What Do Single-view 3D Reconstruction Networks Learn?

Maxim Tatarchenko*[1], Stephan R. Richter*[2], René Ranftl[2], Zhuwen Li[2],
Vladlen Koltun[2], and Thomas Brox[1]

[1]University of Freiburg [2]Intel Labs

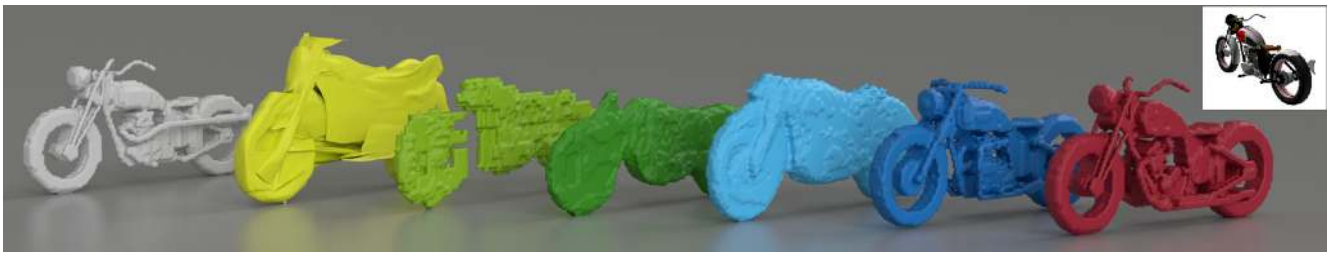

Figure 1. We provide evidence that state-of-the-art single-view 3D reconstruction methods (AtlasNet (light green, 0.38 IoU) [12], OGN (green, 0.46 IoU) [46], Matryoshka Networks (dark green, 0.47 IoU) [37]) do not actually perform reconstruction but image classification. We explicitly design pure recognition baselines (Clustering (light blue, 0.46 IoU) and Retrieval (dark blue, 0.57 IoU)) and show that they produce similar or better results both qualitatively and quantitatively. For reference, we show the ground truth (white) and a nearest neighbor from the training set (red, 0.76 IoU). The inset shows the input image.

Abstract

Convolutional networks for single-view object reconstruction have shown impressive performance and have become a popular subject of research. All existing techniques are united by the idea of having an encoder-decoder network that performs non-trivial reasoning about the 3D structure of the output space. In this work, we set up two alternative approaches that perform image classification and retrieval respectively. These simple baselines yield better results than state-of-the-art methods, both qualitatively and quantitatively. We show that encoder-decoder methods are statistically indistinguishable from these baselines, thus indicating that the current state of the art in single-view object reconstruction does not actually perform reconstruction but image classification. We identify aspects of popular experimental procedures that elicit this behavior and discuss ways to improve the current state of research.

1. Introduction

Object-based single-view 3D reconstruction calls for generating the 3D model of an object given a single image. Consider the motorcycle in Fig. 1. Inferring its 3D structure requires a complex process that combines low-level image cues, knowledge about structural arrangement of parts, and high-level semantic information. We refer to the extremes of this spectrum as *reconstruction* and *recognition*. *Reconstruction* implies reasoning about the 3D structure of the input image using cues such as texture, shading, and perspective effects. *Recognition* amounts to classifying the input image and retrieving the most suitable 3D model from a database, in our example finding a pre-existing 3D model of a motorcycle based on the input image.

While various architectures and 3D representations have been proposed in the literature, existing methods for single-view 3D understanding all use an encoder-decoder structure, where the encoder maps the input image to a latent representation and the decoder is supposed to perform non-trivial reasoning about the 3D structure of the output space. To solve the task, the overall network is expected to incorporate low-level as well as high-level information.

In this work, we analyze the results of state-of-the-art encoder-decoder methods [12, 37, 46] and find that they rely primarily on recognition to address the single-view 3D reconstruction task, while showing only limited reconstruction abilities. To support this claim, we design two pure recognition baselines: one that combines 3D shape clustering and image classification and one that performs image-based 3D shape retrieval. Based on these, we demonstrate

*Equal contribution.

that the performance of modern convolutional networks for single-view 3D reconstruction can be surpassed even without explicitly inferring the 3D structure of objects. In many cases the predictions of the recognition baselines are not only better quantitatively, but also appear visually more appealing, as demonstrated in Fig. 1.

We argue that the dominance of recognition in convolutional networks for single-view 3D reconstruction is a consequence of certain aspects of popular experimental procedures, including dataset composition and evaluation protocols. These allow the network to find a shortcut solution, which happens to be image recognition.

