ABO: Dataset and Benchmarks for Real-World 3D Object Understanding

Jasmine Collins¹, Shubham Goel¹, Kenan Deng², Achleshwar Luthra³, Leon Xu¹,², Erhan Gundogdu², Xi Zhang², Tomas F. Yago Vicente², Thomas Dideriksen², Himanshu Arora², Matthieu Guillaumin², and Jitendra Malik¹

¹ UC Berkeley, ² Amazon, ³ BITS Pilani

Abstract

We introduce Amazon Berkeley Objects (ABO), a new large-scale dataset designed to help bridge the gap between real and virtual 3D worlds. ABO contains product catalog images, metadata, and artist-created 3D models with complex geometries and physically-based materials that correspond to real, household objects. We derive challenging benchmarks that exploit the unique properties of ABO and measure the current limits of the state-of-the-art on three open problems for real-world 3D object understanding: single-view 3D reconstruction, material estimation, and cross-domain multi-view object retrieval.

1. Introduction

Progress in 2D image recognition has been driven by large-scale datasets [15,26,37,43,56]. The ease of collecting 2D annotations (such as class labels or segmentation masks) has led to the large scale of these diverse, in-the-wild datasets, which in turn has enabled the development of 2D computer vision systems that work in the real world.

Theoretically, progress in 3D computer vision should follow from equally large-scale datasets of 3D objects. However, collecting large amounts of high-quality 3D annotations (such as voxels or meshes) for individual real-world objects poses a challenge. One way around the challenging problem of getting 3D annotations for real images is to focus only on synthetic, computer-aided design (CAD) models [10,35,70]. This has the advantage that the data is large in scale (as there are many 3D CAD models available for download online) but many of the models are low quality or untextured and do not exist in the real world. This has led to a variety of 3D reconstruction methods that work well on clear-background renderings of synthetic objects [13,24,46,65] but do not necessarily generalize to real images, new categories, or more complex object geometries [5,6,58].

To enable better real-world transfer, another class of 3D datasets aims to link existing 3D models with real-world images [63,64]. These datasets find the closest matching CAD models for the objects in an image and have human annotators align the pose of each model to best match the
image. While this has enabled the evaluation of 3D reconstruction methods in-the-wild, the shape (and thus pose) matches are approximate. Further, because this approach relies on matching CAD models to images, it inherits the limitations of the existing CAD model datasets (i.e., poor coverage of real-world objects, basic geometries and textures).

The IKEA [41] and Pix3D [57] datasets sought to improve upon this by annotating real images with exact, pixel-aligned 3D models. The exact nature of such datasets has allowed them to be used as training data for single-view reconstruction [21] and has bridged some of the synthetic-to-real domain gap [21]. However, the size of the datasets are relatively small (90 and 395 unique 3D models, respectively), likely due to the difficulty of finding images that exactly match 3D models. Further, the larger of the two datasets [57] only contains 9 categories of objects. The provided 3D models are also untextured, thus the annotations in these datasets are typically used for shape or pose-based tasks, rather than tasks such as material prediction.

Rather than trying to match images to synthetic 3D models, another approach to collecting 3D datasets is to start with real images (or video) and reconstruct the scene by classical reconstruction techniques such as structure from motion, multi-view stereo and texture mapping [12, 54, 55]. The benefit of these methods is that the reconstructed geometry faithfully represents an object of the real world. However, the collection process requires a great deal of manual effort and thus datasets of this nature tend to also be quite small (398, 125, and 1032 unique 3D models, respectively). The objects are also typically imaged in a controlled lab setting and do not have corresponding real images of the object “in context”. Further, included textured surfaces are assumed to be Lambertian and thus do not display realistic reflectance properties.

