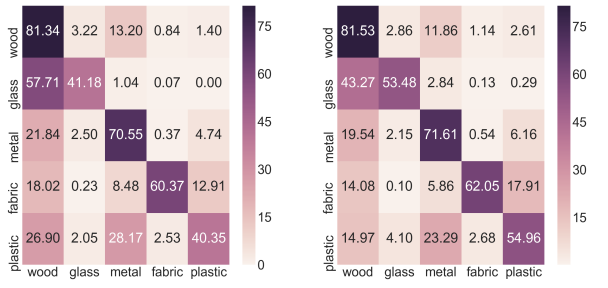


(a) Classification (b) Classification + CRF

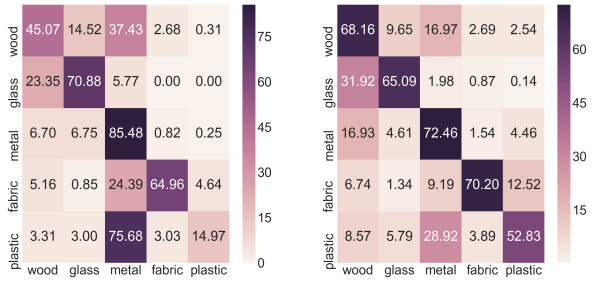


(c) Multitask (d) Multitask + CRF

Figure 5: Confusion matrices for Top-1 classification predictions.



(a) Classification 3 views (b) Classification 9 views



(c) Multitask 3 views (d) Multitask 9 views

Figure 6: Confusion matrices for Top-1 classification predictions.

8. Confusion Matrices for 3 view MVCNN

The confusion matrices for 3 view MVCNNs (1 viewpoint, 3 distances) are in Fig. 6. For reference, the matrices for the 9 view MVCNNs (3 viewpoints, 3 distances) are also shown. Note that confusions are reduced when using 9 views over 3 views. For both Classification and Multitask, fabric performance is relatively unaffected by reduced views while plastic suffers. In Classification 3 views, wood predictions dominate. In Multitask 3 views, metal predictions dominate – as a consequence, glass does relatively well (since glass is typically competing with wood) and plastic does extremely poorly (since plastic parts can often be shaped like metal and our training dataset contains a high number of "plastic or metal" labels relative to "plastic" labels).

9. Confusion Matrix for Contrastive Loss Only

The MVCNN trained with only contrastive loss achieves a mean class top 1 accuracy of 59% (in comparison to 65% with Classification and 66% with Multitask). This variant often confuses plastic for metal, and performs poorly on glass relative to the Classification or Multitask variants. The confusion matrix is shown in Fig. 7.

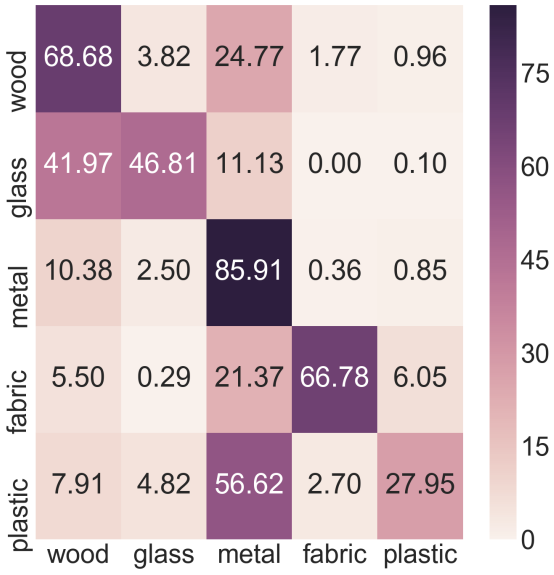


Figure 7: Confusion matrix for Top-1 classification predictions for network trained with contrastive loss only.

10. Sample of Dataset and Predictions

Here we show some samples from both our high-quality expert-annotated benchmark dataset as well as our large crowdsourced training dataset. Please refer to the legend by each shape for labels. The colors are consistent within each figure but may not be across figures.

Figure 8 shows a small sample of our benchmark test shapes with ground truth labels. Figure 9 shows per-point

predictions for each of the 1024 test point samples on these benchmark test shapes. Figure 10 shows per-part predictions after the 1024 point predictions are smoothed with our symmetry-aware CRF. Figure 11 shows a small sample of our MTurk crowdsourced data with the 75 training point samples per shape shown.

