Visual Material Recognition

A Thesis
Submitted to the Faculty
of
Drexel University
by
Gabriel Schwartz
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
December 2017
First and foremost I want to thank my advisor, Dr. Ko Nishino. Without his support and patience I would not be where I am today. He was a constant source of guidance, insight, and motivation throughout my graduate student career. He knew what I was capable of even when sometimes I did not.

My parents have always supported me and I’ve never doubted that they would continue to do so regardless of where life took me. For that I can’t thank them enough. Their hard work and dedication has afforded me with countless opportunities; my education is just one of many.

I’m extremely grateful for my labmates, especially Geoff and Steve. I could always turn them for advice, technical discussions, or just to talk about video games. Simply knowing that they faced the same obstacles and succeeded was a source of hope. I’d also like to thank all of my friends and climbing partners over the years. There’s no better way to relax than a trip into the desert or a long day on the rocks.

This thesis was brought to you in part by the National Science Foundation and the Office of Naval Research. Thanks to their support I’m able to emerge from the Ph.D. program debt-free, which these days is something to be thankful for.
### Contents

<table>
<thead>
<tr>
<th>List of Tables</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>Abstract</td>
<td>xiii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Contributions</td>
<td>5</td>
</tr>
<tr>
<td>2. Related Work</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Material Recognition</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Attributes</td>
<td>8</td>
</tr>
<tr>
<td>2.2.1 Fully-Supervised Attributes</td>
<td>9</td>
</tr>
<tr>
<td>2.2.2 Weakly-Supervised Attributes</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Material Perception and Convolutional Neural Networks</td>
<td>11</td>
</tr>
<tr>
<td>2.4 Dense Prediction</td>
<td>12</td>
</tr>
<tr>
<td>2.5 Context in Visual Recognition</td>
<td>13</td>
</tr>
<tr>
<td>3. Visual Material Traits</td>
<td>14</td>
</tr>
<tr>
<td>3.1 Representing Material Traits</td>
<td>17</td>
</tr>
<tr>
<td>3.1.1 Convolutional Material Trait Features</td>
<td>18</td>
</tr>
<tr>
<td>3.1.2 Supplemental Features</td>
<td>22</td>
</tr>
<tr>
<td>3.1.3 Groupwise Feature Selection</td>
<td>22</td>
</tr>
<tr>
<td>3.2 Recognizing Material Traits</td>
<td>23</td>
</tr>
<tr>
<td>3.3 Using Visual Material Traits</td>
<td>28</td>
</tr>
<tr>
<td>3.3.1 Material Categories from Visual Material Traits</td>
<td>28</td>
</tr>
</tbody>
</table>
CONTENTS

3.3.2 Segmenting Images with Visual Material Traits ........................................... 31

4. **Automatically Discovering Visual Material Attributes** .......................... 35

4.1 Perceptual Distance between Materials ......................................................... 37

4.2 Defining the Material Attribute Space .......................................................... 41

4.3 Training a Material Attribute Classifier ....................................................... 43

4.4 Analysis of Discovered Attributes ............................................................... 46

4.5 From Discovered Attributes to Materials ...................................................... 50

5. **Perceptual Material Attributes in Convolutional Neural Networks** .......... 54

5.1 Perceptual Material Attributes from Local Material Recognition ................. 55

5.1.1 Finding Material Attributes in a Material Recognition CNN .................. 56

5.1.2 Material Attribute-Category CNN ............................................................ 56

5.2 Local Material Database ................................................................................. 59

5.2.1 Material Category Hierarchy ........................................................................ 59

5.2.2 Data Collection and Annotation .................................................................. 61

5.3 Perceptual Material Attributes Discovered in the MAC-CNN ....................... 64

5.3.1 Properties of the Perceptual Material Attributes ...................................... 64

5.3.2 Local Material Recognition ......................................................................... 68

5.4 Novel Material Category Recognition ............................................................. 70

6. **Integrating Local Materials with Global Context** .................................... 74

6.1 Role of Context in Material Recognition ....................................................... 77

6.1.1 Object Context ......................................................................................... 78

6.1.2 Place Context ............................................................................................ 78

6.2 Integrating Context in a Material Segmentation CNN ................................. 80

6.3 Experiments and Comparisons ...................................................................... 84
6.3.1 Material Segmentation Comparisons ........................................... 84
6.3.2 Ablation Studies ................................................................. 86
6.3.3 Qualitative Examples ............................................................ 91
7. CONCLUSION ................................................................................. 94
  7.1 Visual Material Traits .............................................................. 94
  7.2 Perceptual Material Attributes .................................................. 95
  7.3 MAC-CNN and Local Materials Database ..................................... 95
  7.4 Integrating Materials and Context ............................................. 96
  7.5 Future Work ............................................................................. 96
BIBLIOGRAPHY ................................................................................. 98
VITA .................................................................................................. 103
List of Tables

3.1 Selected features for material traits. As “fuzziness” is characterized by fine edge patterns, oriented filters and LBP are useful. Since we define “shiny” only on areas that exhibit specular highlights, it follows that color histograms and learned convolutional filters are important features for this material trait. . . . . . . . . . 23

3.2 Performance breakdown. FS: feature selection, SF: supplemental features, CAE: convolutional auto-encoder features. For the first row we performed direct material category recognition using the concatenation of all feature sets. This shows that the trait representation is indeed providing crucial information. . . . . . . . . . 31

6.1 Material segmentation scores for same-dataset experiments (each method trained and tested only on the given dataset). For MINC models, the table shows the score of the best scoring single model without ensembles. Note that the OpenSurfaces [7] dataset is a subset of MINC, and the FMD is a subset of our local materials database. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84

6.2 Material segmentation scores for cross-dataset experiments (each model trained on one dataset and tested on another). Models are tested without retraining, on overlapping categories only, in order to highlight generalization performance. . . 86

6.3 Accuracy for varying sources of context. Object and place categories each contribute significantly to the overall accuracy, and the combination of the two is even more accurate. As an example of this, sinks (an object) are often metal or ceramic, and bathrooms (a place) often contain metals and ceramics. Bathroom sinks, however, are typically ceramic. Objects and places together can provide information that is not available given either alone. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88

6.4 Accuracy for varying levels of context granularity. Fine-grained places may not appear in many images, but coarse grained categories may offer little in the way of material recognition cues. We find that the finest category granularity offers the best material segmentation performance. In this case, the 205 place categories are both fine-grained and sufficiently well-distributed across training examples. . 88

6.5 Accuracy with context introduced at varying levels. We introduce context at each of the above layers and compute material segmentation accuracy. The accuracy increases as the context is introduced at higher layers in the network, showing that the best level for context introduction is in the upper layers of the network. . 91
List of Figures

3.1 Materials like the plastic in these images exhibit a wide range of appearances depending on the object and scene, making extraction of material information without the use of object information challenging. We propose to locally recognize visual material traits, distinct appearances of material properties such as "translucent," to provide contextual cues for challenging vision tasks including material category recognition and segmentation. ............................................. 15

3.2 When adapted to use aggregated features from local image patches, methods that perform well on full images quickly lose accuracy. This suggests that they are relying heavily on context, including object shape cues, to recognize materials. 16

3.3 Successfully recognized material traits. These image patches were recognized by our framework as exhibiting the indicated material traits. Even at the patch level, we can see the characteristic visual appearances of each material trait. . . . . 17

3.4 These $7 \times 7$ px. convolution filters learned by the CAE represent the top three filters for the listed material traits, ranked by average presence in the testing images. The filters represent characteristic local texture and color patterns. The six filters on the right do not rank in the top three for any material trait. They exhibit significantly less texture variation than the top filters. . . . . . . 21

3.5 Example material trait recognition. Non-masked pixels in (b) and (c) correspond to pixels with high probability ($p > 0.5$) of exhibiting the given trait. Note that the recognized material traits appear consistently across regions of related materials. ................................................................. 24

3.6 Visual material trait recognition accuracy. Material traits are recognized via binary classification on a balanced training and testing set, thus random chance accuracy is 50%. Most traits are recognized well. Difficult material traits, such as metallic and transparent, are challenging due to their object- and environment-dependent appearances. Average accuracy is 78.4%. .................................................. 25

3.7 Material trait frequency distributions. We compute the class-conditional distributions for appearance frequency of each material trait given each material category. These are stored as histograms, examples of which are shown above. Plastic is most often smooth, while stone is very rarely smooth. . . . . . . . . . . . . 26

3.8 Our framework produced false-positive detections of material traits in these patches. For the challenging metallic trait, it is clear that color plays a strong role. The misclassifications generally have a metallic color even though the material is not metal. In some rare cases such as "smooth" there are missing annotations and thus the false positives are actually true positives. . . . . . . . . . . . . 27
3.9 Confusion matrices showing true class vs. predicted class on the Flickr Material
Database and ImageNet images. Average accuracy is 49.2% in (a) and 60.5%
in (b). Though metal and glass both have an appearance that is environment-
dependent, glass is more accurately classified. This is likely due to the tendency
of glass to create characteristic local distortions. ................................. 29

3.10 Three misclassified ImageNet images, with true classes for each prediction is in
parentheses. The left two are a result of confusing appearances (striped and
translucent are more often associated with wood and plastic respectively) while
the rightmost is due to the bounding box poorly fitting the object. ............... 31

3.11 Comparing segmentation with and without material traits. Images on the left
were segmented using the original NCuts algorithm, while those on the right were
segmented with our modified version. Material traits can indicate the difference
between fuzzy grass in the foreground and rocks in the background, despite the
fact that they have similar colors. ............................... 33

4.1 Sample material image patches. Each column contains patches containing the
same material. We would like to obtain a set of attributes that describe what
makes each material look distinct. Asking annotators to simply describe the
patches, however, is an ambiguous question. Patches may look similar even
though the annotator cannot find a concrete word to identify the similarity. In
this chapter, we show that we can probe the human perception of materials by
asking only for binary visual similarity decisions: “Do these two patches look
similar?” ................................................................. 36

4.2 Example projections of materials into a 2D similarity subspace. The locations
of the two material categories corresponding to the axes are marked. We would
expect that, in this case, water would lie furthest along the “water” axis and
likewise with leather. Materials with common visual properties, such as the
smoothness of plastic and glass, lie close to each other. Materials with distinct
visual properties, such as woven fabric and shiny metal, do not. .................. 40

4.3 t-SNE [53] embedding of materials from the raw feature space (a) and from our
discovered attributes (b). We embed a set of material image patches into 2D
space via t-SNE using raw features and predicted attribute probabilities as the
input space for the embeddings. Though t-SNE has been shown to perform well
in high-dimensional input spaces, it fails to separate material categories from the
raw feature space. Material categories are, however, clearly more separable with
our attribute space. ............................................................. 45

4.4 Per-pixel discovered attribute probabilities for four attributes (one per column).
These images show that the discovered attributes exhibit patterns similar to those
of known material traits. The first attribute, for example, appears consistently
within the woven hat and the koala; the second attribute tends to indicate smooth
regions. The last two columns show we are discovering attributes that can appear
both sparsely and densely in an image, depending on the context. These are all
properties shared with visual material traits. ........................................ 47
4.5 Typical per-pixel attribute probabilities based on a random attribute matrix. Unlike the predictions for attributes derived from human perception, these attributes appear randomly within a region and do not reflect any local visual properties. 48

4.6 Correlation between discovered attribute predictions and material traits. Groups of attributes can collectively indicate the presence of a material trait. Metallic, for example, correlates positively with attribute 0 and negatively with attribute 8. 49

4.7 Confusion matrix for material recognition on FMD images. Well-recognized categories, such as foliage, correspond with categories that appeared distinct in human annotations for perceptual distance. Annotators regularly selected foliage patches as appearing different from all other categories. 51

4.8 Accuracy vs. training set size. Accuracy does not continue to increase as we use larger training datasets. This shows that we have successfully extracted as much local information as possible from human perception. 52

5.1 Material Attribute-Category CNN (MAC-CNN) Architecture: We introduce auxiliary fully-connected attribute layers to each spatial pooling layer, and combine the per-layer predictions into a final attribute output via an additional set of weights. The loss functions attached to the attribute layers encourage the extraction of attributes that match the human material representation encoded in perceptual distances. The first set of attribute layers acts as a set of weak learners to extract attributes wherever they are present. The final layer combines them to form a single prediction. 57

5.2 Our proposed material category hierarchy. Categories at the top level (red) separate materials with notable differences in physical properties. Mid-level categories (green) are visually distinct. The lowest level of categories (blue) are fine-grained and may require both physical and visual properties and expert knowledge to distinguish them. In our local materials database, we collect annotations for mid-level categories only, as they correspond to names likely to be familiar with a non-technical audience. We make one exception for concrete and asphalt, as those names are more familiar than the term “composite”. We also add supplemental categories for food, water, and non-water liquids. 60

5.3 Local material patches extracted as the final step in our database creation process. These patches are used to compute human perceptual distances, and also form the training input for our combined material attribute-category CNN. 62

5.4 Example annotation results. Annotators did not hesitate to take advantage of the ability to draw multiple regions, and most understood the guidelines concerning regions crossing object boundaries. As a result, we have a rich database of segmented local material regions. 63
5.5 Attribute Space Embedding via t-SNE [53]: Many categories, such as water, food, foliage, soil, and wood, are extremely well-separated in the attribute space. We find that this separation corresponds roughly with per-category accuracy as shown in Figure 5.8. While other categories do overlap to some extent, they still form separate regions in the space.

5.6 Each column after the first (the input image) shows per-pixel probabilities for an extracted perceptual attribute. The attributes form clearly delineated regions, similar to semantic attributes, and their distributions match as well.

5.7 By performing logic regression from our MAC-CNN extracted attributes to semantic material traits, we are able to extract semantic information from our non-semantic attributes. We can apply logic regression to material attribute predictions on patches in a sliding window to obtain per-pixel semantic material trait information. The per-pixel trait predictions show crisp regions that correspond well with their associated semantic traits. Traits are independent, and thus the maps contain mixed colors. Fuzzy and organic in the lower right image, for example, creates a yellow tint.

5.8 Local material recognition accuracy, by category. Average accuracy is 60.2%. It is clear that some categories, such as metal and glass, are significantly more challenging to recognize locally.

5.9 Images in each column share true material categories. The first three rows are correct predictions, and incorrect predictions (bottom two rows) are shown under the corresponding images. Glass and metal, for example, are both materials whose appearance depends heavily on the surrounding environment. Asphalt and concrete are both common paving materials and it is sensible that they are often confused.

5.10 Applying the MAC-CNN in a sliding-window fashion leads to a set of material category probability maps. These material maps show that we may obtain coherent regions using only small local patches as input. The foliage predictions in the bottom right image are reasonable, as the local appearance is indeed a flower. In the upper right image, the local appearance of the fence resembles lace (a fabric).

5.11 Graphs of novel category recognition accuracy vs. training set size for various held-out categories. The rapid plateau shows that we need only a small number of examples to define a previously-unseen category. The accuracy difference between feature sets shows that the attributes are contributing novel information.
6.1 Material segmentation methods based on large-patch CNNs implicitly rely on the context present in the patch to classify materials. When the context is ambiguous, this leads to errors that can be resolved using the local appearance information. Here, for example, the object is a house in an outdoor scene, but the area surrounding the windows is a painted surface, not glass. Since existing methods do not cleanly separate local appearance and context, they cannot resolve such ambiguities.

6.2 The image above, output from our MAC-CNN, shows material category probabilities for three materials: wood, foliage, and fabric, in the RGB channels respectively. Their method uses only local information; as a result the foliage pattern on the sofa is misclassified as actual foliage. This is an example where scene context is vital in resolving an otherwise ambiguous local material appearance.