Motivated by the lack of large-scale datasets with realistic 3D objects from a diverse set of categories and corresponding real-world multi-view images, we introduce Amazon Berkeley Objects (ABO). This dataset is derived from Amazon.com product listings, and as a result, contains imagery and 3D models that correspond to modern, real-world, household items. Overall, ABO contains 147,702 product listings associated with 398,212 unique catalog images, and up to 18 unique metadata attributes (category, color, material, weight, dimensions, etc.) per product. ABO also includes “360° View” turntable-style images for 8,222 products and 7,953 products with corresponding artist-designed 3D meshes. In contrast to existing 3D computer vision datasets, the 3D models in ABO have complex geometries and high-resolution, physically-based materials that allow for photorealistic rendering. A sample of the kinds of real-world images associated with a 3D model from ABO can be found in Figure 1, and sample metadata attributes are shown in Figure 3. The dataset is released under CC BY-NC 4.0 license and can be downloaded at https://amazon-berkeley-objects.s3.amazonaws.com/index.html.

To facilitate future research, we benchmark the performance of various methods on three computer vision tasks that can benefit from more realistic 3D datasets: (i) single-view shape reconstruction, where we measure the domain gap for networks trained on synthetic objects, (ii) material estimation, where we introduce a baseline for spatially-varying BRDF from single- and multi-view images of complex real world objects, and (iii) image-based multi-view object retrieval, where we leverage the 3D nature of ABO to evaluate the robustness of deep metric learning algorithms to object viewpoint and scenes.

2. Related Work

3D Object Datasets ShapeNet [10] is a large-scale database of synthetic 3D CAD models commonly used for training single- and multi-view reconstruction models. IKEA Objects [42] and Pix3D [57] are image collections with 2D-3D alignment between CAD models and real images, however these images are limited to objects for which there is an exact CAD model match. Similarly, Pascal3D+ [64] and ObjectNet3D [63] provide 2D-3D alignment for images and provide more instances and categories, however
models are procedurally generated shapes that do not correspond to real objects. In contrast, ABO provides shapes and SV-BRDFs created by professional artists for real-life objects that can be directly used for photorealistic rendering.

Table 1 compares the 3D subset of ABO with other commonly used 3D datasets in terms of size (number of objects and classes) and properties such as the presence of real images, full 3D meshes and physically-based rendering (PBR) materials. ABO is the only dataset that contains all of these properties and is much more diverse in number of categories than existing 3D datasets.

3D Shape Reconstruction Recent methods for single-view 3D reconstruction differ mainly in the type of supervision and 3D representation used, whether it be voxels, point clouds, meshes, or implicit functions. Methods that require full shape supervision in the single-view [18, 22, 46, 57, 69] and multi-view [13, 31, 65] case are often trained using ShapeNet. There are other approaches that use more natural forms of multi-view supervision such as images, depth maps, and silhouettes [31, 59, 62, 66], with known cameras. Of course, multi-view 3D reconstruction has long been studied with classical computer vision techniques [27] like multi-view stereo and visual hull reconstruction. Learning-based methods are typically trained in a category-specific way and evaluated on new instances from the same category. Out of the works mentioned, only [69] claims to be category-agnostic. In this work we are interested in how well these ShapeNet-trained networks [13, 22, 46, 69] generalize to more realistic objects.

Material Estimation Several works have focused on modeling appearance from a single image, however realistic datasets available for this task are relatively scarce and small in size. [38] use two networks to estimate a homogeneous BRDF and an SV-BRDF of a flat surface from a single image, using a self-augmentation scheme to alleviate the need for a large training set. However, their work is limited to a specific family of materials, and each separate material requires another trained network. [67] extend the idea of self-augmentation to train with unlabeled data, but their work is limited by the same constraints. [16] use a modified U-Net and rendering loss to predict the SV-BRDFs of flash-lit photographs consisting of only a flat sur-
face. To enable prediction for arbitrary shapes, [40] propose a cascaded CNN architecture with a single encoder and separate decoder for each SV-BRDF parameter. While the method achieves good results on semi-uncontrolled lighting environments, it requires using the intermediate bounces of global illumination rendering as supervision. More recent works have turned towards using multiple images to improve SV-BRDF estimation, but still only with simplistic object geometries. For instance, [17] and [20] use multiple input images with a flash lit light source, but only for a single planar surface. [7] and [8] both use procedurally generated shapes to estimate SV-BRDFs from multi-view images. ABO addresses the lack of sufficient realistic data for material estimation, and in this work we propose a simple baseline method that can estimate materials from single or multi-view images of complex, real-world shapes.

