Material Segmentation from Local Appearance and Global Context

Gabriel Schwartz Ko Nishino
Department of Computer Science, Drexel University
{gbs25,kon}@drexel.edu

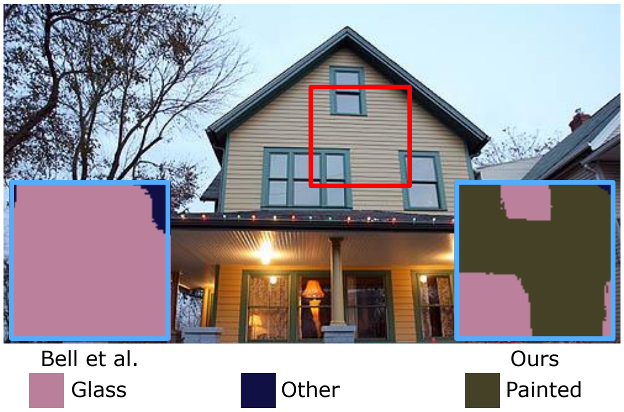

Figure 1. Material segmentation methods derived from object recognition can rely too-heavily on contextual cues to classify materials. When these cues are ambiguous, this leads to errors that can be resolved using the local appearance information. Since existing methods do not cleanly separate local appearance and context, they cannot resolve such ambiguities.

1. Introduction

Material segmentation (recognizing material categories at every pixel) provides valuable information for scene understanding and autonomous interaction. If a robot is tasked to, "Pick up the plastic cup on the table," an object detection framework can provide the location of the table and the cups on it, but we must also provide precise material category labels to identify the desired object. If we are to fully-describe an image we need to know not just what objects are present but also what the objects are made of.

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 [1], materials are a kind of "stuff", fundamentally different than objects ("things").

As seen in Figure 1, when object-focused semantic segmentation methods are applied to materials, they fail because they rely too heavily on properties of the objects involved. At the same 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. Methods that cannot take advantage of scene context at all (such as that of Schwartz and Nishino [9]) will fail to accurately recognize materials when that context is the only distinguishing factor.

In our work, we find that contextual cues like object and place categories provide 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. Existing methods like those of Bell *et al.* [2] or Cimpoi *et al.* [4], however, can only see a portion of these contextual cues inside the large patches and regions on which they are trained.

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.

2. Related Work

Material recognition has been typically performed at the whole-image or image patch level [2, 4, 10, 12]. Such methods have only implicit access to the contextual cues required to recognize materials, and thus must learn to infer these cues (essentially re-learning object recognition) before recognizing materials. This then implies the requirement of an extremely large training dataset to span the product space of materials and objects.

Dense prediction has been extensively studied in the context of object semantic segmentation. Object recognition datasets, such as ImageNet [8] or MS COCO [7], often contain many (80-1,000) categories. Despite this, semantic segmentation methods such as DeepLab [3] focus on only a few coarse-grained categories. An important exception is the recent ADE20k dataset, scene parsing challenge, and associated models [13]. They define a set of 150 categories

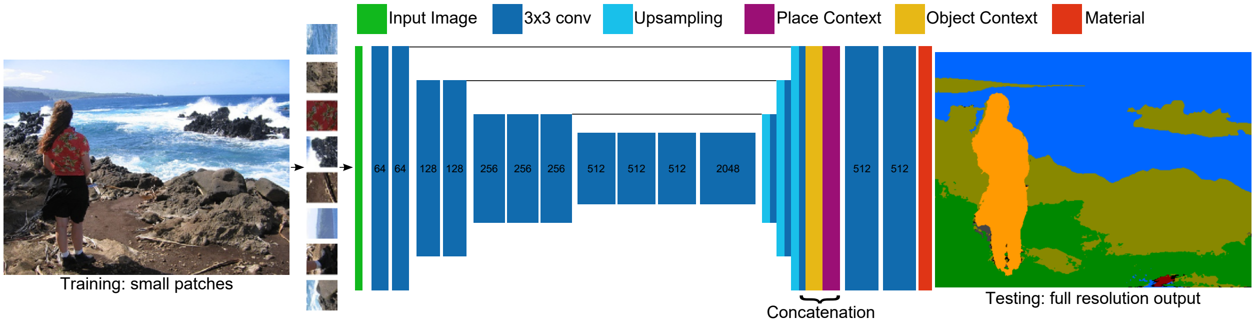

Figure 2. Material segmentation CNN architecture. Our network takes an input image, object category probability map, and a place category probability vector as inputs. Horizontal lines represent additive skip connections, with appropriate zero-padding on the channel axis. During training, the network only sees 48 × 48px image patches to ensure we are separating local material appearance from context. At test time, we may input an image of arbitrary size.

for semantic segmentation. We find the ADE20k models to be ideal sources for the per-pixel object category probabilities used by our method.