6.3 The conditional distributions of materials given ground-truth object categories (top row) and predicted places (bottom row) are highly discriminative. Many context categories exhibit only a small set of materials. Some outliers are inevitable as the ground-truth COCO segmentation masks do not perfectly conform to actual object boundaries in the image.

6.4 Distribution of the ratio of probabilities for predicted vs. true categories given that the prediction was incorrect. We can see that incorrect categories, the ones we would like to change via the use of context, can have much higher probabilities than that of the true category. As a result, simple multiplication with context-conditional distributions will rarely change the classifier's output for the better.

6.5 Material segmentation CNN architecture. Our network takes an input image, object category probability map, and a place category probability vector as inputs. Horizontal lines represent additive skip connections, with appropriate zero-padding on the channel axis. During training, the network only sees $48 \times 48$ px image patches to ensure we are separating local material appearance from context. At test time, we may input an image of arbitrary size.

6.6 Accuracy vs. training set size on the MINC database ($1.0 \approx 2.5$ million patches). We can clearly see that by separating local material appearance from context, we are able to recognize materials more accurately from fewer examples.

6.7 These examples show that context helps disambiguate materials when local information is not sufficient. In the first set of insets, the water has a local appearance similar to asphalt. Global context suggests that this is unlikely. In the second set, we see that the airplane body is incorrectly recognized due to the lack of characteristic specular reflection that locally identifies metal. Again, context fixes this error. Sky is not a material and in this case has the local appearance of water, hence the prediction for those pixels in the second row.
6.8 Sample failure cases. Insets compare the output of our model vs the model of [6] trained on MINC images. Cat fur does not match a material in the MINC database and is unlikely to appear in their database due to the image sources used. Our method predicts fabric instead, which has a similar local visual appearance. The stone statue image contains few contextual cues, but we are able to make reasonable predictions based on the local appearance.

6.9 Additional examples of dense material recognition with context. It is important to note that neither skin nor sky are considered materials within our hierarchy. Skin is a unique case of material that is visible only on one object category (people) in most databases, and the sky is not a material.

6.10 Additional examples of the output of our dense per-pixel material recognition framework.
Abstract
Visual Material Recognition
Gabriel Schwartz
Ko Nishino, Ph.D.

Materials inform many of our interactions with everyday objects. Knowing that a cup is ceramic, we handle it more gently. When sidewalks are covered with snow and ice, we walk differently so as not to slip. If we aim to create an autonomous system, such as a robot, that can manipulate a wide variety of objects or traverse the many different surfaces it may encounter, we will need to be able to provide this material information algorithmically. Visual material recognition is the process of identifying the presence of materials, such as plastic, glass, or metal, in ordinary images. By recognizing these materials, we can obtain valuable cues for general image understanding. Doing so, however, is a challenging problem, as a single material may exhibit many different visual appearances. We can recognize an object based on its characteristic shape, but materials do not have such a singular distinguishing property. In this thesis, we study the problem of visual material recognition by breaking the recognition process down into fundamental and separable components. Our key observation is that the appearance variation which makes materials so challenging to recognize arises from the context in which the materials appear. A smooth white surface does not on its own provide many cues as to the material in question, but when combined with the fact that the surface is on a mug, we may infer that the material is likely ceramic or plastic. In order to take advantage of this observation, we must be able to separate material appearance from the context in which it appears. As a first step, we demonstrate that it is possible to recognize materials from small image patches. These small patches contain only
the appearance of the material, and not that of the surrounding context. We achieve this by using the simple visual material properties which humans use to describe materials, such as “shiny” or “translucent”, as an intermediate representation for the materials themselves. We refer to these properties as visual material traits. Though they prove useful, obtaining annotations for these traits is a challenging and time-consuming process. To address this, we derive an automatic perceptual attribute discovery method that generates classifiers for a set of unknown attributes. By probing the human perception of materials through easily-obtained binary annotations, we may measure the visual similarity of materials and discover attributes that serve the same function as material traits. Finally, having shown that material appearance may be isolated in small local image patches, we introduce a convolutional neural network (CNN)-based framework that integrates local material appearance with global contextual cues. By cleanly separating and combining the material appearance and context, we can take advantage of the strong material cues we show are present in that context to accurately recognize materials with far fewer examples than past attempts at material recognition.
Chapter 1: Introduction

Material recognition – identifying the presence of materials, such as glass or metal, in images – can provide valuable cues for autonomous interaction. Knowing the composition of an object can strongly influence how a robot or other autonomous system may handle it: a plastic knife, for example, can tolerate much less force than a metal one. Materials are also a key component in image understanding and visual-question-answering [3], enabling a robot to, “Pick up the glass cup on the table,” or answer the question, “How many wooden toys are there?” Object recognition can identify the cups, tables, and toys, but to be more specific we need material recognition. We must be able to algorithmically recognize the presence of materials in ordinary images if we are to provide such information to any system.

Recognizing materials has proven to be a challenging problem. Early work, such as that of Liu et al. [37], focused on simple images (one primary material and object of interest, uncluttered scenes, closeup views) and material categories. Even so, the accuracy of their resulting material predictions was relatively low (44.6%). The challenge in recognizing materials visually is largely due to the wide variety of appearances which each material may exhibit. Unlike objects, where, for example, cars tend to exhibit a characteristic shape, materials have no such simple distinguishing properties. One material, such as plastic, may appear in a number of different colors, textures, and reflectances.

The unifying observation in this thesis is that the challenging variation in material appearance arises due to the different contexts in which materials appear. A single material may appear as part of many different objects, and each of those objects may in turn appear in a wide variety of different scenes. As we will show, the type of scene and types of objects
in that scene both strongly constrain and influence the presence and appearance of materials in the image. Metal and ceramic, for example, are two challenging materials to recognize: metal due to the fact that its appearance often depends on its environment, and ceramic due to its lack of distinguishing features. If, however, we know an object is a sink, then we may infer that it is likely made of metal or ceramic. Likewise we may also observe that kitchen sinks are typically metal while bathroom sinks are often ceramic. We refer to such object and scene categories as “context” when they are used to inform material recognition.

We show that we can use our observations concerning materials and context to greatly reduce the number of examples required to accurately recognize a material. In order to achieve this, however, we must first be able to separate material appearance from the surrounding context. Existing material recognition methods do not do so, and instead build frameworks that rely on an entangled combination of material appearance and context with no clear delineation. These methods require very large training datasets to achieve reasonable accuracy. Since they cannot separate the effects of context on material appearance, such methods depend too heavily on the contextual cues and must see all combinations of material and context.

As a first step towards a full integration of untangled materials and context, we demonstrate that we can recognize materials independent of context using small image patches. We refer to this process as local material recognition. Recognizing materials using only the local information contained in a small image patch appears to be a daunting task. Looking at materials closely, it becomes clear how much even our own recognition process can rely on context. Despite this, humans are able to identify visual properties of materials even when we can’t see the surrounding context. We can look at a smooth plastic surface, for example, and see that it is translucent and possibly shiny regardless of the object involved. We refer to
these terms (e.g. “shiny”, “smooth”) describing characteristic material appearance properties as “visual material traits”. We show that these traits can be accurately recognized from only local information and thus that we can recognize materials independent of external context.

Though visual material traits form a useful object-independent intermediate representation for materials, a number of challenges arise when attempting to apply material traits to larger datasets. First, material traits rely on a single, manually-defined set of trait names for annotation and recognition. This is acceptable when dealing with small datasets which may be annotated by a single annotator following their own internal definitions of the traits. If, however, we aim to increase the size of the dataset in question, then this assumption no longer holds. Some of the material traits are intuitive and challenging to precisely define, something that would be required if multiple annotators are to be able to provide consistent annotations. Furthermore, it is difficult to evaluate whether or not any given set of manually-defined material traits is complete. We show that we may address both of these issues by automatically discovering useful visual material attributes. We use the term attributes to highlight the distinction between named material traits and the unnamed properties we discover. We derive a method to define an attribute space that faithfully encodes our own human perceptual representation of materials while simultaneously serving as an intermediate representation for material recognition. Our method produces attributes with the same desirable properties as visual material traits using only a small amount of easily-obtained weak supervision.

Our automatic perceptual attribute discovery method requires only simple supervision and eliminates the need to manually define a set of material traits. The training process is, however, relatively slow and does not scale well to larger datasets. Working well with small amounts of training data is a benefit, but we would ideally like to leverage recent

Chapter 1: Introduction
advances in large-scale end-to-end learning as well. As a step towards this goal, we show that our perceptual material attributes can in fact be discovered within a Convolutional Neural Network (CNN) framework focused on local material recognition (the Material Attribute/Category CNN, MAC-CNN). This enables us to take advantage of potentially larger material datasets. We also find interesting parallels with the material representation in the human material recognition process as observed in neuroscience [25, 22]. In contrast to the intermediate representations formed by our previous attribute methods, the human material recognition process (as well as our MAC-CNN) produces a perceptual representation (material attributes) as a side-product of material category recognition. Our results show that we are able to discover similar perceptual attributes using the MAC-CNN, and we additionally demonstrate the usefulness of perceptual material attributes for transfer learning. To support these experiments, we introduce a new material database focused on local material recognition.

Finally, having shown that we may separate material appearance from context using small local image patches, we introduce a novel material recognition framework that integrates local material appearance and global scene context, in the form of object and place category probabilities, to accurately recognize materials given far fewer examples than required by existing methods. Specifically, we propose a fully-convolutional full-resolution CNN that combines local material appearance and global context to generate per-pixel material category predictions. Our method achieves state-of-the-art accuracy scores on multiple material recognition datasets. Furthermore, we quantitatively investigate the informative properties of various forms of contextual cues as they pertain to the recognition of materials, and evaluate the impact of each form of context we introduce to the recognition process.
1.1 Contributions

This thesis encompasses a number of significant contributions to the field of material recognition. These contributions include:

Methods

- A framework for recognizing local visual material properties (visual material traits)
- An attribute discovery method that automatically builds a set of classifiers for attributes which encode the human perception of materials
- An end-to-end trainable CNN-based framework (MAC-CNN) for unifying discovered attributes and material recognition
- A dense per-pixel material recognition method which integrates local appearance and global context to accurately recognize materials from fewer training examples

Datasets

- Visual Material Traits (trait mask annotations)
- Material Patch Similarities (pairwise binary similarity annotations)
- Local Material Recognition Database (images and associated material masks)
  - A three-level hierarchy of material categories from which material datasets may be built
Chapter 2: Related Work

Our overall goal is to predict the presence of material categories (e.g. fabric, metal, plastic, etc...) in natural images. Here, we will review prior work involving general material recognition methods and relevant image understanding tools, such as semantic/non-semantic attributes and Convolutional Neural Networks (CNN).

2.1 Material Recognition

Textures are visual patterns associated with a specific combination of material, illumination, and surface geometry. Though textures are not materials, texture recognition methods formed the basis for early material recognition methods. Leung and Malik [35] first introduced textons to describe and classify images of textures. A texton represents a particular set of responses for a fixed hand-designed filter bank applied to an image. Texture recognition methods focused on using the distribution of exemplar textons within images to represent texture categories. Later methods, such as that of Varma and Zisserman [54], achieved extremely high accuracy scores (90-100%) on the databases available at that time. These databases, however, typically consisted of extremely specific texture categories like “crumpled paper” or “ribbed paper”, and contained images of flat surfaces, exhibiting solely the texture in question, taken under controlled laboratory conditions. The one exception is a database with only 37 images labeled with 6 categories: air, building, car, road, vegetation, trunk.

Adelson [2] first suggested materials as a distinct concept from objects or simple textures when discussing “things vs. stuff”. “Things” refers to objects, which have been the focus of
much prior work under the field of object recognition. Adelson points out that the world does not just consist of discrete objects, but also includes “stuff”, substances without a natural shape or fixed spatial extent. Ice cream is one example of “stuff” that is not an object but is still a recognizable concept in images. While materials are not equivalent to the “stuff” discussed in his work, the work does lay the foundation for material recognition as a vision problem.

The first collection of material category images for classification originated in Sharan et al. [49] where they introduced a new image database (the Flickr Materials Database or FMD) containing images from the photo sharing website Flickr. The FMD contains a set of images each with a single material annotation and corresponding mask identifying the presence of that material. Building on the FMD, Liu et al. [37] created a framework to recognize these material categories using a modified LDA probabilistic topic model. Hu et al. [26] improved upon the state-of-the-art FMD accuracy using kernel descriptors and large-margin nearest neighbor distance metric learning. Their experiments showed that providing explicit object detection information to material category recognition results in a large improvement in accuracy. Sharan et al. [48] later showed that without information associated with objects (such as the object shape), performance degrades significantly (from 57.1% to 42.6%). Specifically, they note that their material category recognition method depends heavily on non-local features such as edge contours. It is this dependency which we wish to either remove or make explicit with our proposed local material recognition methods. Zhang et al. [60] have shown further-improved performance on the FMD, but they require an auxiliary training dataset which contains a number of images that are extremely similar to those in the FMD.

All prior work discussed above produces a single category prediction for each input im-
age. This inherently assumes that there is only one material of interest in the image, a very restrictive assumption. To relax this assumption, recent work focuses on dense prediction: providing a material category for each pixel in the input image. Bell et al. introduced the OpenSurfaces [7] and MINC [6] datasets to aid in the training of dense material recognition models. With MINC they also describe a simple modification of the VGG CNN architecture of Simonyan and Zisserman [52] to predict their material categories at each pixel. Zhang et al. [59] improved the state-of-the-art accuracy for a subset of the MINC dataset, MINC-2500, using their deep texture encoding network, but their method is limited to single per-patch predictions. Cimpoi et al. [11] aggregate texture descriptors within region proposals, similar to R-CNN [21], for material recognition. They refine their predictions with a dense CRF, but if the region proposals fail to separate two materials their method cannot recover. Wang et al. [55] also demonstrate accurate dense per-pixel material predictions using 4D light field images. These datasets and models have inherent drawbacks involving their category selection and training procedures which we will discuss in later chapters.

2.2 Attributes

Attributes, as used in machine learning and computer vision, are distinctive properties (visual properties, in the context of computer vision) of categories in a classification problem. Attributes are often shared across a sparse subset of the associated categories. In the case of materials, these attributes include visual properties like “shiny”, or “smooth”. As part of our early investigation of local material recognition, we introduce two forms of material attributes as intermediate representations: fully-supervised (Chapter 3) and weakly-unsupervised (Chapter 4) visual material properties.
2.2.1 Fully-Supervised Attributes

Fully-supervised visual attributes have been widely used in object and scene recognition, but largely at the image or scene level. Ferrari and Zisserman [19] introduced a generative model for certain pattern and color attributes, such as “dots”, or “stripes”. The attributes described in their model focus on texture and color, but are not material attributes. A paper cup, for example, may have stripes painted on it, but “striped” is not a property of the paper itself. Kumar et al. [29] proposed a face search engine with their attribute-based FaceTracer framework. FaceTracer uses SVM and AdaBoost to recognize attributes within fixed facial regions. Such fixed regions are not present in materials, which may take on an arbitrary shape unlike the objects which they make up. Farhadi et al. [17] applied attributes to the problem of object recognition. Their results showed an improvement in accuracy over a basic approach using texture features. Lampert et al. [30] also showed that attributes transfer information between disjoint sets of classes. These results suggest that attributes can serve as an intermediate representation for recognition of the categories which exhibit them. Patterson and Hays [42] showed that they could recognize a variety of visual attributes, some of which happen to be general material categories. Their work, however, was not an explicit attempt at recognizing materials.