2. Related work

Historically, single-image 3D reconstruction has been approached via shape-from-shading [6, 16, 57]. More exotic cues for reconstruction are texture [28] and defocus [9]. These techniques only reason about visible parts of a surface using a single depth cue. More general approaches for depth estimation from a single monocular image use multiple cues as well as structural knowledge to infer an estimate of the depth of visible surfaces. Saxena *et al*. [40] estimated depth from a single image by training an MRF on local and global image features. Oswald *et al*. [34] solved the same problem with interactive user input. Hoiem *et al*. [15] used recognition together with simple geometric assumptions to construct 3D models from a single image. Karsch *et al*. [19] proposed a non-parametric framework that uses part- and object-level recognition to assemble an estimate from a database of images and corresponding depth maps. More recently, significant advances have been made in monocular depth estimation by employing convolutional networks [3, 7, 11, 26, 54].

This paper focuses on methods that not only reason about the 3D structure of object parts visible in the input image, but also hallucinate the invisible parts using priors learned from data. Tulsiani *et al*. [47] approached this task with deformable models for specific object categories. Most of the recent methods trained convolutional networks that map 2D images to 3D shapes using direct 3D supervision. A cluster of approaches used voxel-based representations of 3D shapes and generated them with 3D up-convolutions from a latent representation [4, 10, 53]. Several works [13, 38, 46] performed hierarchical partitioning of the output space to achieve computational and memory efficiency, which allows predicting higher-resolution 3D shapes. Johnston *et al*. [17] reconstructed high-resolution 3D shapes with an inverse discrete cosine transform decoder. Wang *et al*. [50] generated meshes by deforming a sphere into a desired shape, assuming a fixed distance between camera and objects. Groueix *et al*. [12] assembled surfaces from small patches. Multiple methods [27, 30, 43, 45] produced multi-view depth maps that are fused together into an output point cloud. Richter *et al*. [37] extended this with nested shapes fused into a single voxel grid. Fan *et al*. [8] directly regressed point clouds. Wu *et al*. [52] learned the mapping from input images to 2.5D sketches in a fully-supervised fashion, and then trained a network to map these intermediate representations to the final 3D shapes. Kong *et al*. [22] use 2D landmark locations together with silhouettes to retrieve and deform CAD models. Pontes *et al*. [35] improved upon this work by using a free-form deformation parametrization to model shape variation.

Tulsiani *et al*. [48] and Niu *et al*. [33] aimed for structural 3D understanding, approximating 3D shapes with a pre-defined set of primitives.

Recently, there has been a trend towards using weaker forms of supervision for single-view 3D shape prediction with convolutional networks. Multiple approaches [20, 36, 49, 55, 59] trained shape regressors by comparing projections of ground-truth and predicted shapes. Kanazawa *et al*. [18] predicted deformations from mean shapes trained from multiple learning signals.

There are only very few datasets available for the task of single-image 3D reconstruction – a consequence of the cost of data collection. Most existing methods use subsets of ShapeNet [1] for training and testing. Recently, Wiles and Zisserman [51] introduced two new synthetic datasets: Blobby objects and Sculptures. The Pix3D dataset [44] provides pairs of perfectly aligned natural images and CAD models. This dataset, however, contains a low number of 3D samples, which is problematic for training deep networks.

3. Reconstruction vs. recognition

Single-view 3D understanding is a complex task that requires interpreting visual data both geometrically and semantically. In fact, these two modes are not disjoint, but span a spectrum from pure geometric reconstruction to pure semantic recognition.

Reconstruction implies per-pixel reasoning about the 3D structure of the object shown in the input image, which can be achieved by using low-level image cues such as color, texture, shading, perspective, shadows, and defocus. This mode does not require semantic understanding of the image content.

Recognition is an extreme case of using semantic priors: it operates on the level of whole objects and amounts to classifying the object in the input image and retrieving a corresponding 3D shape from a database. While it provides a robust prior for reasoning about the invisible parts of objects, this kind of purely semantic solution is only valid if the new object can be explained by an object in the database.

As reconstruction and recognition represent opposing ends of a spectrum, resorting exclusively to either is un-

likely to produce the most accurate 3D shapes, since both ignore valuable information present in the input image. It is thus commonly hypothesized that a successful approach to single-view 3D reconstruction needs to combine low-level image cues, structural knowledge, and high-level object understanding [41].

In the following sections, we argue that current methods tackle the problem predominantly using recognition.

4. Conventional setup

In this section, we analyze current methods for single-view 3D reconstruction and their relation to reconstruction and recognition. We employ a standard setup for single-view 3D shape estimation. We use the ShapeNet dataset [1]. Unlike several recent approaches, which evaluated only on the 13 largest classes, we deliberately use all 55 classes, as was done in [56]. This allows us to investigate how the number of samples within a class influences shape estimation performance. Within each class, the shapes are randomly split into training, validation, and test sets, containing 70%, 10%, and 20% of the samples respectively. Every shape was rendered using the ShapeNet-Viewer from five uniformly sampled viewpoints $(\theta_{azimuth} \in [0°, 360°), \theta_{elevation} \in [0°, 50°))$. The distance to the camera was set such that each rendered shape roughly fits the frame. We rendered RGB images of size 224×224, which were downsampled to the input resolution that is required by each method.