The use of context as a means to reduce ambiguity, whether in materials or other cases, appears promising. Hu *et al*. [5] showed that a simple addition of object category predictions as features could potentially improve material recognition. Iizuka *et al*. [6] use scene place category predictions to improve the accuracy of greyscale image colorization. Shrivastava and Gupta [11] investigate the use of semantic segmentation to augment Faster R-CNN.

3. Materials and Context

We propose a material segmentation CNN (Figure 2) 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. We obtain these probabilities from existing object and place recognition networks. 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 be present.

Figure 3 shows the per-pixel average accuracy of our method compared to that of Bell *et al*. [2] with varying amounts of training data. We are able to accurately recognize materials given limited amounts of labeled training data, and achieve state-of-the-art accuracy.

Figure 4 shows the effect of training patch size on accuracy. Larger patches implicitly contain more contextual cues to which the network may overfit.

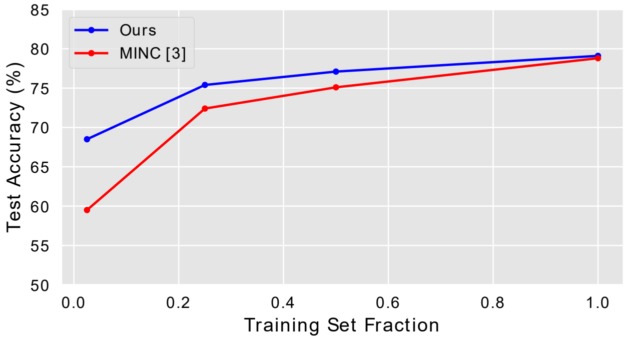

Figure 3. Accuracy vs. training set size on the MINC database (1.0 ≈ 2.5 million patches). We can clearly see that by separating local material appearance from context, we are able to recognize materials more accurately from fewer examples.

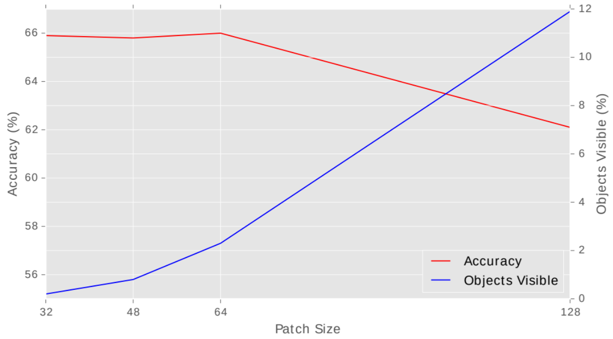

Figure 4. Accuracy vs. training patch size on a small subset of the MINC database. As the patch size increases, the model is free to overfit to the limited amount of available contextual cues.

4. Conclusion

We demonstrate a novel method for separation and integration of local material appearance and global context for accurate material segmentation. Our experimental results show that we are able to take advantage of the strong cues present within scene context to accurately segment materials using significantly less training data than existing methods.

Acknowledgments This work was supported by the Office of Naval Research grant N00014-16-1-2158 (N00014-14-1-0316) and N00014-17-1-2406, and the National Science Foundation award IIS-1421094. The Titan X used for part of this research was donated by the NVIDIA Corporation.

References

[1] E. H. Adelson. On Seeing Stuff: The Perception of Materials by Humans and Machines. In *SPIE*, pages 1–12, 2001.

[2] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material Recognition in the Wild with the Materials in Context Database. In *CVPR*, 2015.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv*, abs/1606.00915, 2016.

[4] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi. Deep Filter Banks for Texture Recognition, Description, and Segmentation. *IJCV*, 118(1):65–94, 2016.

[5] D. Hu, L. Bo, and X. Ren. Toward Robust Material Recognition for Everyday Objects. In *BMVC*, pages 48.1–48.11, 2011.

[6] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. In *SIGGRAPH*, volume 35, pages 110:1–110:11, 2016.

[7] T.-Y. Lin, M. Marie, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[9] G. Schwartz and K. Nishino. Discovering Perceptual Attributes in a Deep Local Material Recognition Network. *arXiv*, abs/1604.01345, 2016.

[10] L. Sharan, R. Rosenholtz, and E. Adelson. Material Perception: What Can You See in a Brief Glance? *Journal of Vision*, 9(8):784, 2009.

[11] A. Shrivastava and A. Gupta. Contextual Priming and Feedback for Faster R-CNN. In *ECCV*, 2016.

[12] H. Zhang and K. Xue, Jia ana Dana. Deep TEN: Texture Encoding Network. In *CVPR*, 2017.

[13] B. Z. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic Understanding of Scenes through ADE20K Dataset. *arXiv*, abs/1608.05442, 2016.