2.2.2 Weakly-Supervised Attributes

The attributes described above were all fully-supervised or “semantic” attributes. A semantic attribute is one to which we can assign a name like “round” or “transparent”. While these attributes are useful, it is difficult to quantify the completeness and consistency of any given attribute set: does the set of attributes contain everything that could help recognize the target categories, and can the appearance (for visual attributes) be agreed upon by
a variety of annotators? Semantic attributes are also task-specific and must be manually defined for each new recognition task.

To address the issues inherent to semantic attributes, a number of unsupervised or weakly-supervised attribute discovery methods have been proposed. Berg et al. [8] described a framework for automatically learning object attributes from web data (images and associated text). This approach learns some localized attributes (as we would require for local material recognition). The required text annotations are, however, image-wide and do not guarantee locality. Patterson and Hays [42] also proposed a process to discover and recognize scene-wide attributes in natural images. While they are able to discover a large amount of attributes, their learned attributes are not local. Rastegari et al. [43] learn a binary attribute representation (binary codes) for images. As with most existing methods, however, these attributes are image-wide and not local. Cimpoi et al. [10] demonstrated a method for learning an arbitrary set of describable texture attributes based on terms derived from psychological studies. As noted by Adelson [2], texture is only one component of material appearance, and cannot alone describe our perception of materials. Though their results demonstrate impressive performance on the FMD, their learned attributes apply only globally. Most relevant to the work discussed in this thesis are the attribute discovery methods of Akata et al. [4] and Yu et al. [58]. Akata et al. [4] formulated attribute discovery as a label embedding problem. Yu et al. [58] proposed a two-step procedure for discovering and classifying attributes based on a similarity matrix. They computed a distance matrix using Euclidean distances in the raw feature space of labeled image patches. In contrast, we embed the material categories in an attribute space derived from our own human visual perception of material similarity.

Chapter 2: Related Work
2.3 Material Perception and Convolutional Neural Networks

As the final step in the scaling of material attribute learning, we discover perceptual material attributes within Convolutional Neural Networks (CNNs). We also formulate the integration of material appearance and context as a CNN. Introduced by LeCun et al. [31] for handwritten digit classification, the convolutional neural network model is a general non-linear model which applies a set of convolution kernels to an image in an hierarchical fashion to generate a category probability vector. The kernel weights are model parameters that are set via non-linear optimization (generally Stochastic Gradient Descent) to attempt to maximize the likelihood of a set of training data.

Recently, Shankar et al. [47] proposed a modified CNN training procedure to improve attribute recognition. Their “deep carving” algorithm provides the CNN with attribute pseudo-label targets, updated periodically during training. This causes the resulting network to be better-suited for attribute prediction. Escorcia et al. [15] show that known semantic attributes can also be extracted from a CNN. They show that attributes depend on features in all layers of the CNN, which will be particularly relevant to our investigation of perceptual material attributes in CNNs (Chapter 5). ConceptLearner, proposed by Zhou et al. [63] uses weak supervision, in the form of images with associated text content, to discover semantic attributes. These attributes correspond to terms within the text that appear in the images. All of these frameworks predict a single set of attributes for an entire image, as opposed to the per-pixel attributes which we introduce in this thesis.

At the intersection of neuroscience and computer vision, Yamins et al. [57] find that feature responses from high-performing CNNs can accurately model the neural response of the human visual system in the inferior temporal (IT) cortex (an area of the human brain that responds to complex visual stimuli). They perform a linear regression from CNN
feature outputs to IT neural response measurements and find that the CNN features are good predictors of neural responses despite the fact that the CNN was not explicitly trained to match the neural responses. Their work focuses on object recognition CNNs, not materials. Hiramatsu et al. [25] take functional magnetic resonance imaging (fMRI) measurements and investigate their correlation with both direct visual information and perceptual material properties (similar to the material traits we introduce in Chapter 3) at various areas of the human visual system. They find that pairwise material dissimilarities derived from fMRI data correlate best with direct visual information (analogous to pixels) at the lower-order areas and with perceptual attributes at higher-order areas. Goda et al. [22] obtain similar findings in non-human primates. These studies suggest the existence of perceptual attributes in human material recognition, but do not actually derive a process to extract them from novel images.

2.4 Dense Prediction

Dense prediction, outputting a value or category prediction for each pixel, has been extensively studied in the context of object recognition and object semantic segmentation. Object recognition datasets, such as ImageNet [46] or MS COCO [36], often contain many (80-1,000) categories. Despite this, state-of-the-art semantic segmentation methods such as DeepLab [9] focus on only a small subset of coarse-grained categories. While we might gain some small contextual cues from such coarse categories, intuitively we would expect that the more detailed the context categories are, the more they will be able to inform material recognition. We show that this is indeed true in Chapter 6. A notable and relevant exception is the recent ADE20k dataset, scene parsing challenge, and associated models [65]. The dataset contains many fully-segmented images, and the challenge defines a set of 150 categories for semantic segmentation. We find the ADE20k models to be ideal sources of

Chapter 2: Related Work
per-pixel object category context information.

2.5 Context in Visual Recognition

The use of context as a means to reduce ambiguity, whether in materials or other cases, appears promising. Hu *et al.* [26] showed that a simple addition of object category predictions as features could potentially improve material recognition. Iizuka *et al.* [27] use scene place category predictions to improve the accuracy of greyscale image colorization. Shrivastava and Gupta [51] investigate the use of semantic segmentation to augment Faster R-CNN. In this case, the semantic segmentation network is trained with R-CNN in a multi-task learning fashion. The semantic segmentation network provides an additional signal for object recognition, but this is not the same thing as context: the semantic segmentation network is producing output for the same type of category (objects) as the main network. Our work, in contrast to these previous methods, takes advantage of multiple sources of context that are not merely additional forms of material recognition.
Chapter 3: Visual Material Traits

In Chapter 6, we show that separating materials from their surrounding context allows us to combine them with accurately-recognized information about said context for improved accuracy. Prior to doing so, we must first show that we can indeed recognize materials in the absence of global context like object shape or scene properties. We refer to such recognition in the absence of context as **local material recognition**.

Recognizing materials is an inherently challenging problem, made more so by our goal of local material recognition. As Figure 3.2 shows, previous material recognition frameworks rely heavily on context cues present in large image patches. As the patch size is reduced, materials become more difficult to recognize for their frameworks. One contributing factor to this difficulty is the intra-class appearance variation present in typical material categories. A car, for example, often has a very distinct boundary shape that allows for its identification as an object. On the other hand metal, a material present in most cars, can take on a variety of appearances depending on the surroundings. Figure 3.1 contains a visual example of such variation. Each image contains a sample of plastic material, but the material appearance varies based on the object and the surrounding scene conditions.

Looking at the images in Figure 3.1, one can see that plastic tends to have properties that are associated with a distinct visual appearance, such as “smooth” and “translucent”. Our key observation is that these visual properties are recognizable even when the surrounding objects and scenes are not visible. We can use these properties to tackle the challenging variations in material appearance and recognize materials independent of context. In general, material properties can include tactile ones such as “hard,” or purely visual ones such as...
Figure 3.1: Materials like the plastic in these images exhibit a wide range of appearances depending on the object and scene, making extraction of material information without the use of object information challenging. We propose to locally recognize visual material traits, distinct appearances of material properties such as "translucent," to provide contextual cues for challenging vision tasks including material category recognition and segmentation.
Figure 3.2: When adapted to use aggregated features from local image patches, methods that perform well on full images quickly lose accuracy. This suggests that they are relying heavily on context, including object shape cues, to recognize materials.
Figure 3.3: Successfully recognized material traits. These image patches were recognized by our framework as exhibiting the indicated material traits. Even at the patch level, we can see the characteristic visual appearances of each material trait.

“shiny.” We model the local visual appearance of these characteristic material properties as a novel intermediate representation: visual material traits.

Experimental results show that visual material traits can be recognized accurately from small (32 × 32) image patches, as high as 93.1% with an average accuracy of 78.4%. To express more complex concepts, such as material categories, we may treat the distribution of material traits in a region as an image descriptor and generate a per-image material category prediction. Furthermore, material traits learned from one dataset can be recognized and used to extract material information from an entirely different set. This is in contrast with past methods [48, 26] that train and test on images taken from a single source. These results show that the representation generalizes well. We also demonstrate the use of material traits in mid-level image understanding tasks by augmenting segmentation algorithms with per-pixel material information.

3.1 Representing Material Traits

Figure 3.3 shows examples of the visual material traits recognized by our framework. Even at the local level of the example images, each visual material trait corresponds to the ap-
pearance of a characteristic material property. Ideally, recognition of these material traits will enable us to extract crucial material information from any image.

The key contribution of our material traits is their ability to encode per-pixel material information without relying on object-specific features. Material traits provide a compact, local, and discriminative encoding of material properties. To obtain a representation for these material traits, we must avoid introducing any dependence on object information in the recognition process. We accomplish this by learning the best convolutional features to describe material trait patches in an unsupervised setting. Convolutional features are ideal for this purpose as they can be applied at any point in an image, and do not encode object boundary contours. We supplement these unsupervised features with selected low-level features to describe appearance patterns that cannot be learned by the unsupervised model.

3.1.1 Convolutional Material Trait Features

Expressing the appearance of material traits poses a challenge. While intuitive, traits such as “fuzzy” can be hard to quantify. While we may attempt to do so using only existing designed features, the space of images that may be represented using these features is incomplete (as shown by our feature selection results).

Rather than rely solely on handcrafted features, we determine features associated with each material trait through unsupervised feature learning. Unsupervised learning builds a generative model for images by finding simple components that can be combined to reproduce them. Constraints, such as sparsity, force optimal model components to also act as discriminative features for classification.

Our goal is to recognize per-pixel, object-independent visual material traits. To this end, we choose to learn convolutional features so that we may extract them at any pixel in an
image. By operating in fixed local neighborhoods, convolutional features ensure that we do not encode object boundary contours. These boundary contours are the primary source of undesired object-dependent features in previous frameworks [48, 26].

We build upon the convolutional auto-encoder (CAE) model [39] to learn the feature kernels. The model defines images as the weighted sum of convolution kernel responses. Optimal filters under our model are defined by the following objective function:

\[ C = T_r + \alpha T_w + \beta T_s. \tag{3.1} \]

The objective contains three terms: a reconstruction error term \( T_r \), a weight-decay (smoothness) term \( T_w \), and a sparsity term \( T_s \). The weight-decay and sparsity terms have corresponding weights \( \alpha \) and \( \beta \), and each term acts as a constraint to help produce useful features.

Reconstruction error for \( N \) images is the squared-difference between the input images \( \mathbf{I} \) and their reconstructions \( \mathbf{R} \) using the learned features,

\[ T_r = \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{I}_i - \mathbf{R}_i \|_2^2. \tag{3.2} \]

Since the features are convolution kernels, the reconstructed images \( \mathbf{R} \) are described in terms
of the encoding in feature space $\mathbf{E}_i$ by

\begin{align*}
\mathbf{E}_i &= h(\mathbf{W} \ast \mathbf{I}_i + \mathbf{b}_e), \tag{3.3} \\
\mathbf{R}_i &= \mathbf{W}' \ast \mathbf{E}_i + \mathbf{b}_r, \tag{3.4} \\
h(x_i) &= \begin{cases} x_i & \text{if } 0 \leq x_i \leq 1 \\ 0 & \text{if } x < 0 \\ 1 & \text{if } x > 1 \end{cases} \tag{3.5}
\end{align*}

with $\ast$ representing convolution with a set of filters $\mathbf{W}$, along with bias terms $\mathbf{b}_e$ and $\mathbf{b}_r$ for the encoding and reconstruction, respectively. Some formulations force the reconstruction filters $\mathbf{W}'$ to be the transpose of the encoding filters $\mathbf{W}$. We, however, found that allowing them to be separately optimized resulted in more diverse features.

The non-linear encoding function $h(x_i)$ in Equation 3.3 contains a linear region between 0 and 1. If allowed, the combination of small encoding weights and large decoding weights could force any inputs to encode solely into this linear region. Such an encoding would result in a trivially perfect reconstruction. Weight decay, $T_w = \|\mathbf{W}\|_2^2 + \|\mathbf{W}'\|_2^2$, is a term that prevents this trivial solution by ensuring that the weights do not take on exceedingly large values.

By definition, discriminative image features do not appear everywhere in an image. Figure 3.3 shows that certain material traits, particularly “shiny,” exhibit strong local appearance cues. Sparsity constraints express this property well. Sparse features are features that are only present in a small fraction of the possible locations in each image, as measured by their presence in the encoding $\mathbf{E}_i$. As in Lee et al. [33], we enforce sparsity by penalizing
Figure 3.4: These $7 \times 7$px. convolution filters learned by the CAE represent the top three filters for the listed material traits, ranked by average presence in the testing images. The filters represent characteristic local texture and color patterns. The six filters on the right do not rank in the top three for any material trait. They exhibit significantly less texture variation than the top filters.

the difference between mean filter activations and a small constant $p$:

$$T_s = \left\| p - \frac{1}{N} \sum_{i=1}^{N} E_i \right\|^2_2. \quad (3.6)$$

To further constrain the learning process and obtain a discriminative feature set, we force a fixed number of the features to be oriented first-order Gaussian filters. Learning these filters alone will satisfy both sparsity and reconstruction constraints, but their discriminative power is limited. As shown in Table 3.1, edge filters are selected roughly half as often as the CAE-learned features.

We optimize the full objective function using L-BFGS with automatically-generated symbolic gradient evaluation.

Figure 3.4 shows a selection of the top convolution filters by the CAE, ranked by average presence in the corresponding material trait images. The filters were learned from whitened material trait image patches. The top filters appear to represent the presence or absence of...
specific local texture patterns. For comparison, the non-ranked features on the right exhibit far less texture variation.

### 3.1.2 Supplemental Features

Cybenko [12] showed that artificial neural networks, including auto-encoders such as the CAE, are capable of approximating any continuous function defined on $\mathbb{R}^n$. There are, however, local features such as HOG that are not continuous and thus cannot be learned by the CAE. These discrete features may encode important properties of material traits, such as the strong local patterns in woven material. To address this, we supplement the learned features with Local Binary Patterns (LBP), HOG features and color histograms. We do not use other low-level features, such as the edge slices and ribbons of Sharan et al. [48], as they encode object-specific information and cannot be extracted on a per-pixel basis.

The results of our feature selection process show that these additional features supplement rather than replace the CAE-learned features. As will be shown in Table 3.1 in the following analysis of feature selection, CAE features are selected on average as often as any of the supplemental features. Furthermore, our analysis in Table 3.2 shows that the CAE features play a crucial role in the application of material traits.

### 3.1.3 Groupwise Feature Selection

We would like to obtain a feature set that generalizes well to new datasets. To avoid over-fitting and improve generalization, we perform feature subset selection on the supplemental and CAE-learned features. Our final feature set contains a small number of groups of conceptually related features. Rather than separate the groups into individual elements, we select the best combination of groups to recognize each trait. This process takes advantage of the fact that two individually useless features can have predictive power when grouped to-
Table 3.1: Selected features for material traits. As “fuzziness” is characterized by fine edge patterns, oriented filters and LBP are useful. Since we define “shiny” only on areas that exhibit specular highlights, it follows that color histograms and learned convolutional filters are important features for this material trait.

<table>
<thead>
<tr>
<th>Trait</th>
<th>CAE</th>
<th>Oriented</th>
<th>HOG</th>
<th>LBP</th>
<th>Color Histograms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shiny</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fuzzy</td>
<td>•</td>
<td></td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Transparent</td>
<td>•</td>
<td></td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
</tbody>
</table>

(13 Material Traits)

| Total Uses | 7 | 4 | 6 | 9 | 7 |

together [23]. We are able to exhaustively evaluate all combinations of groups (CAE features, oriented edges, HOG, LBP, color histograms), selecting those that maximize performance on a validation set. Feature groups are not further divided, thus, for example, either all HOG features are included or none are.