2D/3D Image Retrieval Learning to represent 3D shapes and natural images of products in a single embedding space has been tackled by [39]. They consider various relevant tasks, including cross-view image retrieval, shape-based image retrieval and image-based shape retrieval, but all are inherently constrained by the limitations of ShapeNet [10] (cross-view image retrieval is only considered for chairs and cars). [36] introduced 3D object representations for fine-grained recognition and a dataset of cars with real-world 2D imagery (CARS-196), which is now widely used for deep metric learning (DML) evaluation. Likewise, other datasets for DML focus on instances/fine categories of few object types, such as birds [60], clothes [44], or a few object categories [50].

Due to the limited diversity and the similar nature of query and target images in existing retrieval benchmarks, the performance of state-of-the-art DML algorithms are near saturation. Moreover, since these datasets come with little structure, the opportunities to analyze failure cases and improve algorithms are limited. Motivated by this, we derive a challenging large-scale benchmark dataset from ABO with hundreds of diverse categories and a proper validation set. We also leverage the 3D nature of ABO to measure and improve the robustness of representations with respect to changes in viewpoint and scene. A comparison of ABO and existing benchmarks for DML can be found in Table 2.

3. The ABO Dataset

Dataset Properties The ABO dataset originates from worldwide product listings, metadata, images and 3D models provided by Amazon.com. This data consists of 147,702 listings of products from 576 product types sold by various Amazon-owned stores and websites (e.g. Amazon, PrimeNow, Whole Foods). Each listing is identified by an item ID and is provided with structured metadata corresponding to information that is publicly available on the listing’s main webpage (such as product type, material, color, and dimensions) as well as the media available for that product. This includes 398, 212 high-resolution catalog images, and, when available, the turntable images that are used for the “360º View” feature that shows the product imaged at 5º or 15º azimuth intervals (8, 222 products).

3D Models ABO also includes 7,953 artist-created high-quality 3D models in glTF 2.0 format. The 3D models are oriented in a canonical coordinate system where the “front” (when well defined) of all objects are aligned and each have a scale corresponding to real world units. To enable these meshes to easily be used for comparison with existing methods trained on 3D datasets such as ShapeNet, we have collected category annotations for each 3D model and mapped them to noun synsets under the WordNet [47] taxonomy. Figure 4 shows a histogram of the 3D model categories.

Catalog Image Pose Annotations We additionally provide 6-DOF pose annotations for 6, 334 of the catalog images. To achieve this, we develop an automated pipeline for pose
estimation based on the knowledge of the 3D model in the image, off-the-shelf instance masks [28, 34], and differentiable rendering. For each mask $M$, we estimate $R \in SO(3)$ and $T \in \mathbb{R}^3$ such that the following silhouette loss is minimized

$$R^*, T^* = \arg\min_R \|DR(R, T) - M\|$$

where $DR(\cdot)$ is a differentiable renderer implemented in PyTorch3D [52]. Examples of results from this approach can be found in Figure 2. Unlike previous approaches to CAD-to-image alignment [57, 63] that use human annotators in-the-loop to provide pose or correspondences, our approach is fully automatic except for a final human verification step.

**Material Estimation Dataset** To perform material estimation from images, we use the Disney [9] base color, metallic, roughness parameterization given in glTF 2.0 specification [25]. We render 512x512 images from 91 camera positions along an upper icosphere of the object with a 60° field-of-view using Blender’s [14] Cycles path-tracer. To ensure diverse realistic lighting conditions and backgrounds, we illuminate the scene using 3 random environment maps out of 108 indoor HDRIs [23]. For these rendered images, we generate the corresponding ground truth base color, material, roughness, and normal maps along with the object depth map and segmentation mask. The resulting dataset consists of 2.1 million rendered images and corresponding camera intrinsics and extrinsics.