All 3D shapes have a consistent canonical orientation and are represented as 128^3 voxel grids. Using high-resolution ground truth (compared to the conventionally used 32^3 voxel grids) is crucial for evaluating a method's ability to reconstruct fine detail. Evaluating on a higher resolution than 128^3 does not offer additional benefits, since the performance of state-of-the-art methods saturates at this level [37, 46], while training and evaluation become much more costly. We follow standard procedure and measure shape similarity with the mean Intersection over Union (mIoU) metric, aggregating predictions within semantic classes [4, 8, 13, 37, 42, 46, 55].

4.1. Existing approaches

We base our experiments on modern convolutional networks that predict high-resolution 3D models from a single image. A taxonomy of approaches arises by categorizing them based on their output representation: voxel grids, meshes, point clouds, and depth maps. To this end, we chose state-of-the-art methods that cover the dominant output representations or have clearly shown to outperform other related representations for our evaluation.

We use Octree Generating Networks (OGN) [46] as the representative method that predicts the output directly on a voxel grid. Compared to earlier works [4] that operate on this output representation, OGN allows predicting higher-resolution shapes by using octrees to represent the occupied space efficiently. We evaluate AtlasNet [12] as the representative approach for surface-based methods. AtlasNet predicts a collection of parametric surfaces and constitutes the state-of-the-art among methods that operate on this output representation. It was shown to outperform the only approach that directly produces point clouds as output [8], as well as another octree-based approach [13]. Finally, we evaluate the current state-of-the-art in the field, Matryoshka Networks [37]. Matryoshka Networks use a shape representation that is composed of multiple, nested depth maps, which are volumetrically fused into a single output object.

For IoU-based evaluation of the surface predictions from AtlasNet, we project them to depth maps, which we further fuse to a volumetric representation. In our experiments, this approach reliably closed holes in the reconstructed surfaces while retaining fine details. For surface-based evaluation metrics, we use the marching cubes algorithm [29] to extract meshes from volumetric representations.

4.2. Recognition baselines

We implemented two straightforward baselines that approach the problem purely in terms of recognition. The first is based on clustering of the training shapes in conjunction with an image classifier; the second performs database retrieval.

Clustering. In this baseline, we cluster the training shapes into K sub-categories using the K-means algorithm [31]. Since using 128^3 voxelizations as feature vectors for clustering is too costly, we run the algorithm on 32^3 voxelizations flattened into a vector. Once the cluster assignments are determined, we switch back to working with high-resolution models.

Within each of the K clusters, we calculate the mean shape as

$$\hat{m}_k = \frac{1}{N_k} \sum_{n=0}^{N_k} v_n, \tag{1}$$

where v_n is one of the N_k shapes belonging to the k-th cluster. We threshold the mean shapes at τ_k, where the optimal τ_k value is determined by maximizing the average IoU over the models belonging to the k-th cluster:

$$\tau_k = \arg\max_{\tau} \frac{1}{N_k} \sum_{n=0}^{N_k} \text{IoU}(\hat{m}_k > \tau, v_n), \tag{2}$$

where the thresholding operation is applied per voxel. We enumerate τ in the interval $[0.05, 0.5]$ with a step size of 0.05 to find the optimal threshold. We set $K = 500$.

Since correspondences between images and 3D shapes are known for the training set, images can be readily matched with the respective cluster k. Subsequently, we

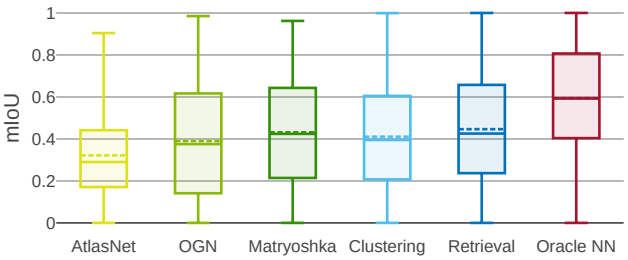

Figure 2. Comparison by mean IoU over the dataset. The box corresponds to the second and third quartile. The solid line in the box depicts the median; the dashed line the mean. Whiskers mark the minimum and maximum values, respectively.

train a 1-of-K classifier that assigns images to cluster labels. At test time, we set the mean shape of the predicted cluster as the inferred solution. For classification, we use the ResNet-50 architecture [14], pre-trained on the ImageNet dataset [5], and fine-tuned for 30 epochs on our data.