Table 3.1 shows the results of our feature selection process. Features are selected fairly evenly and, as the full table shows, in disjoint sets. A particular case of note is the “shiny” material trait. Since we focus on recognizing visual material traits without dependence on object-specific information, “shiny” is synonymous with specular highlights. This may be seen clearly in Figure 3.3. While there are visual cues, such as contoured reflections on a car body, that may lead an observer to call a material “shiny,” these features are specific to the object and do not directly indicate the material trait. As a result of this, color histograms and learned convolutional filters prove to be more useful features for this material trait.

3.2 Recognizing Material Traits

For training and testing, we annotate images in the Flickr Materials Database (FMD) [49] with masks indicating regions that exhibit each material trait. From these regions, we extract 45,500 annotated patches\(^1\). We use balanced sets of positive and negative examples.

\(^1\)Our implementation uses, but is not restricted to, 32 × 32px. patches.
to train randomized decision forest (RDF) classifiers for each material trait. Though we use the same dataset as methods that include object information, our feature set and recognition process explicitly avoid object dependence.

Figure 3.5 shows the recognition results for two material traits on an image from the Berkeley Segmentation Dataset (BSDS) [38]. Note that the main object in the image, a Koala, was not present in the Flickr dataset. The FMD does not, in fact, contain any animals or any examples of animal fur. Despite this, characteristic properties of the fur and plants are accurately recognized.

Figure 3.6 contains recognition accuracies for each of the 13 material traits. Since we
Figure 3.6: Visual material trait recognition accuracy. Material traits are recognized via binary classification on a balanced training and testing set, thus random chance accuracy is 50%. Most traits are recognized well. Difficult material traits, such as metallic and transparent, are challenging due to their object- and environment-dependent appearances. Average accuracy is 78.4%.
Figure 3.7: Material trait frequency distributions. We compute the class-conditional distributions for appearance frequency of each material trait given each material category. These are stored as histograms, examples of which are shown above. Plastic is most often smooth, while stone is very rarely smooth.

predict material traits independently, and the training and testing data are balanced, random chance performance is 50% accuracy. Most material traits are recognized very accurately, however, some are challenging. “Metallic” and “transparent” have the two lowest recognition rates (66.4% and 67.0%). The appearance of these material properties depends heavily on the environment surrounding the object. In the case of a reflective metal surface or a clear glass sphere, the appearance is determined entirely by the object and its environment. As we explicitly avoid object dependence, we cannot expect to model these particular material traits with the same level of accuracy as others. Despite this, “metallic” and “transparent” are still recognized better than chance.

Material traits, as a form of visual attribute, should represent a discriminative set of appearances. To investigate this, we compute the class-conditional distributions of material traits given material categories. We use the ten categories of the FMD for this
test. For each image in each category, we sample material traits uniformly across the
masked material region in the image. Figure 3.7 shows selected distributions from the
set \( \{ p(t_i|m_j) | i \in 1 \ldots 13, j \in 1 \ldots 10 \} \). The resulting distributions do, in fact, represent the
characteristic properties of their respective material categories. Stone is often rough but
very rarely smooth (there are a small number of polished stone examples in the training
data), plastic is smooth, and foliage is organic. As material traits are purely visual, they
can occasionally produce false positives, as seen in \( p(\text{soft}|\text{stone}) \). While stone is not soft,
porous stones may have a soft appearance.

Figure 3.8 shows a set of false positive material trait recognition results. “Shiny,” with
its characteristic bright highlights, is prone to be recognized in over-exposed image regions.
Results for “metallic” show that color is a strong cue for this material trait. Though the
patches are metallic in color, the material is not in fact metallic. These are limitations of the
representation. There are a few cases where the material trait annotations are incomplete,
generally for the pervasive “smooth” material trait.
3.3 Using Visual Material Traits

Our analysis shows that we may accurately recognize material traits. The material trait distributions also show that material traits encode discriminative material information. Each material category exhibits characteristic class-conditional material trait distributions. From these results, we expect to be able to inform higher-level processes with material information from material traits. Material trait distributions allow us to recognize material categories in arbitrary images without dependence on prior object knowledge. We also demonstrate a preliminary application of material traits to the problem of segmentation.

3.3.1 Material Categories from Visual Material Traits

Sharan et al. [48] showed that material category recognition depends on object-specific information. Despite this, our class-conditional trait distributions suggest that the information encoded in material traits does provide a discriminative set of features for material category recognition. We rely on these visual material trait distributions to encode and recognize material categories.

We recognize material categories from material traits by training SVM classifier on the material trait distributions. Distributions are computed from material traits recognized in uniformly sampled random patches within material regions. We select features and train material trait classifiers using half of the FMD for training, then predict their class-conditional distributions. We further supplement the distributions, in a cascade fashion, with the output of a RDF classifier trained to directly predict the material category of a patch using our feature set. The cascade process is responsible for improvements in the more recognizable categories such as foliage (11% improvement), with minor changes in other categories. Accuracy without the cascade process is 46.5%.
Figure 3.9: Confusion matrices showing true class vs. predicted class on the Flickr Material Database and ImageNet images. Average accuracy is 49.2% in (a) and 60.5% in (b). Though metal and glass both have an appearance that is environment-dependent, glass is more accurately classified. This is likely due to the tendency of glass to create characteristic local distortions.

Using the computed class-conditional distributions, we train an SVM classifier with a histogram intersection kernel to recognize material categories. The histogram intersection kernel, defined as

\[ k(x,y) = \sum \min(x_i, y_i), \]

for histogram feature vectors \( x \) and \( y \) with elements \( x_i \) and \( y_i \), measures the similarity between two normalized histograms [5]. As the material trait distributions are histograms, they are ideally suited for the histogram intersection kernel SVM.

Figure 3.9 shows the average and per-class accuracy for our method on the FMD. We split the dataset of 1000 images in half for training and testing. Our accuracy (49.2%) does not surpass the final results of Sharan et al. (57.1%) but again, their method relies heavily on features that encode the shape of the objects. We do find that our method achieves
higher accuracy than that of theirs (42.6%) when object context is removed. These results show that material traits provide important information to the material recognition process.

To demonstrate the ability of material traits to generalize well between datasets, we collected a second set of material images from a different source: ImageNet [13]. ImageNet obtains images from a variety of sources; they are thus more diverse than solely Flickr images. We collected 3480 images from ImageNet via searches for each material category. Images without bounding boxes were discarded.

To evaluate the use of material traits for material recognition on this ImageNet dataset, we first train material trait classifiers on the full set of FMD images. We then split the ImageNet images evenly into training and test sets and compute the distributions of recognized material traits on the training and test sets. We train an SVM classifier with the histogram intersection kernel of Equation 3.7 using the distribution of material traits on the training set.

Figure 3.9 shows the average accuracy for our method on this dataset. The average accuracy of 60.5% on ImageNet images shows that material traits encode material information that depends on neither the particular type of object exhibiting a material, nor the scene context in which that material appears. While Hu et al. [26] do not provide an exact value, visual inspection of their results indicates an accuracy of roughly 60% as well.

Figure 3.10 contains three misclassification examples from ImageNet images. The stone in the first image has brown color stripes characteristic of wood. The glass in the second image looks translucent due to condensation, and translucent is a trait associated with plastic more than glass. The final image is a misclassification due to localization. The ImageNet database only provides object bounding boxes, not masks. This box contains mostly smooth regions and light colors, traits representative of paper.
Figure 3.10: Three misclassified ImageNet images, with true classes for each prediction is in parentheses. The left two are a result of confusing appearances (striped and translucent are more often associated with wood and plastic respectively) while the rightmost is due to the bounding box poorly fitting the object.

Table 3.2: Performance breakdown. FS: feature selection, SF: supplemental features, CAE: convolutional auto-encoder features. For the first row we performed direct material category recognition using the concatenation of all feature sets. This shows that the trait representation is indeed providing crucial information.

<table>
<thead>
<tr>
<th>FS</th>
<th>Traits SF</th>
<th>CAE</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>34.2%</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td></td>
<td>43.5%</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>42.5%</td>
</tr>
<tr>
<td>•</td>
<td>•</td>
<td>•</td>
<td>49.2%</td>
</tr>
</tbody>
</table>

We ran a set of tests, summarized in Table 3.2, to examine the impact of each major component of the material trait and category recognition process. The first row, accuracy when performing direct category recognition, with all features, without material traits, shows that the trait representation provides crucial information for the material recognition process. By excluding either CAE-learned features or supplemental features (HOG, LBP, Color Histograms) from the trait recognition process, we see that both feature sets are necessary in order to best represent material categories.

3.3.2 Segmenting Images with Visual Material Traits

Segmenting images is a challenging process partially because the concept of a good segmentation is subjective. In the Berkeley Segmentation Dataset (BSDS) benchmark of Mar-
tin et al. [38], evaluation relies on multiple human segmentations as ground truth, since each one is a potentially correct solution. Visual material traits, with their accurate encoding of characteristic and intuitive material properties, should contribute valuable contextual cues to this process.

As an investigation of the potential for image segmentation via material traits, we augment the Normalized Cuts (NCuts) algorithm of Shi and Malik [50] with material trait information. In their method, they treat image segmentation as a graph partitioning problem and show that the optimal solution can be obtained from the solution to a generalized eigensystem, specifically, the eigenvector $y_2$ corresponding to the second-smallest eigenvalue (the smallest eigenvalue is trivially 1 due to the properties of the matrices involved):

$$ (D - W) y = \lambda Dy , $$

where $W$ is a matrix of weights representing pairwise pixel similarities and $D$ is a diagonal matrix containing the sum of all weights for a given pixel. We add an additional term,

$$ \exp \left\{ -\frac{||t_i - t_j||_2^2}{\sigma_T} \right\} , $$

(3.9)

to the similarity score function used to obtain $W$. $t_i$ represents the predicted per-trait probabilities for pixel $i$ in the image and $\sigma_T$ is a scaling parameter. This term should cause pixels that exhibit similar material traits to be grouped together in the segmentation.

Figure 3.11 shows images segmented using the original NCuts algorithm and our modified version. The first example shows that material traits can help discriminate between regions exhibiting different material properties (fuzzy grass and rocks). The expanded border around the penguin in the second segmentation is likely due to the fact that the traits
Figure 3.11: Comparing segmentation with and without material traits. Images on the left were segmented using the original NCuts algorithm, while those on the right were segmented with our modified version. Material traits can indicate the difference between fuzzy grass in the foreground and rocks in the background, despite the fact that they have similar colors.
are recognized in part using learned convolution kernels. The size of these kernels is likely to be an important parameter for good segmentations. These results show that contextual cues from material traits can indicate regions of similar materials that should be merged, or regions that should be split despite similar color or texture.
Chapter 4: Automatically Discovering Visual Material Attributes

Material trait recognition (Chapter 3) relies on a set of fully labeled material trait examples. This assumption hinders scaling the method to larger training datasets. We also do not have a complete, mutually-agreeable vocabulary for describing materials and their visual characteristics. This makes scaling with multiple annotators difficult. Considering the images in the first column of Figure 4.1, for instance, one annotator may call them fuzzy and others may call them fluffy. People may also be inconsistent in annotating material traits. Some may only annotate the patches in the second column as smooth and others may only see them as translucent. Cimpoi et al. [10] alleviate these problems for texture recognition by preparing a pre-defined vocabulary. They may do so by focusing on apparent texture patterns like stripes and dots. Materials underlie these texture patterns (i.e., the stripes or dots on a plastic cup are still plastic) and do not follow such a vocabulary.

In this chapter we show that we can address these challenges by automatically discovering locally-recognizable material attributes. We achieve this by exploiting human perception of visual material similarity. Humans are able to reliably assess material similarity from local visual information [20]. We expect that people judge material similarity based on characteristics equivalent to visual material traits. For instance, a person would perceive an image patch of wool to be similar to that of sheep fur as both look fuzzy. Humans can look at the images in Figure 4.1 and see that images in each column share visual properties without necessarily being able to identify them. By analyzing human assessments of the visual similarities of different local material image patches, we should be able to build a classifier to recognize these implicit local visual attributes. We show that such assessments
Figure 4.1: Sample material image patches. Each column contains patches containing the same material. We would like to obtain a set of attributes that describe what makes each material look distinct. Asking annotators to simply describe the patches, however, is an ambiguous question. Patches may look similar even though the annotator cannot find a concrete word to identify the similarity. In this chapter, we show that we can probe the human perception of materials by asking only for binary visual similarity decisions: “Do these two patches look similar?”

We use crowdsourcing (Amazon Mechanical Turk) to determine the visual similarity of material categories as seen by humans. For this, we show image patches of different materials as references and ask whether other image patches from other materials look similar. These results are aggregated to compute pairwise visual distances between material categories. The idea is to identify a space of material attributes that preserves these pairwise distances while permitting reliable recognition of the attributes on local image patches. For this, we first convert the distance matrix into a category-attribute matrix that realizes desirable characteristics such as sparsity. We then train a joint attribute classifier that predicts, on average for each category, the desired attribute likelihoods. Our formulation requires no supervised labeling of attributes on training data.

There are an infinite number of random non-semantic local attributes which can be used
to recognize materials. In our work, we are specifically discovering those that underlie human perception. We show that humans agree on a common perception of similar materials, and that we can in fact encode this perception in our discovered attributes. The discovered attributes show clear correlations with known visual material traits. Due to the constraints we impose on the learning process, the discovered attributes also exhibit similar spatial sparsity patterns to those that are characteristic of known traits. This is in contrast with random attributes which exhibit no such patterns. Our framework requires only simple annotations that can be quickly and consistently collected. Unlike previous methods, we do not rely on visual properties that are only visible at a global image level.

4.1 Perceptual Distance between Materials

Our goal is to discover a set of attributes that exhibit the desirable properties of material traits. We want to achieve this without relying on fully-supervised learning. Known material traits, such as “smooth” or “rough,” represent visual properties shared between similar materials. We expect that attributes that preserve this similarity will satisfy our goal. We propose to define a set of attributes based on the perceived distances between material categories. By working with distances rather than similarities, we avoid any need to assume a particular similarity function. For this, we obtain a measurement of these distances from human annotations.

From a high-level perspective, our attribute discovery consists of three steps:

1. Measure perceptual distances between materials

2. Define an attribute space based on perceptual distances

3. Train classifiers to reproduce this space from image patches

Defining perceptual distance between material categories poses a challenge. If each material
had a single typical appearance (e.g., if metal was always shiny and gray), we could simply compute the difference between these typical appearances. This, however, is not the case: materials may exhibit a wide variety of appearances, even sharing appearances between categories (what we refer to as material appearance variability). An image patch from a leaf, for example, may appear similar to certain fabrics or plastics.

Directly measuring distances via human annotation would be ideal, as we have an intuitive understanding of the differences between materials. As Sharan et al. [48] showed, this understanding persists even in the absence of object cues. It is, however, also a difficult task to obtain these distances. Given two query image patches, annotators would have to decide how different the patches look on a consistent quantitative scale. We would instead like to ask simple questions that can be reliably answered.

We propose that instead of asking how different patches look, we reduce the question to a binary one: “Do these patches look similar or not?” We expect that this will give us sufficient information to obtain consistent and sensible perceptual information. Our underlying assumption for this claim is that if two image patches look similar, they do so as a result of at least one shared visual material trait.

To transform a set of binary similarity annotations into pairwise distances, we represent each material as a point defined by the average probabilities of similarity to each material category. The pairwise distances between these points define the material perceptual distance matrix. This process treats each material category as a point in a space of typical (but not necessarily realizable) material appearances. The resulting distance between a pair of materials depends on joint similarity with all material categories, including the pair in question, and is thus robust to material appearance variability.