**4. Experiments**

**4.1. Evaluating Single-View 3D Reconstruction**

As existing methods are largely trained in a fully supervised manner using ShapeNet [10], we are interested in how well they will transfer to more real-world objects. To measure how well these models transfer to real object instances, we evaluate the performance of a variety of these methods on objects from ABO. Specifically we evaluate 3D-R2N2 [13], GenRe [69], Occupancy Networks [46], and Mesh R-CNN [22] pre-trained on ShapeNet. We selected these methods because they capture some of the top-performing single-view 3D reconstruction methods from the past few years and are varied in the type of 3D representation that they use (voxels in [13], spherical maps in [69], implicit functions in [46], and meshes in [22]) and the coordinate system used (canonical vs. view-space). While all the models we consider are pre-trained on ShapeNet, GenRe trains on a different set of classes and takes as input a silhouette mask at train and test time.

To study this question (irrespective of the question of cross-category generalization), we consider only the subset of ABO models objects that fall into ShapeNet training categories. Out of the 63 categories in ABO with 3D models, we consider 6 classes that intersect with commonly used ShapeNet classes, capturing 4,170 of the 7,953 3D models. Some common ShapeNet classes, such as “airplane”, have no matching ABO category; similarly, some categories in ABO like “air conditioner” and “weights” do not map well to ShapeNet classes.

For this experiment, we render a dataset (distinct from the ABO Material Estimation Dataset) of objects on a blank background from a similar distribution of viewpoints as in the rendered ShapeNet training set. We render 30 viewpoints of each mesh using Blender [14], each with a 40° field-of-view and such that the entire object is visible. Camera azimuth and elevation are sampled uniformly on the surface of a unit sphere with a $-10^\circ$ lower limit on elevations to avoid uncommon bottom views.

GenRe and Mesh-RCNN make their predictions in “view-space” (i.e. pose aligned to the image view), whereas R2N2 and Occupancy Networks perform predictions in canonical space (predictions are made in the same category-specific, canonical pose despite the pose of the object in an image). For each method we evaluate Chamfer Distance and Absolute Normal Consistency and largely follow the evaluation protocol of [22].

**Results** A quantitative comparison of the four methods

<table>
<thead>
<tr>
<th></th>
<th>Chamfer Distance (↓)</th>
<th>Absolute Normal Consistency (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bench</td>
<td>chair</td>
</tr>
<tr>
<td>3D R2N2 [13]</td>
<td>2.46/0.85</td>
<td>1.46/0.77</td>
</tr>
<tr>
<td>Occ Nets [46]</td>
<td>1.72/0.51</td>
<td>0.72/0.39</td>
</tr>
<tr>
<td>GenRe [69]</td>
<td>1.54/2.86</td>
<td>0.89/0.79</td>
</tr>
<tr>
<td>Mesh R-CNN [22]</td>
<td>1.05/0.09</td>
<td>0.78/0.13</td>
</tr>
</tbody>
</table>

Table 3. Single-view 3D reconstruction generalization from ShapeNet to ABO. Chamfer distance and absolute normal consistency of predictions made on ABO objects from common ShapeNet classes. We report the same metrics for ShapeNet objects (denoted in gray), following the same evaluation protocol. All methods, with the exception of GenRe, are trained on all of the ShapeNet categories listed.
we considered on ABO objects can be found in Table 3. We also re-evaluated each method’s predictions on the ShapeNet test set from R2N2 [13] with our evaluation protocol and report those metrics. We observe that Mesh R-CNN [22] outperforms all other methods across the board on both ABO and ShapeNet in terms of Chamfer Distance, whereas Occupancy Networks performs the best in terms of Absolute Normal Consistency. As can be seen, there is a large performance gap between all ShapeNet and ABO predictions. This suggests that shapes and textures from ABO, while derived from the same categories but from the real world, are out of distribution and more challenging for the models trained on ShapeNet. Further, we notice that the lamp category has a particularly large performance drop from ShapeNet to ABO. Qualitative results suggest that this is likely due to the difficulty in reconstructing thin structures. We highlight some qualitative results in Figure 5, including one particularly challenging lamp instance.