Retrieval. Our retrieval baseline is inspired by the work of Li *et al*. [25], which learns to embed images and shapes in a joint space. The embedding space is constructed from the pairwise similarity matrix of all 3D shapes in the training set by compressing each row of the matrix to a low-dimensional descriptor via Multi-Dimensional Scaling [24] with Sammon mapping [39]. To compute the similarity of two arbitrary shapes, Li *et al*. employ the lightfield descriptor [2]. To embed images in the space spanned by the shape descriptors, a convolutional network [23] is trained to map images to the descriptor given by the corresponding shape in the training set. During training, the network optimizes the Euclidean distance between predicted and ground-truth descriptors.

We adapt the work of Li *et al*. in several ways. As with our clustering baseline, we determine the similarity between two shapes via the IoU of their 32^3 voxel grid representation. We then compute a low-dimensional descriptor via principal component analysis. We further use a larger descriptor (512 vs. 128) and a network with larger capacity (ResNet-50 [14], pre-trained on ImageNet [5], without fixing any layers during fine-tuning). Finally, instead of minimizing the Euclidean distance, we maximize the cosine similarity between descriptors during training.

Oracle nearest neighbor. To gain more insight into the characteristics of the dataset, we evaluate an Oracle Nearest Neighbor (Oracle NN) baseline. For each of the test 3D shapes, we find the closest shape from the training set in terms of IoU. This method cannot be applied in practice, but gives an upper bound on how well a retrieval method can solve the task.

4.3. Analysis

We start by conducting a standard comparison of all methods in terms of their mean IoU scores. The results are summarized in Fig. 2. We find that state-of-the-art methods, despite being backed by different architectures, perform at a remarkably similar level. Interestingly, the retrieval baseline, a pure recognition method, outperforms all other approaches both in terms of mean and median IoU. The simple clustering baseline is competitive and outperforms both AtlasNet and OGN. We further observe that a perfect retrieval method (Oracle NN) performs significantly better than all other methods. Strikingly, the variance in the results is extremely high (between 35% and 50%) for all methods. This implies that quantitative comparisons that rely solely on the mean IoU do not provide a full picture at this level of performance. To shed more light on the behavior of the methods, we proceed with a more detailed analysis.

Per-class analysis. The similarity in average accuracy cannot be attributed to methods specializing in different subsets of classes. In Fig. 3 we observe consistent relative performance between methods across different classes. The retrieval baseline achieves the best results for 30 out of 55 classes. The classes are sorted from left to right in ascending order according to the performance of the retrieval baseline. The variance is high for all classes and all methods.

One might assume that the per-class performance depends on the number of training samples that are available for a class. However, we find no correlation between the number of samples in a class and its mean IoU score; see Fig. 4. The correlation coefficient between the two quantities is close to zero for all methods. This implies that there is no justification for only using 13 out of the 55 classes, as was done in many prior works [4, 8, 12, 37, 46, 55].

The quantitative results are backed by qualitative results shown in Fig. 5. For most classes, there is no significant visual difference between the predictions of the decoder-based methods and our clustering baseline. Clustering fails when the sample is far from the mean shape of the cluster, or when the cluster itself cannot be described well by the mean shape (this is often the case for chairs or tables because of thin structures that get averaged out in the mean shape). The predictions of the retrieval baseline look more appealing in most cases due to the presence of fine details, even though these details are not necessarily correct. We provide additional qualitative results in the supplementary material.

Statistical evaluation. To further investigate the hypothesis that convolutional networks bypass true reconstruction via image recognition, we visualize the histograms of IoU scores for individual object classes in Fig. 6. For histograms of all 55 classes we refer to the supplementary material. Although the distributions differ between classes, the within-