Formally, given a set of $N$ reference images with material category $c_n \in \{1\ldots K\}$, we
obtain binary similarity decisions \( s_n \in \{0, 1\}^K \) for each reference image against a set of sample images from each category. We represent each material category in the space of typical material category appearances as \( K \)-dimensional vectors \( p_k \):

\[
p_k = \frac{1}{N_k} \sum_{n|c_n = k} s_n, \tag{4.1}
\]

where \( N_k = |\{c_n|c_n = k\}|. \) Entries \( d_{kk'} \) in the \( K \times K \) pairwise distance matrix \( D \) are then defined as:

\[
d_{kk'} = \|p_k - p_{k'}\|_2. \tag{4.2}
\]

We obtain the required set of binary similarity annotations through Amazon Mechanical Turk (AMT). Each task presents annotators with a reference image patch of a given material category (unknown to the annotator) and a row of random image patches, one from each material category. We use patches from images of the 10 material categories from the Flickr Materials Database of Sharan et al. [49]. Annotators are directed to select image patches that look similar to the reference. Examples of suggested similar image patches are given based on known material traits. Each set of patches is shown to 10 annotators, and final results are obtained from a vote where at least 5 annotators must agree that the patches look similar. We collect similarity decisions for 10,000 reference image patches.

The 2D projection in Figure 4.2 shows that the similarity values obtained from the AMT annotations agree with our own intuitive understanding of material appearance. The plot shows the locations of material categories projected into one 2D subspace of the 10-dimensional space of typical material appearances. We would expect that the two materials corresponding to the typical materials in each subspace will lie close to their respective axes.
Figure 4.2: Example projections of materials into a 2D similarity subspace. The locations of the two material categories corresponding to the axes are marked. We would expect that, in this case, water would lie furthest along the “water” axis and likewise with leather. Materials with common visual properties, such as the smoothness of plastic and glass, lie close to each other. Materials with distinct visual properties, such as woven fabric and shiny metal, do not.
In this case, water is most similar with itself, but is also similar to glass. Leather is likewise most similar with itself, but also similar to fabric.

To show that we do in fact obtain a consistent distance matrix, we compute the difference between the distance matrix computed with all annotations versus that from only \( n \) of the \( N \) total annotations. The difference drops quickly (within the first few hundred samples of 10,000), showing that annotators agree on a single common set of perceptual distances.

4.2 Defining the Material Attribute Space

Discovering attributes given only a desired distance matrix poses a challenge. A straightforward approach would be to directly train classifiers to predict attributes that encode the distance matrix. This would be a particularly under-constrained problem as we do not even know which attributes to associate with which categories.

We instead propose to separate attribute association and classifier learning into two steps. First, we discover attributes in an abstract form by discovering a mapping between categories and attribute probabilities. We ensure that the mapping preserves the pairwise perceptual material distances, and then train classifiers to predict the presence of these attributes on image patches.

As described in Section 3, we obtain a distance matrix \( D \) from crowdsourced similarity answers for \( K \) material categories \( C = \{1 \ldots K\} \). Using \( D \), we find a mapping that indicates which attributes are associated with which categories. The number of attributes we discover is arbitrary, and we refer to it as \( M \). The mapping is encoded in the \( K \times M \) category-attribute matrix \( A \). We restrict values in \( A \) to lie in the interval \([0, 1]\) so that we may treat them as conditional probabilities.

We impose two constraints on the category attribute mapping. \( A \) should map categories to attributes in a way that preserves the measured distances in \( D \), and the mapping should
contain realizable values. If the values in $A$ are not plausible, we will not be able to recognize the attributes on image patches. For example, one potential attribute mapping would be to assign each attribute to a single category. Attribute recognition then becomes the same as the intractable problem of material category recognition on single image patches.

We formulate the attribute discovery process as a minimization problem over category-attribute matrices $A$:

$$A^* = \arg \min_A d(D; A) + w_A \kappa_A(A) \quad (4.3)$$

with hyperparameter $w_A$. $d$ describes how well the current estimate of $A$ encodes the pairwise perceptual differences between material categories, and $\kappa_A$ is a constraint that makes the discovered attribute associations exhibit a realizable distribution.

The category-attribute matrix that best encodes the desired pairwise distances will minimize the following term defined over rows $a_k$ of the matrix $A$:

$$d(D; A) = \sum_{k,k' \in C} (\|a_k - a_{k'}\|_2 - D_{kk'})^2 . \quad (4.4)$$

To discover realizable attributes, we encode our own prior knowledge that recognizable attributes exhibit a particular distribution and sparsity pattern. We observe that semantic attributes, specifically visual material traits, have a Beta-distributed association with material categories. Generally, a material category will either strongly exhibit a trait or it will not exhibit it at all. Intermediate cases occur when a material category exhibits a particularly wide variation in appearance. Fabric, for example, sometimes has a clear “woven” pattern but, in the case of silk or other smooth fabrics, does not. We would like the values in $A$ to be Beta-distributed to match the distribution of known material trait associations.
The canonical method for matching two distributions is to minimize a divergence measure between them. To incorporate this into a minimization formulation, we need a differentiable measurement for the unknown empirical distribution of values in \( A \). We choose the KL-divergence and Gaussian kernel density estimator. The Gaussian kernel density estimate at point \( p \) is:

\[
q(p; A) = \frac{1}{KM} \sum_{k,m} (2\pi h^2)^{-\frac{1}{2}} \exp \left\{ -\frac{(a_{km} - p)^2}{2h^2} \right\}
\] (4.5)

The KL-divergence between the distribution of the values in the category-attribute matrix \( A \) and the target Beta distribution \( \beta(p; a, b) \) with \( a = b = 0.5 \) can then be written as:

\[
\kappa_A(A) = \sum_{p \in P} \beta(p; a, b) \ln \left( \frac{\beta(p; a, b)}{q(p; A)} \right).
\] (4.6)

### 4.3 Training a Material Attribute Classifier

We now must derive classifiers that recognize the attributes defined by the category-attribute mapping. As attributes are not defined semantically, we cannot ask for further annotation to label training patches with attributes. Instead, we propose a model and a set of constraints that will enable us to predict our discovered attributes on any material image patches.

We do not know \textit{a priori} any particular semantics or structure associated with the attributes, thus we model our attributes using a general two-layer non-linear model [12]. We constrain the predictions such that they reproduce the desired values in the attribute matrix (in expectation) while also separating material categories when possible.

Formally, given a training set of \( N \) image patches represented by \( D \)-dimensional raw feature vectors \( x_n \) with corresponding material categories \( c_n \in C \), we train a model \( f \) with parameters \( \Theta \) that maps an image patch to \( M \) attribute probabilities: \( f(x_n; \Theta) : \mathbb{R}^D \rightarrow \mathbb{R}^M \)
[0, 1]^M. Given an intermediate layer with dimensionality $H$ and parameters $W_1 \in \mathbb{R}^{H \times D}$, $W_2 \in \mathbb{R}^{M \times H}$, $b_1 \in \mathbb{R}^H$, $b_2 \in \mathbb{R}^M$ the prediction for an instance $x_n$ is defined as:

$$f(x_n; \Theta) = h(W_2 h(W_1 x_n + b_1) + b_2)$$

$$h(x) = \min(\max(x, 0), 1). \quad (4.7)$$

As additional regularization, used only during training, we mask out a random fraction of the weights used in the model to discourage overfitting (akin to dropout [24]).

We formulate the full classifier training process as a minimization problem:

$$\Theta^* = \arg\min_{\Theta} r(X; A, \Theta) + w_1 \kappa(X; \Theta) - w_2 \pi(X; A, \Theta), \quad (4.8)$$

with hyperparameters $w_1$ and $w_2$. $r$ (Equation 4.9) is a data term indicating the difference between predicted and expected attribute probabilities. $\kappa$ and $\pi$ (Equations 4.10 and 4.11) are, respectively, constraints on the distribution of attribute predictions and on the pairwise separation of material categories.

The category-attribute matrix encodes the probabilities that each category will exhibit each attribute. We represent this in our classifier training by matching the mean predicted probability for each attribute to the given entry in the category-attribute matrix:

$$r(X; A, \Theta) = \sum_{k \in C} \left\| a_k - \frac{1}{N_k} \sum_{i: c_i = k} f(x_i; \Theta) \right\|^2_2. \quad (4.9)$$

Equation 4.9 directly encodes the desired behavior of the classifier, but it alone is under-
Figure 4.3: t-SNE [53] embedding of materials from the raw feature space (a) and from our discovered attributes (b). We embed a set of material image patches into 2D space via t-SNE using raw features and predicted attribute probabilities as the input space for the embeddings. Though t-SNE has been shown to perform well in high-dimensional input spaces, it fails to separate material categories from the raw feature space. Material categories are, however, clearly more separable with our attribute space.

We have observed that, similar to category-attribute associations, predicted probabilities for known material traits are also Beta-distributed. Local image regions exhibiting a trait will have uniformly high probability for that trait, only decreasing around the trait region edges. We constrain the predicted probabilities such that they are Beta-distributed. Using the formulation discussed in Section 4.2, we again minimize a KL-divergence of a kernel density estimate:

$$\kappa(X; \Theta) = \sum_{p \in P} \beta(p; a, b) \ln \left( \frac{\beta(p; a, b)}{q(p; f(X; \Theta))} \right),$$  \hspace{1cm} (4.10)

where $f(X; \Theta)$ represents the $N \times M$ matrix of attribute probability predictions for the training dataset, and $q, a, b$ are defined as in Equation 4.6.

One of the goals for our attribute representation is to discover attributes that allow for
material classification. If this were our only goal, we could simply maximize the distance between the predicted attributes for all pairs of different material categories. This would conflict with our goal of preserving human perception, as material categories do not always exhibit different appearances. We instead modify this separation by weighting each component of the distance based on the values in the category-attribute matrix:

\[
\pi(X; A, \Theta) = \sum_{i,j \in N|c_i \neq c_j} p_{ij} \mathbf{p}_{ij}^T \mathbf{p}_{ij} \tag{4.11}
\]

\[
p_{ij} = \left(2 |a_{c_i} - a_{c_j}| - 1\right) \left(f(x_i; \Theta) - f(x_j; \Theta)\right).
\]

This separates the material categories in attribute space only when the attributes dictate that there is a perceptual difference.

### 4.4 Analysis of Discovered Attributes

To analyze the properties of attributes discovered by our framework, we follow the procedures outlined above to collect annotations and discover a set of attributes. Since both learning steps involve minimization of a non-linear, non-convex function, we rely on existing optimization tools\footnote{Specifically, L-BFGS with box constraints for \(A\) and stochastic gradient descent for \(\Theta\).} to find suitable estimates. As a raw feature set, we use the local features we developed for material trait recognition (Section 3.1).

If our attributes described a space that successfully separates material categories, we would expect categories to form clusters in the attribute space. To verify this, we compute a 2D embedding of a set of labeled image patches. For the embedding, we use the t-SNE method of van der Maaten and Hinton [53]. t-SNE attempts to generate an embedding that matches the distributions of neighboring points in the high- and low-dimensional spaces.
Figure 4.4: Per-pixel discovered attribute probabilities for four attributes (one per column). These images show that the discovered attributes exhibit patterns similar to those of known material traits. The first attribute, for example, appears consistently within the woven hat and the koala; the second attribute tends to indicate smooth regions. The last two columns show we are discovering attributes that can appear both sparsely and densely in an image, depending on the context. These are all properties shared with visual material traits.
Figure 4.5: Typical per-pixel attribute probabilities based on a random attribute matrix. Unlike the predictions for attributes derived from human perception, these attributes appear randomly within a region and do not reflect any local visual properties.

In Figure 4.3, we represent image patches by their raw feature vectors (a) and predicted attribute probability vectors (b), and compare the 2D embeddings resulting from each. Material categories are separated much more clearly in our attribute space than in the raw feature space.

Part of the usefulness of visual material traits, as we have shown above, is derived from the fact that they each represent a particular intuitive visual material property. This is evident in the spatial sparsity pattern of the traits, specifically the fact that they appear in regions and not randomly within an image. Traits such as “shiny” are highly localized, while others such as “woven” or “smooth” exist as coherent regions within a particular material instance. Figure 4.4 shows examples of per-pixel attribute probabilities predicted from our discovered attribute classifiers. The attributes exhibit both sparse and dense spatial patterns that are consistent within local regions. Dense attributes generally correspond with smooth image regions. Sparse attributes often indicate localized surface features such as specific texture patterns.

For comparison, in Figure 4.5 we visualize per-pixel predictions for an attribute classifier
Figure 4.6: Correlation between discovered attribute predictions and material traits. Groups of attributes can collectively indicate the presence of a material trait. Metallic, for example, correlates positively with attribute 0 and negatively with attribute 8.

We aimed to discover attributes similar to the visual material traits that underlie human perception. We thus expect that the discovered attributes exhibit a correlation with known traits. Figure 4.6 shows the correlation between 13 discovered attributes and 13 known material traits using attributes predicted on labeled material trait image patches. Collectively, we can indeed describe material traits using the discovered attributes. Visually similar traits, such as rough and woven, show similar correlations with the attributes. Discovered attributes are also consistent with the semantic properties of material traits. Rough and
smooth are mutually exclusive traits, and we see that discovered attributes that positively correlate with smooth do not generally correlate with rough.

We quantitatively evaluate the discovered attributes using logic regression [45]. Given a set of image patches with known traits, we predict our discovered attributes as binary values for use as input variables in a logic regression model for material traits. Logic regression from 30 attributes alone (no other features) achieves comparable accuracy to our trait-based method and its complex feature set. These results show that the discovered attributes do collectively encode intuitive visual material properties.

4.5 From Discovered Attributes to Materials

Seeing that discovered attributes encode visual material properties, we would expect them to also serve as an intermediate representation for material category recognition. To test this, we follow our local material recognition procedure (Section 3.3.1), substituting our discovered attributes in place of labeled material traits. We compute the histograms of these predicted probabilities across the material region and use them as input for a histogram kernel SVM. As we focus on local attributes, these previous local results (and those of Sharan et al. [48] on scrambled images) serve as the correct baseline.

To compare with our previous results using material traits, we compute average material recognition accuracy on the Flickr Materials Database (FMD). All results are computed using $M = 30$ discovered attributes and 5-fold cross-validation unless otherwise specified.

Our attributes achieve an average accuracy of 48.9% ($\sigma = 1.2\%$) on FMD images using only local information. This is comparable to our results and those of Sharan et al. [48] (using only local information) even though we are discovering attributes using only weak supervision.

Figure 4.7 shows a confusion matrix for FMD images. In agreement with previous work,
Figure 4.7: Confusion matrix for material recognition on FMD images. Well-recognized categories, such as foliage, correspond with categories that appeared distinct in human annotations for perceptual distance. Annotators regularly selected foliage patches as appearing different from all other categories.
metal is the most challenging category to identify. Foliage is very well-recognized. This follows from the results of our measurements of human perception, as annotators consistently found that foliage image patches looked different from all other material categories. Fabric was previously somewhat challenging to recognize locally, and we see that paper is also challenging in this case. It is possible that subtle cues separating paper and plastic were not visible to the annotators.

Figure 4.8 shows that accuracy reaches a plateau as the training dataset size increases. We also compute accuracy for varying values of $M$ and find that past $M = 30$, there is little
(<0.1%) gain in accuracy from additional attributes. These plateaus indicate that we are in fact extracting as much perceptual material information as we can from the available data.
Chapter 5: Perceptual Material Attributes in Convolutional Neural Networks

We have shown that we may use visual material traits to enable local material recognition, and we may further scale this attribute-based recognition process by automatically discovering perceptual material attributes. Our previous methods consider attributes separately from category recognition. The attributes are used solely as an intermediate representation for material categories. Similarly for conventional object and scene recognition, attributes like “sunset” or “natural,” have also been extracted for use as independent features. Shankar et al. [47] generate pseudo-labels to improve the attribute prediction accuracy of a Convolutional Neural Network, and Zhou et al. [63] discover concepts from weakly-supervised image data. In both cases, the attributes are considered on their own, not within the context of higher-level categories. In object and scene recognition, however, recent work shows that semantic attributes seem to arise in networks that are trained end-to-end for category recognition [64].