### 4.2. Material Prediction

To date, there are not many available datasets tailored to the material prediction task. Most publicly available datasets with large collections of 3D objects [10, 12, 19] do not contain physically-accurate reflectance parameters that can be used for physically-based rendering to generate photorealistic images. Datasets like PhotoShape [51] do contain such parameters but are limited to a single category. In contrast, the realistic 3D models in ABO are artist-created and have highly varied shapes and SV-BRDFs. We leverage this unique property to derive a benchmark for material prediction with large amounts of photorealistic synthetic data. We also present a simple baseline approach for both single- and multi-view material estimation of complex geometries.

**Method**

To evaluate single-view and multi-view material prediction and establish a baseline approach, we use a U-Net-based model with a ResNet-34 backbone to estimate SV-BRDFs from a single viewpoint. The U-Net has a common encoder that takes an RGB image as input and has a multi-head decoder to output each component of the SV-BRDF separately. Inspired by recent networks in [7, 17], we align images from multiple viewpoints by projection using depth maps, and bundle the original image and projected image pairs as input data to enable an analogous approach for the multi-view network. We reuse the single-view architecture for the multi-view network and use global max pooling to handle an arbitrary number of input images. Similar to [16], we utilize a differentiable rendering layer to render the flash illuminated ground truth and compare it to similarly rendered images from our predictions to better regularize the network and guide the training process. Ground truth material maps are used for direct supervision.

Our model takes as input 256x256 rendered images. For training, we randomly subsample 40 views on the icosphere for each object. In the case of the multi-view network, for each reference view we select its immediate 4 adjacent views as neighboring views. We use mean squared error as the loss function for base color, roughness, metalness, surface normal and render losses. Each network is trained for 17 epochs using the AdamW optimizer [45] with a learning rate of 1e-3 and weight decay of 1e-4.

**Results**

Results for the single-view network (SV-net) and multi-view network (MV-net) can be found in Table 4. The multi-view network has better performance compared to single-view network in terms of the base color, roughness, metalness, and surface normal prediction tasks. The multi-view network is especially better at predicting properties that affect view-dependent specular components like roughness and metalness.

We also run an ablation study on our multi-view network without using 3D structure to align neighboring views to reference view (denoted as MV-net: no projection). First, we observe that even without 3D structure-based alignment, the network still outperforms the single-view network on roughness and metallic predictions. Comparing to the multi-view network, which uses 3D structure-based alignment, we can see structure information leads to better performance for all parameters. We show some qualitative results from the test set in Figure 6.

As a focus of ABO is enabling real-world transfer, we also test our multi-view network on catalog images of objects from the test set using the pose annotations gathered by the methodology in Section 3, and use the inferred material parameters to relight the object (Figure 7). Despite the domain gap in lighting and background, and shift from synthetic to real, our network trained on rendered images makes reasonable predictions on the real catalog images. In one case (last row), the network fails to accurately infer the true base color, likely due to the presence of self-shadow.

### 4.3. Multi-View Cross-Domain Object Retrieval

Merging the available catalog images and 3D models in ABO, we derive a novel benchmark for object retrieval.
Figure 6. **Qualitative material estimation results for single-view (SV-net) and multi-view (MV-net) networks.** We show estimated SV-BRDF properties (base color, roughness, metallicness, surface normals) for each input view of an object compared to the ground truth.