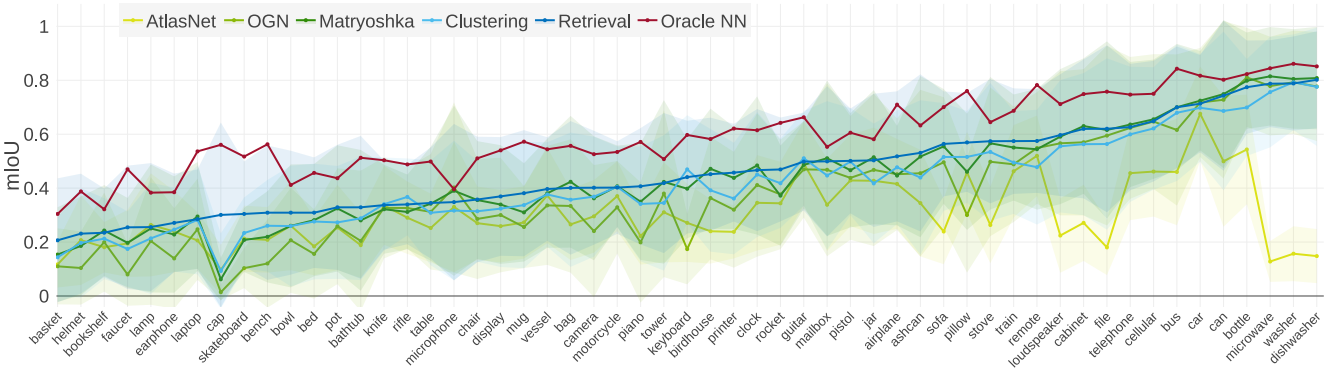

Figure 3. Comparison by mIoU per class. Overall, the methods exhibit consistent relative performance across different classes. The retrieval baseline produces the best reconstructions for the majority of classes. The variance is high for all classes and methods.

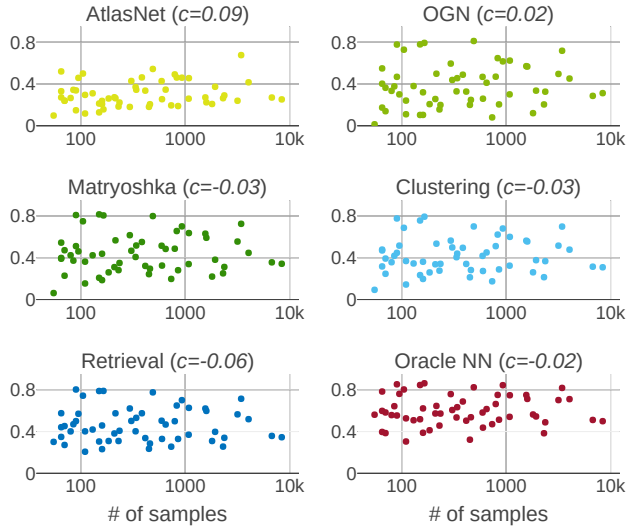

Figure 4. mIoU versus number of training samples per class. We find no correlation between the number of samples within a class and the mIoU score for this class. The correlation coefficient c is close to zero for all methods.

class distributions of decoder-based methods and recognition baselines are surprisingly similar.

For reference, we also plot the results of the Oracle NN baseline, which, for many classes, differs substantially. To verify this observation rigorously, we perform the Kolmogorov-Smirnov test [32] on the 50-binned versions of the histograms for all classes and all pairs of methods. The null hypothesis assumes that two distributions exhibit no statistically significant difference. We visualize the results of the test in the rightmost part of Fig. 6. Every cell of the heat map shows the number of classes for which the statistical test does not allow to reject the null hypothesis, *i.e.*, where the p-value is larger than 0.05. We find that for decoder-based methods and recognition baselines the null hypothesis cannot be rejected for the vast majority of classes.

5. Problems

In the preceding section we provided evidence that current methods for single-view 3D object reconstruction predominantly rely on recognition. Here we discuss aspects of popular experimental procedures that may need to be reconsidered to elicit more detailed reconstruction behavior from the models.

5.1. Choice of coordinate system

The vast majority of existing methods predict output shapes in an object-centered coordinate system, which aligns objects of the same semantic category to a common orientation. Aligning objects this way makes it particularly easy to find spatial regularities. It encourages learning-based approaches to recognize the object category first, and refine the shape later if at all.

Shin *et al*. [42] studied how the choice of coordinate frames affects reconstruction performance and generalization abilities of learning-based methods, comparing object-centered and viewer-centered coordinate frames. They found that a viewer-centered frame leads to significantly better generalization to object classes that were not seen during training, a result that can only be achieved when a method operates in a geometric reconstruction regime.

To validate these conclusions, we repeated the experimental evaluation (Sec. 4) in a viewer-centered coordinate frame. We attempted to extend the clustering baseline with a viewpoint prediction network which would regress the azimuth and elevation angles of the camera w.r.t. the canonical frame. This naive approach failed because the canonical frame has a different meaning for each object class, implying that the viewpoint network needs to use class information in order to solve the task. For the retrieval baseline, we retrained the method, treating each training view as a separate sample. To avoid artifacts from rotating voxelized shapes, we synthesized ground-truth shapes by rotating and then voxelizing the original meshes, resulting in