Recent neuroscience studies also reveal that human material perception, in fact, relies on internal representations that correspond to semantic material attributes. Hiramatsu et al. and Goda et al. [25, 22] have investigated how visual information is transformed in the brain during the human and animal recognition of materials. They find that the material representation in our visual system shifts from raw image features at lower levels (V1/V2) to perceptual properties (such as matte, colorful, fuzzy, shiny, etc.) in higher-level brain regions dedicated to recognition (FG/CoS). All of these recent observations both in computer vision and neuroscience share the common trend that attributes inherently arise
from recognition tasks, as opposed to forming a discrete step in that recognition process.

We would like to take advantage of the benefits of end-to-end learning to incorporate automatically-discovered attributes with material recognition in one seamless process. Material attribute recognition, however, is not easily scalable. In the past we relied on semantic attributes, such as “shiny” or “fuzzy”, that needed careful annotation by a consistent annotator as their appearance may not be readily agreed upon. We addressed the difficulty in annotation scaling by automatically discovering perceptual material attributes from weak supervision. The training process for this method does not, however, scale well to large datasets. In this chapter, we will investigate a novel CNN architecture that recognizes materials from small local image patches while producing perceptual material attributes as an auxiliary output. We also introduce a novel material database with material categories drawn from a materials-science-based category hierarchy.

5.1 Perceptual Material Attributes from Local Material Recognition

Now, we will show that perceptual material attributes we previously discovered arise naturally in a material recognition framework. This agrees with the findings of Hiramatsu et al. [25] which indicate that perceptual attributes form an integral component of the human material recognition process. Based on correlations between Convolutional Neural Network (CNN) feature maps and human visual system neural output discovered by Yamins et al. [57], a CNN architecture appears to be a very suitable framework in which to discover attributes analogous to those in human material perception. We must derive our own method to realize material attributes, however, as their work focuses on object recognition and does not extract any attributes. In this case, our perceptual attributes are particularly relevant. In this section we derive a novel framework to discover perceptual attributes similar to the ones we describe in Chapter 4 inside a material recognition CNN framework.
5.1.1 Finding Material Attributes in a Material Recognition CNN

A simple experiment to verify the presence of perceptual attributes in a CNN trained to recognize materials would be to add an attribute prediction layer at the top of the network, immediately before the final material category probability softmax layer. If we could predict attributes from this layer without affecting the material recognition accuracy, it would suggest that the attributes were indeed present in the network. We implemented this approach with the goal of predicting our previous perceptual attributes and found that, while the material accuracy was unaffected, the attribute predictions were less accurate than those of the relatively simple attribute-only model (mean average error of 0.2 vs 0.08).

The key issue with the straightforward approach is that it is not an entirely faithful model to the process described in [25]. They note that the human neural representation of material categories transitions from visual (raw image features) to perceptual (visual properties like “shiny”) in an hierarchical fashion. This implies, in agreement with findings of Escorcia et al. [15], that attributes require information from multiple levels of the material recognition network. We show that this is indeed the case by successfully discovering the attributes using input from multiple layers of the material recognition network.

5.1.2 Material Attribute-Category CNN

We need a means of extracting attribute information at multiple levels of the network. Simply combining all feature maps from all network layers and using them to predict attributes would be computationally-impractical. Rather than directly using all features at once, we augment an initial CNN designed for material classification with a set of auxiliary fully-connected layers attached to the spatial pooling layers. This allows the attribute layers to use information from multiple levels of the network without needing direct access to ev-
Figure 5.1: Material Attribute-Category CNN (MAC-CNN) Architecture: We introduce auxiliary fully-connected attribute layers to each spatial pooling layer, and combine the per-layer predictions into a final attribute output via an additional set of weights. The loss functions attached to the attribute layers encourage the extraction of attributes that match the human material representation encoded in perceptual distances. The first set of attribute layers acts as a set of weak learners to extract attributes wherever they are present. The final layer combines them to form a single prediction.

ery feature map. We treat the additional layers as a set of weak learners, each auxiliary layer discovering the attributes available at the corresponding level of the network. This is similar to the deep supervision of Lee et al. [32]. Their auxiliary loss functions, however, simply propagate the same classification targets (via SVM-like loss functions) to the lower layers. Rather than propagating gradients, our attribute layers discover perceptual material attributes.

For the auxiliary layer loss functions, we introduce a modified form of the perceptual attribute loss function (Equation 4.8) to the outputs of each auxiliary fully-connected layer. Specifically, assuming the output of a given pooling layer $i$ in the network for image $j$ is $h_{ij}$, and given categories $C$, $|C| = K$ and a set of sample points $P \in (0,1)$ for density estimation, we add the following auxiliary loss functions:

$$
u_i = \frac{1}{K} \sum_{k \in C} \left\| a_k - \frac{1}{N_k} \sum_{j \mid c_j = k} f \left( W_i^T h_{ij} + b_i \right) \right\|_1$$

(5.1)
\[ d_i = \sum_{p \in P} \beta(p; a, b) \ln \left( \frac{\beta(p; a, b)}{q(p; f(W_i^T h_{ij} + b_i))} \right), \]  

where \( f(x) = \min(\max(x, 0), 1) \) clamps the outputs within \((0, 1)\) to conform to attribute probabilities, and weights \( W_i, b_i \) represent the auxiliary fully-connected layers we add to the network. \( a_k \) represents a row in the category-attribute mapping matrix derived as in Section 4.2. Equation 5.1 causes the attribute layer to discover attributes which match the perceptual distances measured from human annotations. As certain attributes are expected to appear at different levels of the network, some layers will be unable to extract them. This implies that their error should be sparse, either predicting an attribute well or not at all. For this reason we use an L1 error norm. Equation 5.2, applied only to the final attribute layer, encourages the distribution of the attributes to match those of known semantic material traits. It takes the form of a KL-divergence between a Beta distribution (empirically observed to match the distribution of semantic attribute probabilities), and a Kernel Density Estimate \( q(\cdot) \) of the extracted attribute probability density sampled at points \( p \in P \).

The material recognition portion of our MAC-CNN is inspired by the VGG-16 network of Simonyan and Zisserman [52], but with critical modifications to enable the recognition of materials and discovery of their attributes from small local image patches. Beyond the auxiliary loss functions derived above, we also remove the last set of pooling and convolution layers from the VGG-16 model. VGG models were designed for object recognition and use large (224×224px) patches as input. The final pooling and convolution layers cannot be applied in our case, as the feature maps would be reduced to single vectors by the excessive downsampling. We use their trained convolutional weights as initialization where applicable, and add new fully-connected layers for material classification. Figure 5.1 shows our architecture for material attribute discovery and category recognition. We refer to this
network as the Material Attribute-Category CNN (MAC-CNN).

5.2 Local Material Database

In order to train the category recognition portion of the MAC-CNN, we need a proper local material recognition dataset. We find existing material databases lacking in a few key areas necessary to properly perform local material recognition. Previous material recognition datasets [48, 7, 6] have relied on ad-hoc choices regarding the selection and granularity of material categories. When patches are involved, as in [6], the patches can be as large as 24% of the image size surrounding a single pixel identified as corresponding to a material. These patches are large enough to include entire objects. These issues make it difficult to separate challenges inherent to material recognition from those related to general recognition tasks. We also find that image diversity is still lacking in modern datasets. For these reasons, we introduce a new local material recognition dataset to support the experiments in this paper.

5.2.1 Material Category Hierarchy

Material categories in existing datasets have not been carefully selected. Examples of this issue include the proposed material categories “mirror” (actually an object), and “brick” (an object or group of objects). Existing categories also confuse materials and their properties, for example, separating “stone” from “polished stone”. To address the issue of material category definition, we propose a more carefully-selected set of material categories for local material recognition. We derive a taxonomy of materials based on their properties from materials science [1] and create a hierarchy based on the generality of each material family. Figure 5.2 shows an example of one tree of the hierarchy.

Our hierarchy consists of a set of three-level material trees. The highest level corresponds to major structural differences between materials in the category. Metals are conductive,
Figure 5.2: Our proposed material category hierarchy. Categories at the top level (red) separate materials with notable differences in physical properties. Mid-level categories (green) are visually distinct. The lowest level of categories (blue) are fine-grained and may require both physical and visual properties and expert knowledge to distinguish them. In our local materials database, we collect annotations for mid-level categories only, as they correspond to names likely to be familiar with a non-technical audience. We make one exception for concrete and asphalt, as those names are more familiar than the term “composite”. We also add supplemental categories for food, water, and non-water liquids.
polymers are composed of long chain molecules, ceramics have a crystalline structure, and composites are fusions of materials either bonded together or in a matrix. We define the mid-level (also referred to as entry-level [41]) categories as groups that separate materials based primarily on their visual properties. Rubber and paper are flexible, for example, but paper is generally matte and rubber exhibits little color variation. The lowest level, fine-grained categories, can often only be distinguished via a combination of physical and visual properties. Silver and steel, for example, may be challenging to distinguish based solely on visual information.

Such a hierarchy is sufficient to cover most natural and manmade materials. For the sake of completeness, we also add three supplemental mid-level categories to our data collection process: food, water, and non-water liquids\(^1\). Though they do not follow the strict hierarchy established above, these categories appear in many natural images. Water and food are both “stuff”, not “things”, to use the terminology of Adelson [2], and as a result we consider them materials for practical purposes.

5.2.2 Data Collection and Annotation

The mid-level set of categories forms the basis for a crowdsourced annotation pipeline to obtain material regions from which we may extract local material patches (Figure 5.3). We employ a multi-stage process to efficiently extract both material presence and segmentation information for a set of images.

The first stage asks annotators to identify materials present in the image. Given a set of images with materials identified in each image, the second stage presents annotators with a user interface that allows them to draw multiple regions in an image. Each annotator is given a single image-material pair and asked to mark regions where that material is present.

\(^1\)Non-water liquids were identified extremely rarely in typical images and thus are not considered in subsequent experiments.
Figure 5.3: Local material patches extracted as the final step in our database creation process. These patches are used to compute human perceptual distances, and also form the training input for our combined material attribute-category CNN.

While not required, our interface allows users to create and modify multiple disjoint regions in a single image. Images undergo a final validation step to ensure no poorly drawn or incorrect regions are included.

Each image in the first stage is shown to multiple annotators and a consensus is taken to filter out unclear or incorrect identifications. While sentinels and validation were not used to collect segmentations in other datasets, ours is intended for local material recognition. This implies that identified regions should contain only the material of interest. During collection, annotators are given instructions to keep regions within object boundaries, and we validate the final image regions to insure this.

Image diversity is an issue present to varying degrees in current material image datasets. The Flickr Materials Database (FMD) [49] contains images from Flickr which, due to the nature of the website, are generally more artistic in nature. The OpenSurfaces and Materials in Context datasets [7, 6] attempt to address this, but still draw from a limited variety of sources. We source our images from multiple existing image datasets spanning the space of indoor, outdoor, professional, and amateur photographs. We use images from the PASCAL
Figure 5.4: Example annotation results. Annotators did not hesitate to take advantage of the ability to draw multiple regions, and most understood the guidelines concerning regions crossing object boundaries. As a result, we have a rich database of segmented local material regions.

VOC database [16], the Microsoft COCO database [36], the FMD [49], and the imagenet database [46].

Examples in Figure 5.4 show that our annotation pipeline successfully provides properly-segmented material regions within many images. Many images also contain multiple regions. While the level of detail for provided regions varies from simple polygons to detailed material boundaries, the regions all contain single materials.
5.3 Perceptual Material Attributes Discovered in the MAC-CNN

To verify that the perceptual attributes we seek are indeed present in and can be extracted with our MAC-CNN, we augment our dataset with annotations to compute the necessary perceptual distances described in Section 4.1. Using our dataset and these distances, we derive a category-attribute matrix $A$ and train an implementation of the MAC-CNN described in Sec. 5.1.2.

We train the network on $\sim 200,000$ $48 \times 48$ image patches extracted from segmented material regions. Optimization is performed using mini-batch stochastic gradient descent with momentum. The learning rate is decreased by a factor of 10 whenever the validation error increases, until the learning rate falls below $1 \times 10^{-8}$.

5.3.1 Properties of the Perceptual Material Attributes

We examine the properties of our perceptual material attributes by visualizing how they separate materials, computing per-pixel attribute maps to verify that the attributes are being recognized consistently, and linking the non-semantic attributes with known semantic material traits (“fuzzy”, “smooth”, etc...) to visualize semantic content. Figs. 5.5, 5.6, and 5.7 are generated using a test set of held-out images.

A 2D embedding of material image patches shows that the perceptual attributes (Figure 5.5) separate material categories. A number of materials are almost completely distinct in the attribute space, while a few form overlapping but still distinguishable regions. Foliage, food, and water form particularly clear clusters. The quality of the clusters matches the per-category recognition rates in Figure 5.8, with accurately-recognized categories forming more separate clusters.

Visualizations of per-pixel attribute probabilities in Figure 5.6 show that the attributes
Figure 5.5: Attribute Space Embedding via t-SNE [53]: Many categories, such as water, food, foliage, soil, and wood, are extremely well-separated in the attribute space. We find that this separation corresponds roughly with per-category accuracy as shown in Figure 5.8. While other categories do overlap to some extent, they still form separate regions in the space.

are spatially consistent. While overfitting is difficult to measure for weakly-supervised attributes, we use spatial consistency as a proxy. Spatial consistency is an indicator that the attributes are not overly-sensitive to minute changes in local appearance, something that would appear if overfitting were present. The attributes exhibit correlation with the materials that induced them: attributes with a strong presence in a material region in one image often appear similarly in others. The visualizations also clearly show that the attributes are representing more than trivial properties such as “flat color” or “bumpy texture”.

Logic regression [45] is a method for building trees that convert a set of boolean variables into a probability value via logical operations (AND, OR, NOT). It is well-suited for collections of binary attributes such as ours. Results of performing logic regression (Figure 5.7) from extracted attribute predictions to known semantic material traits (such as fuzzy, shiny, smooth etc...) show that our MAC-CNN attributes encode material traits with the roughly same average accuracy (77%) as the our previous attributes. We may also predict per-pixel trait probabilities in a sliding window fashion, showing that the attributes are encoding both
Figure 5.6: Each column after the first (the input image) shows per-pixel probabilities for an extracted perceptual attribute. The attributes form clearly delineated regions, similar to semantic attributes, and their distributions match as well.
Figure 5.7: By performing logic regression from our MAC-CNN extracted attributes to semantic material traits, we are able to extract semantic information from our non-semantic attributes. We can apply logic regression to material attribute predictions on patches in a sliding window to obtain per-pixel semantic material trait information. The per-pixel trait predictions show crisp regions that correspond well with their associated semantic traits. Traits are independent, and thus the maps contain mixed colors. Fuzzy and organic in the lower right image, for example, creates a yellow tint.
Figure 5.8: Local material recognition accuracy, by category. Average accuracy is 60.2%. It is clear that some categories, such as metal and glass, are significantly more challenging to recognize locally.