Figure 8: Small selection of benchmark shapes. Ground truth labels are shown. Please refer to legend for each shape for labels.

Figure 9: MVCNN point-predictions on benchmark shapes. Please refer to legend for each shape for labels.

Figure 10: MVCNN+CRF part-predictions on benchmark shapes. Please refer to legend for each shape for labels. (Note that colors are not consistent with Figs 8, 9)

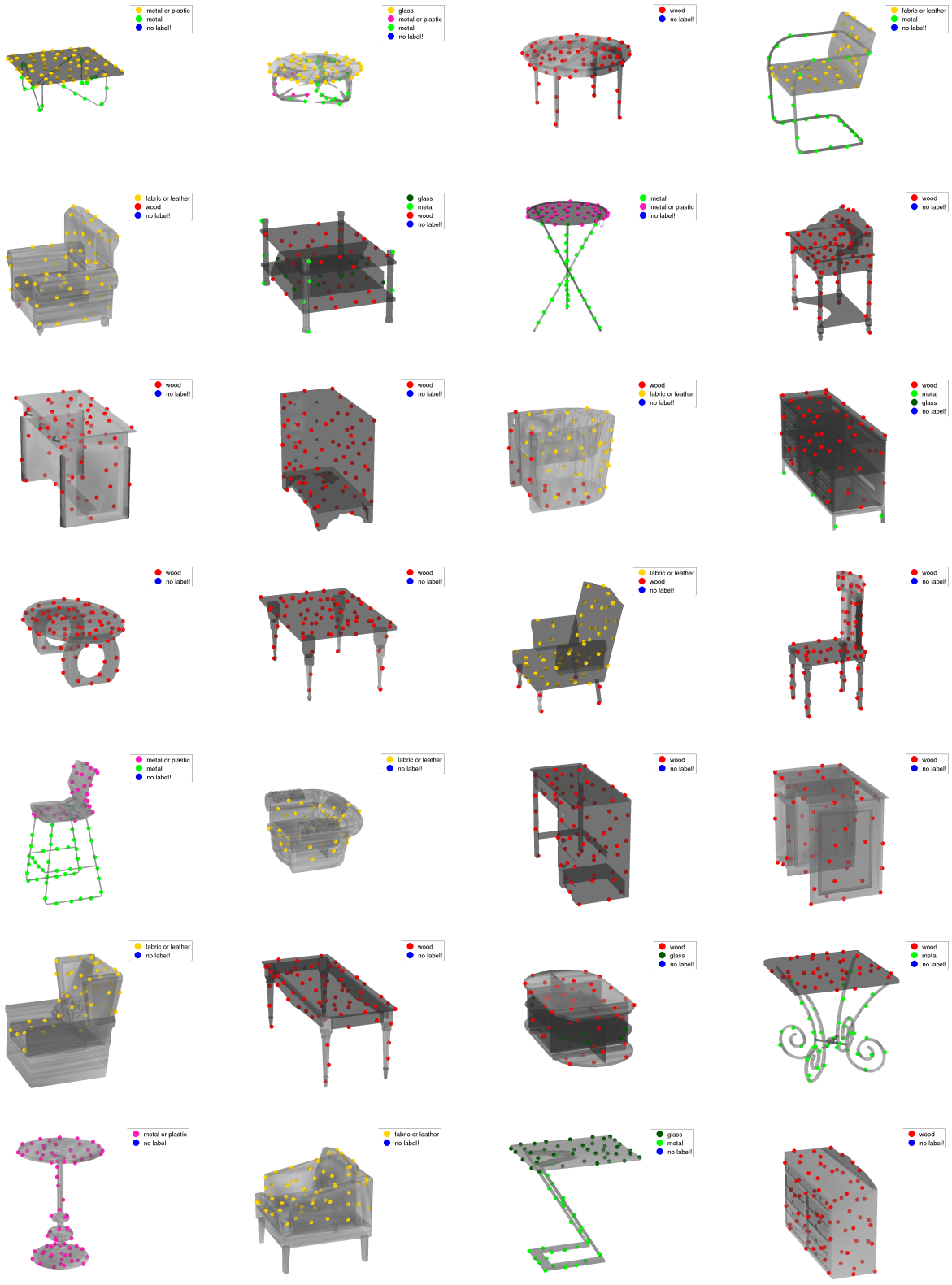


Figure 11: Small sample of crowdsourced dataset. 75 training point samples with ground truth labels per shape are shown. Please refer to legend for each shape for labels.