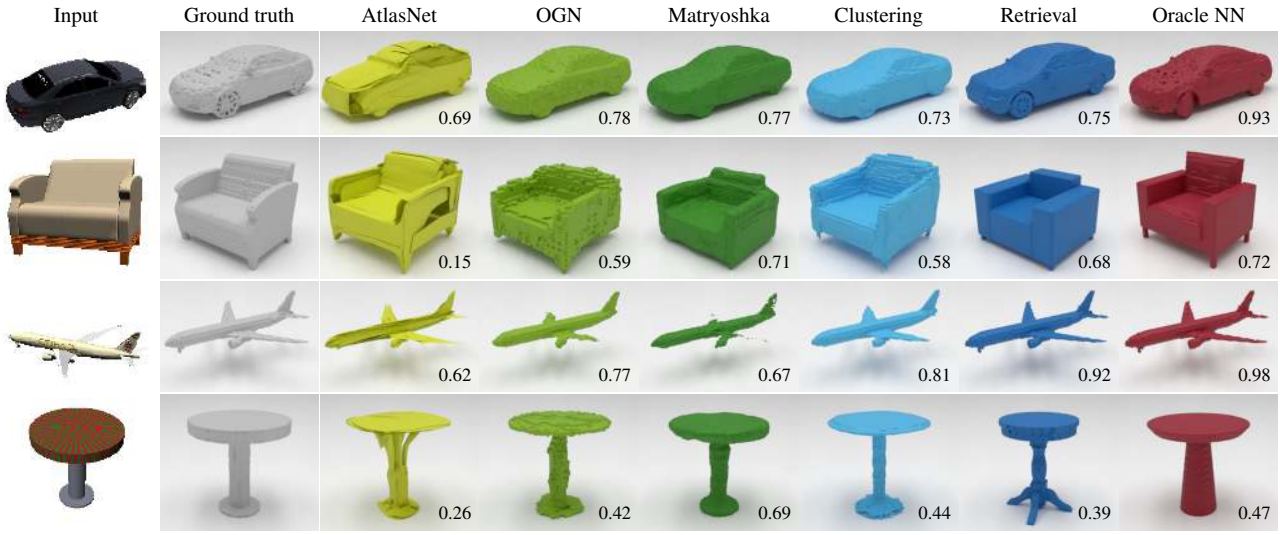

Figure 5. Qualitative results. Our clustering baseline produces shapes at a quality comparable to state-of-the-art approaches. Our retrieval baseline returns high-fidelity shapes by design, although details may not be correct. Numbers in the bottom right corner of each sample indicate the IoU.

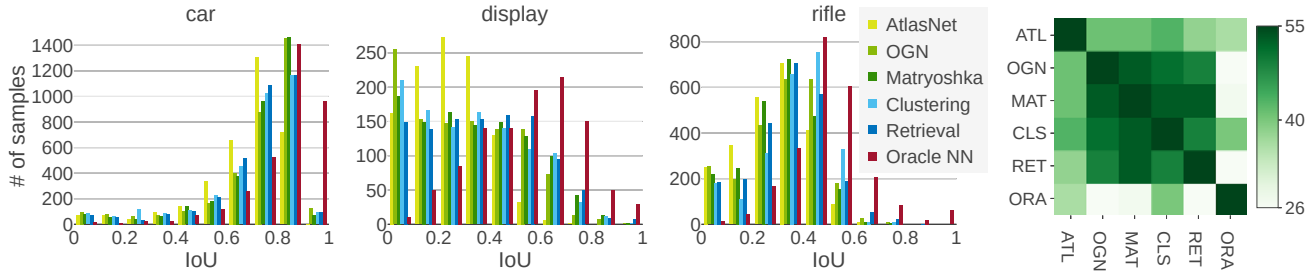

Figure 6. Left: Distribution of IoUs for selected classes. Within-class distributions for decoder-based methods and explicit recognition baselines are similar. The distributions of the Oracle NN differ for most of the classes. Right: A heat map of the number of classes for which the pairwise Kolmogorov-Smirnov test fails to reject the null hypotheses of the two distributions being the same.

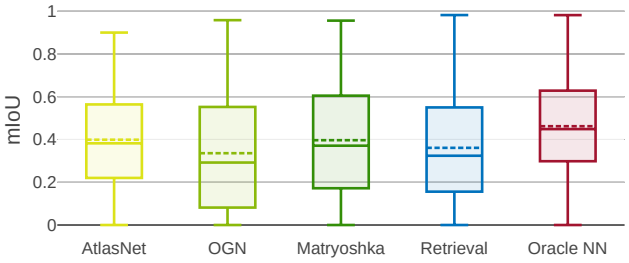

Figure 7. Mean IoU in viewer-centered mode. The retrieval baseline does not perform as well in this mode.

a distinct target shape for each view of each object. Results are shown in Fig. 7, where we observe a mild decrease in performance for OGN and Matryoshka networks, and a larger drop for the retrieval baseline. For the retrieval setting, the viewer-centered setup is computationally more demanding, as different views of the same object now refer to different shapes to be retrieved. Consequently, less learning capacity is available for each individual object.