5.3.2 Local Material Recognition

If perceptual material attributes are naturally present in the material classification network, we must be able to extract them without compromising the network’s ability to recognize materials. Our results in Section 5.3 show that we can extract the perceptual attributes in the combined material-attribute network. We compare local material recognition accuracy with and without the auxiliary attribute loss functions to verify the second requirement. Figure 5.8 shows a detailed breakdown of accuracy with the auxiliary layers.
Figure 5.9: Images in each column share true material categories. The first three rows are correct predictions, and incorrect predictions (bottom two rows) are shown under the corresponding images. Glass and metal, for example, are both materials whose appearance depends heavily on the surrounding environment. Asphalt and concrete are both common paving materials and it is sensible that they are often confused.
The average accuracy is 60.2% across all categories. Foliage is the most accurately recognized, consistent with past material recognition results in which foliage is the most visually-distinct category. Paper is the least well-recognized category. Unlike the artistic closeup images of the FMD, many of the images in our database come from ordinary images of scenes. Paper, in these situations, shares its appearance with a number of other materials such as fabric. Figure 5.9 shows some individual patch examples for correct and incorrect predictions.

It is important to note that we are recognizing materials directly from single small image patches, with none of the region-based aggregation or large patches used before and by other methods [6]. This is a much more challenging task as the available information is restricted.

We find that the average material category accuracy does not change when the attribute layers are removed. While the attribute layers are auxiliary, they are connected to spatial pooling layers at every level and thus the attribute constraints affect the entire network. If the attributes were not in fact encoding visual material properties, constraining the network to extract them would negatively affect the material recognition performance.

We will properly address fully-dense per-pixel material recognition in Chapter 6. We are, however, able to use the same attribute/material CNN to produce per-pixel material probability predictions in a sliding window fashion. Results in Figure 5.10 show that we may still generate reasonable material probability maps even from purely local information.

5.4 Novel Material Category Recognition

One prominent application of attributes is in novel category recognition tasks. Examples of these tasks include one-shot [18] or zero-shot learning [30]. Zero-shot learning allows recognition of a novel category from a human-supplied list of applicable semantic attributes. Since our attributes are non-semantic, zero-shot learning is not applicable here. We may,
Figure 5.10: Applying the MAC-CNN in a sliding-window fashion leads to a set of material category probability maps. These material maps show that we may obtain coherent regions using only small local patches as input. The foliage predictions in the bottom right image are reasonable, as the local appearance is indeed a flower. In the upper right image, the local appearance of the fence resembles lace (a fabric).

However, investigate the generalization of our attributes through a form of one-shot learning, in which we use image patches extracted from a small number of images to learn a novel category.

To evaluate the application of perceptual material attributes for novel category recognition, we train a set of attribute/material networks on modified datasets each containing a single held-out category. No examples of the held-out category are present during training. The corresponding row of the category-attribute matrix is also removed. The same number of attributes are defined based on the remaining categories.

For the novel category training, we use a balanced dataset consisting of unseen examples of training categories and a matching number of images from the held-out category. We also separate a number of images of the held-out category as final testing samples. We train a simple binary classifier (a linear SVM) to distinguish between the training categories and
\textbf{Figure 5.11}: Graphs of novel category recognition accuracy vs. training set size for various held-out categories. The rapid plateau shows that we need only a small number of examples to define a previously-unseen category. The accuracy difference between feature sets shows that the attributes are contributing novel information.

The held-out category based on either their attribute probabilities, material probabilities, or both, computed on patches extracted from each input image. We measure the effectiveness of novel category recognition by the fraction of final held-out category samples properly identified as belonging to that category.

Figure 5.11 shows plots of novel category recognition effectiveness as the number of training examples for the held-out category varies. We can see that the accuracy plateaus quickly, indicating that the attributes provide a compact and accurate representation for novel material categories. The number of images we are required to extract patches from to obtain reasonable accuracy is generally quite small (on the order of 10) compared to full material category recognition frameworks which require hundreds of examples. Furthermore, we include accuracy for the same predictions based on only material probabilities instead of attribute probabilities, as well as using a concatenation of both. This clearly shows that the extracted attributes can expose novel information in the MAC-CNN that would not
ordinarily be available.
Chapter 6: Integrating Local Materials with Global Context

Our material recognition methods introduced in previous chapters focus on local material recognition in single patches. Materials, however, are present everywhere in an image, even where objects cannot be discerned. To take full advantage of material recognition, we must instead predict materials everywhere they appear, in other words, at the per-pixel level. Our earlier methods can be adapted to produce dense predictions in a sliding window fashion, but this is inefficient. Furthermore, local material recognition is subject to limitations which we discuss further in this chapter. Predicting a dense per-pixel material map is not a new problem; we refer to the problem here specifically as “material segmentation” in order to make the distinction clear that we are recognizing materials, and not arbitrary semantic categories.

A straightforward approach to material segmentation is to simply train a semantic segmentation model with material categories. This ignores the fact that materials are not just another form of categories that can be substituted for objects. Objects are defined primarily by their shape, not by their material. As a result, recognizing an object requires that one marginalize out any variation in material, since, for example, plastic cups and glass cups must both be recognized as cups. Unlike objects, materials have no inherent shape: one can describe something as “horse-shaped”, for example, but not “metal-shaped”. Following the terms of Adelson [2], materials are a kind of “stuff”, fundamentally different than objects (“things”).

As seen in Figure 6.1, when semantic segmentation methods based on large-patches fail, they do so because they rely too heavily on properties of the objects involved. At the same
Figure 6.1: Material segmentation methods based on large-patch CNNs implicitly rely on the context present in the patch to classify materials. When the context is ambiguous, this leads to errors that can be resolved using the local appearance information. Here, for example, the object is a house in an outdoor scene, but the area surrounding the windows is a painted surface, not glass. Since existing methods do not cleanly separate local appearance and context, they cannot resolve such ambiguities.
Figure 6.2: The image above, output from our MAC-CNN, shows material category probabilities for three materials: wood, foliage, and fabric, in the RGB channels respectively. Their method uses only local information; as a result the foliage pattern on the sofa is misclassified as actual foliage. This is an example where scene context is vital in resolving an otherwise ambiguous local material appearance.

time, we cannot simply ignore objects or other contextual cues, as we are not always able to distinguish one material from another purely based on local visual appearance. A white ceramic sink and a white plastic cup, for example, appear very similar locally. As we see in Figure 6.2, methods like our previous local material recognition frameworks that cannot take advantage of scene context when it is needed will fail to accurately recognize materials when that context is the only distinguishing factor.

As we show in Section 6.1 below, contextual information like object and place categories provides very strong cues as to which materials are present in an image. Given the strength of these cues, we would expect that a material recognition method could take advantage of them to recognize materials from relatively few training examples. If, for example, we knew mugs are often made of ceramic, then we would not need to see all examples of ceramic mugs to accurately recognize that material. Existing methods like those of Bell et al. [6] or Cimpoi et al. [11], however, can only see a portion of these contextual cues inside the large patches and regions on which they are trained. As a result, they require very large training
datasets to see all possible combinations of materials and the limited context available.

We show that we can indeed learn to segment materials from relatively few training examples so long as we properly separate material appearance and scene context. By providing these two components as separate streams of information, we are able to make the strong material recognition cues present in the context explicit. As a result, our material segmentation framework does not need to see every possible combination of material and context. We propose a material segmentation CNN that produces full-resolution dense material maps using only small local image patches and explicit external scene context in the form of object and place category probabilities. By training on small local image patches, we ensure that we are separating the material appearance from global scene context, and by introducing the global context from external sources, we ensure that the context contains the strong material recognition cues we know to exist.

Our experimental results show state-of-the-art material segmentation accuracy on multiple datasets. We investigate in detail the contribution of each of the contextual cues used in our model, as well as the effects of the granularity of the context categories involved. Most importantly, we show that by separating local material appearance and global context, we can learn to accurately recognize materials with less training data than existing methods.

6.1 Role of Context in Material Recognition

We have an intuitive understanding that if, for example, one sees an image of a car, it is likely to contain glass (windows), rubber (tires) and metal (body) materials. Likewise, if we know that an image was taken in a park, we are likely to see foliage and wood materials. We refer to categories like “car” and “park” as context when they are used to inform the prediction of a different domain of category, such as materials. We can quantitatively evaluate these discriminative cues by computing the conditional probability distributions of materials given
each possible category of context. If our intuition is correct, then these distributions should have a low entropy relative to the corresponding discrete uniform distribution. In our work, we focus on two forms of context categories most likely to provide useful information for material recognition: objects and places.

6.1.1 Object Context

We can get an initial idea as to how discriminative context is by using ground-truth object masks and corresponding materials to compute the conditional distribution $p(M|O)$, where $M$ is the material category and $O$ is the object category. We use our local materials database as it contains images from databases that contain object category map annotations (particularly, MS COCO [36]). To compute the conditional probabilities, we take each image with material annotations and find the object exhibiting each material as indicated by the COCO ground truth. Figure 6.3 shows conditional material probabilities $p(M|O = o)$ for two selected object categories $o$. The entropy\(^1\) for the discrete uniform distribution over 16 categories is 2.77, and as shown in Figure 6.3 the entropy given object categories is much lower.

6.1.2 Place Context

Scenes may contain multiple objects and each object tends to exhibit only a small set of materials. This contributes to the discriminative nature of the object-conditional material probabilities. In contrast, places are single scene-wide properties and can encompass many objects and materials. Despite this, we expect that places can still provide useful cues to disambiguate local materials. Ceramic and paper, for example, often appear in the form of flat white surfaces. Without seeing a specular highlight, it may be difficult to distinguish the two materials. If, however, we know that the image patch originates from an image of

\[^1\] \(H(p) = -\sum p_i \ln p_i\)

Chapter 6: Integrating Local Materials with Global Context
Figure 6.3: The conditional distributions of materials given ground-truth object categories (top row) and predicted places (bottom row) are highly discriminative. Many context categories exhibit only a small set of materials. Some outliers are inevitable as the ground-truth COCO segmentation masks do not perfectly conform to actual object boundaries in the image.
a classroom, it is much more likely that the patch contains paper.

We can evaluate the discriminative power of places by using predictions from the MIT Places CNN [62]. Figure 6.3 contains examples of the conditional distributions $p(M|P=p)$ for a few choices of place category $p$. While they are not uniformly as discriminative as object categories, they still do provide useful cues. Botanical gardens, for example, tend to contain plants as one would expect, and images of crosswalks contain asphalt, metal, and rubber (roads, cars).

6.2 Integrating Context in a Material Segmentation CNN

We might expect that even implicit scene context would reduce the amount of training examples required to recognize materials. As we have seen from the distributions above, context provides a very strong cue for the material categories present in an image. Certain combinations of context, such as "bathroom" and "sink", are almost sufficient on their own to identify a material. Even when it is not, context can help to disambiguate a material with an otherwise-challenging visual appearance, for example, showing that a white ceramic cup is in fact made of ceramic and not paper. These observations suggest that scene context, such as objects and places, should allow us to recognize materials given a relatively small set of training examples. Previous methods, however, do not show this behavior.

Existing methods, such as that of Bell et al. [6] or Cimpoi et al. [11], use large image patches, or texture descriptors aggregated within regions, to recognize material categories. These methods may only implicitly take advantage of contextual cues, such as object shape, as it happens to appear within the patches or regions on which they are trained. As a result, the models learn to recognize materials as an entangled combination of appearance and context. This suggests that such methods can only produce accurate material predictions when trained on large datasets that contain many examples of all combinations of materials.
Figure 6.4: Distribution of the ratio of probabilities for predicted vs. true categories given that the prediction was incorrect. We can see that incorrect categories, the ones we would like to change via the use of context, can have much higher probabilities than that of the true category. As a result, simple multiplication with context-conditional distributions will rarely change the classifier’s output for the better.

The conditional probability plots in Section 6.1 appear to provide very strong material recognition cues. One might expect that given such strong cues, we could simply multiply the appropriate conditional distributions with the category probability output of any existing material recognition method (essentially a simple Bayesian prior). A simple experiment with the output our MAC-CNN shows that unfortunately, the output of such methods is not suitable for this approach. Looking at the ratio of predicted probability for mispredicted classes vs. that of the true class in Figure 6.4, we can see that the mispredictions often have
significantly higher probability than the true category. As a result, a simple multiplication with a conditional probability vector will rarely change the predicted class.

Our key observation is that in order to take advantage of the strong conditional distributions of materials given their surrounding context, we must provide the scene context as a separate signal during the learning process. Doing so allows us to learn a classifier that factorizes material recognition into two separate components – local material appearance and global scene context – and thereby learn to recognize materials with far fewer training examples than existing methods.

To achieve this desired factorization of local material appearance and context within a material segmentation CNN, we must cleanly separate them in the network formulation. We accomplish this by designing the network such that during training, the network is only able to see small local image regions. This serves to remove implicit dependency on context as it may appear in larger patches. With such a network, we may then incorporate a separate source of scene context to complete the factorization. We propose to provide this scene context in the form of contextual category probabilities, e.g. places or objects, to faithfully encode the distributions observed in Section 6.1.

Given such separate streams of material appearance and scene context, the next logical question is how do we combine them to accurately recognize materials from fewer training examples. One potential approach would be to introduce the context at the lowest level of the network, by concatenating it as an additional feature channel. This could be viewed as a fusion of top-down and bottom-up feedback which has proven successful in other domains [34, 14, 40, 51]. Unlike with these contextual feedback methods, however, we propose to provide context as a separate stream known to only contain contextual information (object and place category probabilities). Introducing context at this low level is likely to result in overfitting.
Figure 6.5: Material segmentation CNN architecture. Our network takes an input image, object category probability map, and a place category probability vector as inputs. Horizontal lines represent additive skip connections, with appropriate zero-padding on the channel axis. During training, the network only sees $48 \times 48$ px image patches to ensure we are separating local material appearance from context. At test time, we may input an image of arbitrary size.

as opposed to beneficial feedback, since the context already provides such a strong cue for material recognition. This is supported by our experimental results which show that context is best introduced at higher levels in the network.

Figure 6.5 shows our proposed CNN architecture for the integration of material appearance and context. We train on $48 \times 48$ px local image patches and corresponding per-pixel object category probabilities. While our architecture resembles existing VGG-based [52] U-Net architectures [44], we make some critical modifications to enable us to separate material appearance and context. Our architecture has fewer pooling layers than typical large-patch CNNs, which cannot be trained on such small input patches. Additionally, the network is fully-convolutional by design, rather than as a post-process, allowing us to introduce context at the proper place in the network (at the higher levels). Dilated convolutions, a tool often used in object semantic segmentation networks, are not helpful when the goal is to represent local material appearance as their large receptive fields extend far past the boundaries of the local training patches. As a final step for full-image testing, we post-process the predictions with a fully-connected CRF implemented as a CNN layer [28, 61].
Table 6.1: Material segmentation scores for same-dataset experiments (each method trained and tested only on the given dataset). For MINC models, the table shows the score of the best scoring single model without ensembles. Note that the OpenSurfaces [7] dataset is a subset of MINC, and the FMD is a subset of our local materials database.

<table>
<thead>
<tr>
<th>Method</th>
<th>MINC AVG</th>
<th>MINC MCA</th>
<th>Local Material Database AVG</th>
<th>Local Material Database MCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINC</td>
<td>78.8%</td>
<td>70.4%</td>
<td>71.1%</td>
<td>69.2%</td>
</tr>
<tr>
<td>Ours</td>
<td>79.1%</td>
<td>68.7%</td>
<td>73.8%</td>
<td>69.7%</td>
</tr>
</tbody>
</table>

**AVG**: per-pixel average accuracy  
**MCA**: mean class accuracy

### 6.3 Experiments and Comparisons

For all results in this section, we are evaluating material segmentation performance, not patch classification performance. We use two metrics: per-pixel average accuracy (AVG) and mean class accuracy (MCA). Per-pixel average accuracy measures overall recognition accuracy, while mean class accuracy aggregates accuracy scores by class before averaging.