Figure 7. **Qualitative multi-view material estimation results on real catalog images.** Each of the multiple views is aligned to the reference view using the catalog image pose annotations.

with the unique ability to measure the robustness of algorithms with respect to viewpoint changes. Specifically, we leverage the renderings described in Section 3, with known azimuth and elevation, to provide more diverse views and scenes for training deep metric learning (DML) algorithms. We also use these renderings to evaluate the retrieval performance with respect to a large gallery of catalog images from ABO. This new benchmark is very challenging because the rendered images have complex and cluttered indoor backgrounds (compared to the cleaner catalog images) and display products with viewpoints that are not typically present in the catalog images. These two sources of images are indeed two separate image domains, making the test scenario a multi-view cross-domain retrieval task.

**Method** To compare the performance of state-of-the-art DML methods on our multi-view cross-domain retrieval benchmark, we use PyTorch Metric Learning [2] implementations that cover the main approaches to DML: NormSoftmax [68] (classification-based), ProxyNCA [48] (proxy-based) and Contrastive, TripletMargin, NTXent [11] and Multi-similarity [61] (tuple-based). We leveraged the Powerful Benchmarker framework [1] to run fair and controlled comparisons as in [49], including Bayesian hyperparameter optimization.

We opted for a ResNet-50 [29] backbone, projected it to 128D after a LayerNorm [4] layer, did not freeze the BatchNorm parameters and added an image padding transformation to obtain undistorted square images before resizing to 256x256. We used batches of 256 samples with 4 samples per class, except for NormSoftmax and ProxyNCA where we obtained better results with a batch size of 32 and 1 sample per class. After hyperparameter optimization, we trained all losses for 1000 epochs and chose the best epoch based on the validation Recall@1 metric, computing it only every other epoch.

Importantly, whereas catalog and rendered images in the training set are balanced (188K vs 111K), classes with and without renderings are not (4K vs. 45K). Balancing them in each batch proved necessary to obtain good performance: not only do we want to exploit the novel viewpoints and scenes provided by the renderings to improve the retrieval performance, but there are otherwise simply not sufficiently many negative pairs of rendered images being sampled.
multi-view networks for SV-BRDF material estimation of 3D reconstruction approaches, and that both view- and meshes despite sampling them from the same distribution canonical-space methods do not generalize well to ABO are a challenging test set for ShapeNet-trained meshes. In this work we introduced ABO, a new dataset to help bridge the gap between real and synthetic 3D worlds. We demonstrated that the set of real-world derived 3D models in ABO are a challenging test set for ShapeNet-trained 3D reconstruction approaches, and that both view- and canonical-space methods do not generalize well to ABO models despite sampling them from the same distribution of training classes. We also trained both single-view and multi-view networks for SV-BRDF material estimation of

<table>
<thead>
<tr>
<th></th>
<th>Rendered images</th>
<th>Catalog</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recall@k (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-trained</td>
<td>5.0</td>
<td>11.4</td>
</tr>
<tr>
<td>Constrastive</td>
<td>28.6</td>
<td>48.9</td>
</tr>
<tr>
<td>Multi-similarity</td>
<td>23.1</td>
<td>41.9</td>
</tr>
<tr>
<td>NormSoftmax</td>
<td><strong>30.0</strong></td>
<td><strong>50.2</strong></td>
</tr>
<tr>
<td>NTXent</td>
<td>23.9</td>
<td>42.6</td>
</tr>
<tr>
<td>ProxyNCA</td>
<td>29.4</td>
<td>50.0</td>
</tr>
<tr>
<td>TripletMargin</td>
<td>22.1</td>
<td>41.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 4$</th>
<th>$k = 8$</th>
<th>$k = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-trained</td>
<td>5.0</td>
<td>8.1</td>
<td>11.4</td>
<td>15.3</td>
<td>18.0</td>
</tr>
<tr>
<td>Constrastive</td>
<td>28.6</td>
<td>38.3</td>
<td>48.9</td>
<td>59.1</td>
<td><strong>39.7</strong></td>
</tr>
<tr>
<td>Multi-similarity</td>
<td>23.1</td>
<td>32.2</td>
<td>41.9</td>
<td>52.1</td>
<td>38.0</td>
</tr>
<tr>
<td>NormSoftmax</td>
<td><strong>30.0</strong></td>
<td><strong>40.3</strong></td>
<td><strong>50.2</strong></td>
<td>60.0</td>
<td>35.5</td>
</tr>
<tr>
<td>NTXent</td>
<td>23.9</td>
<td>33.0</td>
<td>42.6</td>
<td>52.0</td>
<td>37.5</td>
</tr>
<tr>
<td>ProxyNCA</td>
<td>29.4</td>
<td>39.5</td>
<td>50.0</td>
<td>60.1</td>
<td>35.6</td>
</tr>
<tr>
<td>TripletMargin</td>
<td>22.1</td>
<td>31.1</td>
<td>41.3</td>
<td>51.9</td>
<td>36.9</td>
</tr>
</tbody>
</table>