5.2. Evaluation metric

Intersection over union. The mean IoU is commonly used as the primary quantitative measure for benchmarking single-view reconstruction approaches. This can be problematic if it is used as the sole metric to argue for the merits of an approach, since it is only indicative of the quality of a predicted shape if it reaches sufficiently high values. Low to mid-range scores indicate a significant discrepancy between two shapes.

An example is shown in Fig. 8, which compares a car model to different shapes in the dataset and illustrates their similarity in terms of IoU scores. As shown in the figure, even an IoU of 0.59 allows for considerable deviation from the ground-truth shape. For reference, note that 75% of the predictions by the best performing approach, our retrieval baseline, have an IoU below 0.66; 50% are below 0.43 (*c.f.* Fig. 2).

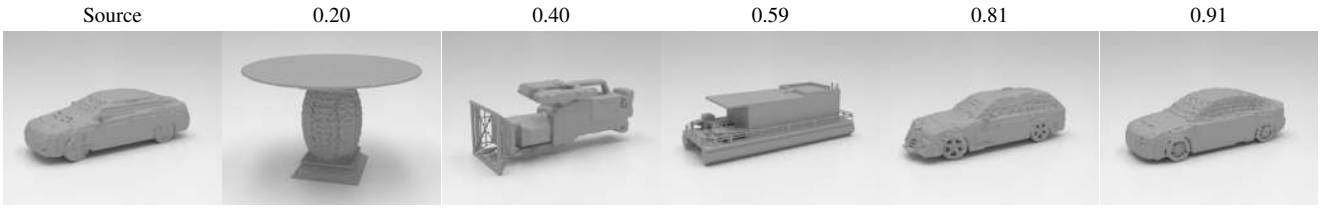

Figure 8. IoU between a source shape and various target shapes. Low to mid-range IoU values are a poor indicator of shape similarity.

Figure 9. The Chamfer distance is sensitive to outliers. Compared to the source, both target shapes exhibit non-matching parts that are equally wrong. While the F@1% is 0.56 for both shapes, the Chamfer distance differs significantly.

All information about an object's shape is situated on its surface. However, for voxel-based representations with a solid interior, the IoU is dominated by the interior parts of objects. As a consequence, even seemingly high IoU values may poorly reflect the actual surface similarity.

Moreover, while IoU can easily be evaluated for a volumetric representation, there is no straightforward way to evaluate it for point clouds. A good measure should allow comparing different 3D representations within the same unified framework. Point-based measures are most suitable for this, because a point cloud can be obtained from any other 3D representation via (a) surface point sampling for meshes, (b) per-pixel reprojection for depth maps, or (c) running the marching cubes algorithm followed by point sampling for voxel grids.

Chamfer distance. Some recent methods use the Chamfer Distance (CD) for evaluation [8, 12, 44]. Although it is defined on point clouds and by design satisfies the requirement of being applicable (after conversion) to different 3D representations, it is a problematic measure because of its sensitivity to outliers. Consider the example in Fig. 9. Both target chairs perfectly match the source chair in the lower part and are completely wrong in the upper part. However, according to the CD score, the second target is much better than the first. As this example shows, the CD measure can be significantly perturbed by the geometric layout of outliers. It is affected by how far the outliers are from the reference shape. We argue that in order to reliably reflect real reconstruction performance, a good quantitative measure should be robust to the detailed geometry of outliers.

F-score. Motivated by the insight that both IoU and CD can be misleading, we propose to use the F-score [21], an established and easily interpretable metric that is actively used in the multi-view 3D reconstruction community. The F-score explicitly evaluates the distance between object surfaces and is defined as the harmonic mean between precision and recall. Precision measures the accuracy of the reconstruction by counting the percentage of reconstructed points that lie within a certain distance to the ground truth. Recall measures the completeness of the reconstruction by counting the percentage of points on the ground truth that lie within a certain distance to the reconstruction. The strictness of the F-score can be controlled by varying the distance threshold d. The metric has an intuitive interpretation: the percentage of points (or surface area) that was reconstructed correctly.

We plot the F-score of viewer-centered reconstructions for different distance thresholds d in Fig. 10 (left). At $d = 2\%$ of the side length of the reconstructed volume, the absolute F-score values are in the same range as the current mIoU scores, which, as we argued before, is not indicative of the prediction quality. We therefore suggest evaluating the F-score at distance thresholds of 1% and below.

In Fig. 10 (right), we show the percentage of models with an F-score of 0.5 or higher at a threshold $d = 1\%$. Only a small number of shapes is reconstructed accurately, indicating that the task is still far from solved. Our retrieval baseline is no longer a clear winner, further showing that a reasonable solution in viewer-centered mode is harder to get using a pure recognition method.