In all experiments, we use the output of the MIT Places CNN [62] (with standard 10-crop prediction) as our place context probability vector, and the output of the DilatedNet model from ADE20K [65] as the object context probability map. Both models require the full image as input, and we store the resulting probability vectors and maps prior to training. During training, we extract small local patches from images and object probability maps at the same locations.

#### 6.3.1 Material Segmentation Comparisons

Table 6.1 shows a number of accuracy metric scores comparing our method with the previous state-of-the-art MINC material segmentation model. All scores are same-dataset scores, i.e. each method was trained and tested exclusively on the given dataset. For MINC, we test on the 1789 released segmented test images.

Our model achieves state-of-the-art material segmentation accuracy on multiple datasets.
More importantly, we observe a significantly larger gap in accuracy between our model and others given less training data. We sample approximately 250,000 local patches from our local materials database (Chapter 5), an order of magnitude smaller than the size of the MINC training set. This clearly shows that large-patch-based methods require extremely large amounts of data to see all possible combinations of material and context within the patch. In contrast, even in the presence of limited training data, we are able to reliably recognize materials by properly separating and integrating the contributions of scene and object context.

It is important to note that we achieve the reported scores on the MINC dataset from partially unreliable data. The MINC training set only includes annotations for single clicks within each image, sufficient information to train a patch-based method. Our method, however, is designed from the ground up as a fully-convolutional method predicting a full-resolution category map directly from the network. As a result, we require label maps, not clicks, during training. To obtain approximate label maps from the available MINC data, we take the provided clicks and use them to initialize the unary potential map for a fully-connected CRF [28]. The map is initialized to the uniform distribution over $K = 23$ categories everywhere except clicks, where it is initialized to the 1-hot vector for the corresponding clicked category. The output of the CRF given this unary map and a pairwise bilateral kernel on the input image becomes our training label map. As a result, only some of our training data is known ground truth; the rest are imputed.

Table 6.2 shows cross-dataset scores, where a model is trained on one dataset and tested on another. Models are evaluated on the set of categories shared by both datasets. We test the models without retraining, in order to focus on generalization performance.

---

We investigated training a fully-convolutional network given only annotated center pixels but the resulting network performed poorly.
Table 6.2: Material segmentation scores for cross-dataset experiments (each model trained on one dataset and tested on another). Models are tested without retraining, on overlapping categories only, in order to highlight generalization performance.

<table>
<thead>
<tr>
<th>Train Model</th>
<th>Test Model</th>
<th>MINC</th>
<th>Local Materials</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINC</td>
<td>MINC</td>
<td>N/A</td>
<td>54.0%</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td><strong>57.9%</strong></td>
</tr>
<tr>
<td>Local Materials</td>
<td>MINC</td>
<td>61.3%</td>
<td>N/A</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td><strong>61.9%</strong></td>
<td></td>
</tr>
</tbody>
</table>

As Bell et al. [6] and others have noted, training on balanced data is important if we want a model that does not learn trivial solutions based on inevitable dataset biases. Doing so in the case of densely-annotated training data, however, is particularly challenging. In the case of simple single-click annotations, it is sufficient to sample a class uniformly, then sample an example from that class, again uniformly [6]. For dense annotations, selecting an optimally-balanced subset of training examples is a challenging combinatorial optimization problem: each sample contributes a positive integer number of pixels to possibly multiple categories. We obtain approximately-balanced training examples by assigning each patch a weight based on the average frequency at which each pixel’s category appears in the training data.

6.3.2 Ablation Studies

Training Data Quantity  Plots in Figure 6.6 show the per-pixel average accuracy of our method as we vary the amount of available training data, using the MINC dataset [6]. The difference in accuracy becomes more significant as the amount of available training data gets smaller, showing that we are clearly able to exploit scene context for accurate material predictions from less data.

For all subsequent experiments, we use our local materials database introduced in Chap...
Figure 6.6: Accuracy vs. training set size on the MINC database (1.0 ≈ 2.5 million patches). We can clearly see that by separating local material appearance from context, we are able to recognize materials more accurately from fewer examples.

Context Sources We quantitatively evaluate the importance of context for material segmentation by training our model with all subsets of possible context sources as well as no context at all. Results in Table 6.3 show that each form of context independently improves the material segmentation accuracy. Furthermore, the accuracy with both forms of context is higher than either alone. Consider the following example: given the fact that an object is a sink, the material for pixels within that object region is likely to be either metal or ceramic; likewise, if we do not know the object category but know that the image was taken in a bathroom, metal and ceramic materials are likely to be present. Both pieces of information improve our material segmentation accuracy. If, however, we know both that the object is a sink and the image was taken in a bathroom, it is then much more likely that the sink is ceramic. In this way, we gain more information from the combination of context sources
Table 6.3: Accuracy for varying sources of context. Object and place categories each contribute significantly to the overall accuracy, and the combination of the two is even more accurate. As an example of this, sinks (an object) are often metal or ceramic, and bathrooms (a place) often contain metals and ceramics. Bathroom sinks, however, are typically ceramic. Objects and places together can provide information that is not available given either alone.

<table>
<thead>
<tr>
<th>Context</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>63.5%</td>
</tr>
<tr>
<td>Objects</td>
<td>68.9%</td>
</tr>
<tr>
<td>Places</td>
<td>70.1%</td>
</tr>
<tr>
<td>Both</td>
<td>73.8%</td>
</tr>
</tbody>
</table>

Table 6.4: Accuracy for varying levels of context granularity. Fine-grained places may not appear in many images, but coarse-grained categories may offer little in the way of material recognition cues. We find that the finest category granularity offers the best material segmentation performance. In this case, the 205 place categories are both fine-grained and sufficiently well-distributed across training examples.

<table>
<thead>
<tr>
<th>Hierarchy Level</th>
<th>Accuracy</th>
<th>Entropy (training data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>63.8%</td>
<td>2.51</td>
</tr>
<tr>
<td>Mid</td>
<td>64.9%</td>
<td>2.40</td>
</tr>
<tr>
<td>Low</td>
<td>66.3%</td>
<td>2.27</td>
</tr>
<tr>
<td>All</td>
<td>70.1%</td>
<td>1.91</td>
</tr>
</tbody>
</table>

than we could otherwise obtain from the sources on their own.

Place Category Granularity  As part of their SUN database for scene and object recognition, Xiao et al. [56] define a hierarchy of place categories. This hierarchy raises the question of whether any particular context granularity is more or less useful for material recognition. On one hand, having an extremely fine set of place categories might mean that few training examples would appear from certain places. At the other extreme, the coarsest division of places could only provide very general cues as to which materials may be present.

To evaluate the importance of place granularity, we compute material segmentation accuracy scores using only place context at each level of the SUN places hierarchy. We adapt their hierarchy to the place categories recognized by the MIT places CNN and treat nodes...
Figure 6.7: These examples show that context helps disambiguate materials when local information is not sufficient. In the first set of insets, the water has a local appearance similar to asphalt. Global context suggests that this is unlikely. In the second set, we see that the airplane body is incorrectly recognized due to the lack of characteristic specular reflection that locally identifies metal. Again, context fixes this error. Sky is not a material and in this case has the local appearance of water, hence the prediction for those pixels in the second row.
within each level of the hierarchy as place categories. The highest level is the simple division of indoors vs. outdoors, mid-level categories deal with distinctions such as commercial and residential buildings, or mountains and forests, and the lowest level includes smaller groups such as entertainment or religious places. Results in Table 6.4 show that accuracy increases with place category granularity: more detailed place categories provide more discriminative information for material recognition.

**Object Context Spatial Resolution** Unlike place categories, which we represent with a single set of category probabilities for an entire image, object probability maps have distinct values at every pixel. It is possible that the fine detail present in the object category maps could produce more accurate material segmentations merely by providing boundary shape information rather than by constraining the space of possible materials. We show that this is in fact not the case by first training our model with full places and full object context resolution, then testing with reduced spatial resolution for the object probability maps. We reduce spatial resolution by blurring, downsampling, then upsampling the object context map. After running this experiment at downsampling factors $d \in \{2, 4, 8, 16\}$, results show that spatial resolution has essentially no effect on material segmentation accuracy$^3$. If we were in fact relying on the fine detail in the object map, we would see a decrease in accuracy which is not present.

**Context Introduction Level** Results in Table 6.5 show that the highest level is indeed the ideal place at which to introduce global context. If introduced at lower levels, the network is free to overfit to the context and sacrifices test accuracy as a result. For this experiment alone we do not use pre-trained weights to initialize the network. If we did, as in

\[\text{Accuracy was within } \pm 0.2\% \text{ of } 73.8\%, \text{ the full-resolution accuracy, with no correlation between downsample factor and accuracy difference.}\]
Table 6.5: Accuracy with context introduced at varying levels. We introduce context at each of the above layers and compute material segmentation accuracy. The accuracy increases as the context is introduced at higher layers in the network, showing that the best level for context introduction is in the upper layers of the network.

<table>
<thead>
<tr>
<th>CNN Layer</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>pool1</td>
<td>62.0%</td>
</tr>
<tr>
<td>pool2</td>
<td>62.8%</td>
</tr>
<tr>
<td>pool3</td>
<td>64.1%</td>
</tr>
<tr>
<td>conv4_3</td>
<td>66.0%</td>
</tr>
<tr>
<td>fr_conv1</td>
<td>66.4%</td>
</tr>
</tbody>
</table>

Figure 6.8: Sample failure cases. Insets compare the output of our model vs the model of [6] trained on MINC images. Cat fur does not match a material in the MINC database and is unlikely to appear in their database due to the image sources used. Our method predicts fabric instead, which has a similar local visual appearance. The stone statue image contains few contextual cues, but we are able to make reasonable predictions based on the local appearance.

Our other experiments, then the accuracy would be artificially reduced as the added context would invalidate the pre-trained weights above it.

6.3.3 Qualitative Examples

We can readily see in Figure 6.7 that the context helps disambiguate materials that may be difficult to recognize from only local information. When metal does not exhibit specular highlights or reflections, as is the case with the airplane body, the flat white surface offers little in the way of local recognition cues. Knowing that the current pixel belongs to a plane removes this ambiguity. Likewise, in the natural scene with elephants, the combination of
high-frequency waves and specular reflection causes the water to appear like concrete. Scene context makes it clear that concrete would be unlikely in this case. In general, the predictions are accurate subject to the limitations of the training data. Skin is not a material in our local materials database, and thus skin is often classified as the surrounding fabric. Sky is not a material and the predictions for sky are determined largely by context (ex. metal water over the ocean). Figure 6.8 shows a failure case of our method in the presence of an unknown material. Despite this, our method is able to use the local appearance to predict a visually-similar material rather than simply “other”. Figure 6.9 and Figure 6.10 contain further examples of dense material predictions from our framework.
Figure 6.10: Additional examples of the output of our dense per-pixel material recognition framework.
Chapter 7: Conclusion

We have investigated the close relationship between materials and the contextual cues which surround them. We first showed, with two distinct methods, that we may isolate material appearance in small local image patches. Following parallels between computer vision and neuroscience, we integrated our perceptual attribute discovery method with a local material recognition CNN to recognize materials and their perceptual attributes in large-scale image databases. To support this, we also introduced a new material recognition database with carefully-selected categories aimed at local material recognition. We then introduced external sources for scene and object context to this separated local material appearance to accurately recognize materials whether or not context is present, from fewer examples than existing methods.

7.1 Visual Material Traits

First we introduced visual material traits, named visual material properties like “shiny” or “fuzzy”, to enable the recognition of materials from small local image patches alone. In support of this, we also released a set of annotations for the FMD [49] containing binary masks for each trait. We also show that we may aggregate material traits within image regions to produce object-level material category predictions.

Limitations As a purely-local method, visual material traits cannot take advantage of the contextual cues that we have shown to be required for accurate material recognition. The traits are context-independent, and they maintain this property even when aggregated within a large object region which might otherwise contain significant context. Despite this,
the traits themselves are still useful on their own as a local visual descriptor.

7.2 Perceptual Material Attributes

To address scalability and consistency issues that arise in visual material traits, we derived a method that produces a set of classifiers for an unnamed set of visual material attributes. We probe our own human visual perception of materials to derive constraints that ensure these classifiers produce attributes which serve the same function as material traits given only easily-obtained weak supervision.

**Limitations**  As with material traits, perceptual material attributes are context-independent properties and thus suffer from the same drawbacks relating to the lack of contextual cues. While the annotation phase of our method is significantly more scalable than visual material traits, the discovery of the classifiers is still relatively slow and not applicable to modern large-scale image databases.

7.3 MAC-CNN and Local Materials Database

We proposed a single framework that integrates weakly-supervised attribute discovery with local material recognition. Our proposed CNN architecture allows us to discover perceptual material attributes within a local material recognition network. To evaluate the framework, and to address issues present in existing material recognition databases, we also built a new material image database from carefully-chosen material categories. The accuracy of unseen category recognition based solely on our discovered attributes and few sample images shows that the attributes form a compact representation for novel materials. We find the parallels between our own human visual perception of materials and the material attributes discovered in the MAC-CNN architecture particularly interesting.
Limitations  This method is again a local material recognition method and subject to the appropriate limitations.

7.4 Integrating Materials and Context

We demonstrated a novel method for separation and integration of local material appearance and global context for accurate material segmentation. Our experimental results show that by separating local material appearance and global context, we are able to take advantage of the strong cues present within the scene context to accurately recognize materials at the per-pixel level using significantly less training data than existing methods which rely on context implicitly being present in large input patches. We also investigate interesting properties of the contextual information used by our method, including the ideal hierarchical level at which we should introduce context, and how much object and place categories individually contribute to the overall recognition accuracy.

Limitations  While we show improved material segmentation accuracy with the addition of external context, to do so we require a source of said context. This is, however, readily obtained from existing place and object recognition frameworks and requires no re-training of the context networks. Since our method is end-to-end fully-convolutional and outputs full-resolution material maps by design, we also require full-resolution training data. Despite this, the training data need not be densely-annotated; simple polygonal regions are sufficient.

7.5 Future Work

The limitations discussed above provide promising avenues for future work. At a somewhat straightforward level, it would be interesting to investigate the combination of perceptual material attributes with our material segmentation framework. The same contextual cues
we use to aid material segmentation might result in the discovery of more discriminative perceptual material attributes. The material hierarchy we introduced in Chapter 5 also contains fine-grained categories for which we did not collect annotations due to practical concerns. Given sufficient manpower and time, however, annotations for the fine-grained categories we proposed could provide even more detailed material information for autonomous systems. In Chapter 6, we showed that a simple CRF-based method works well to infer dense material annotations from sparse clicks, so long as there are sufficient clicks in the image. Given the larger sizes of single-click vs. dense- or region-based material databases, we consider the adaptation of single-click annotations for use with dense methods to be a promising avenue for future study.
Bibliography


Vita

Gabriel Schwartz received his Bachelor of Science in Computer Science and Master of Science in Computer Science from Drexel University in 2011.

Selected Publications

• Material Recognition from Local Appearance in Global Context

• Discovering Perceptual Attributes in a Deep Local Material Recognition Network
  Gabriel Schwartz and Ko Nishino, Biological and Artificial Vision (Workshop held in conjunction with ECCV), 2016

• Automatically Discovering Local Visual Material Attributes
  Gabriel Schwartz and Ko Nishino, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015

• Visual Material Traits: Recognizing Per-Pixel Material Context
  Gabriel Schwartz and Ko Nishino, Color and Photometry in Computer Vision (Workshop held in conjunction with ICCV), 2013

• 3D Geometric Scale Variability in Range Images: Features and Descriptors