Table 5. Test performance of state-of-the-art deep metric learning methods on the ABO retrieval benchmark. Retrieving products from rendered images highlights performance gaps that are not as apparent when using catalog images.

**Results** As shown in Table 5, the ResNet-50 baseline trained on ImageNet largely fails at the task (Recall@1 of 5%). This confirms the challenging nature of our novel benchmark. DML is thus key to obtain significant improvements. In our experiments, NormSoftmax, ProxyNCA and Contrastive performed better (≈ 29%) than the Multi-similarity, NTXent or TripletMargin losses (≈ 23%), a gap which was not apparent in other datasets, and is not as large when using cleaner catalog images as queries. Moreover, it is worth noting that the overall performance on ABO is significantly lower than for existing common benchmarks (see Table 2). This confirms their likely saturation [49], the value in new and more challenging retrieval tasks, and the need for novel metric learning approaches to handle the large scale and unique properties of our new benchmark.

Further, the azimuth ($\theta$) and elevation ($\phi$) angles available for rendered test queries allow us to measure how performance degrades as these parameters diverge from typical product viewpoints in ABO’s catalog images. Figure 8 highlights two main regimes for both azimuth and elevation: azimuths beyond $|\theta| = 75^\circ$ and elevations above $\phi > 50^\circ$ are significantly more challenging to match, consistently for all approaches. Closing this gap is an interesting direction of future research on DML for multi-view object retrieval. For one, the current losses do not explicitly model the geometric information in training data.

**5. Conclusion**

In this work we introduced ABO, a new dataset to help bridge the gap between real and synthetic 3D worlds. We demonstrated that the set of real-world derived 3D models in ABO are a challenging test set for ShapeNet-trained 3D reconstruction approaches, and that both view- and canonical-space methods do not generalize well to ABO meshes despite sampling them from the same distribution of training classes. We also trained both single-view and multi-view networks for SV-BRDF material estimation of complex, real-world geometries - a task that is uniquely enabled by the nature of our 3D dataset. We found that incorporating multiple views leads to more accurate disentanglement of SV-BRDF properties. Finally, joining the larger set of products images with synthetic renders from ABO 3D models, we proposed a challenging multi-view retrieval task that alleviates some of the limitations in diversity and structure of existing datasets, which are close to performance saturation. The 3D models in ABO allowed us to exploit novel viewpoints and scenes during training and benchmark the performance of deep metric learning algorithms with respect to the azimuth and elevation of query images.

While not considered in this work, the large amounts of text annotations (product descriptions and keywords) and non-rigid products (apparel, home linens) enable a wide array of possible language and vision tasks, such as predicting styles, patterns, captions or keywords from product images. Furthermore, the 3D objects in ABO correspond to items that naturally occur in a home, and have associated object weight and dimensions. This can benefit robotics research and support simulations of manipulation and navigation.

**Acknowledgements** We thank Pietro Perona and Frederic Devernay. This work was funded in part by an NSF GRFP (#1752814) and the Amazon-BAIR Commons Program.
References


[58] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do