We observe that AtlasNet often produces qualitatively good surfaces. It even outperforms the Oracle NN baseline on more liberal (above 2%) thresholds, as shown in Fig. 10 (left). Perceptually, humans tend to judge quality by global and semi-global features and tolerate if parts are slightly wrong in position or shape. We observe that AtlasNet, which was trained to optimize surface correspondence, rarely completely misses parts of the model, but tends to produce poorly localized parts. This is reflected in the high-performance range analysis, shown in Fig. 10 (right), where AtlasNet trails all other approaches.

Analyzing precision and recall separately provides additional insights into each method's behavior. In Fig. 11 we see that OGN and Matryoshka Networks outperform Oracle NN in terms of precision. However, both Oracle NN and

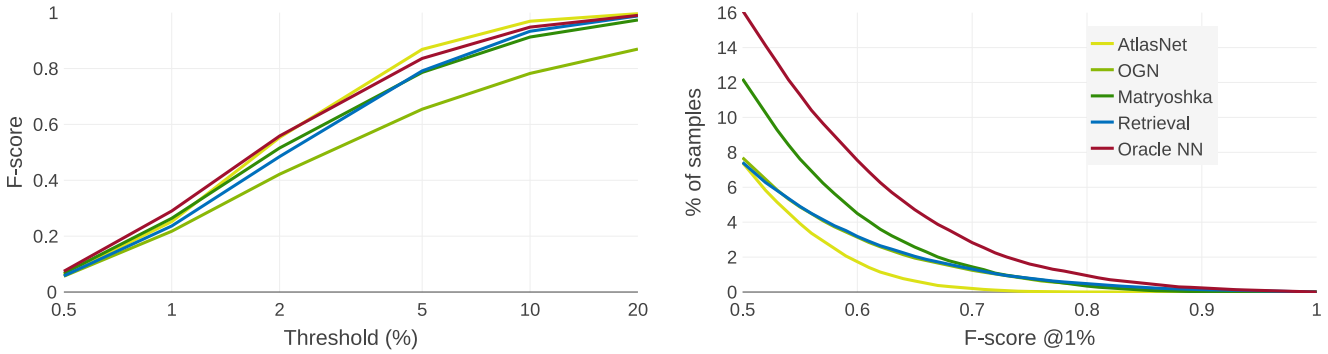

Figure 10. F-score statistics in viewer-centered mode. Left: F-score for varying distance thresholds. Right: percentage of reconstructions with F-score above a value specified on the horizontal axis, with a distance threshold $d = 1\%$.

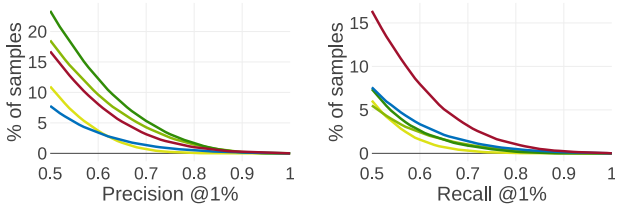

Figure 11. Percentage of samples with precision (left) and recall (right) of 0.5 or higher. Existing CNN-based methods show good precision but miss parts of objects, which results in lower recall.

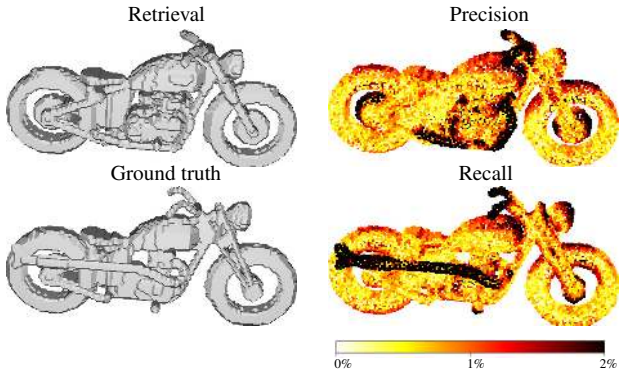

Figure 12. Visualizing precision and recall provides detailed information about which object parts were reconstructed correctly. Colors encode the normalized distance between shapes (as used for the distance threshold).

the retrieval baseline show higher recall. This is supported by qualitative observations that OGN and Matryoshka Networks tend to produce incomplete models.

Both recall and precision can be easily visualized to gain further insights, as illustrated in Fig. 12.

5.3. Dataset

The problem of networks finding a semantic shortcut solution is closely related to the choice of training data. The ShapeNet dataset has been used extensively because of its size. However, its particular composition – single objects of representative types, aligned to a canonical reference frame – enables recognition models to masquerade as reconstruction. In Fig. 2, we demonstrate that a retrieval solution (Oracle NN) outperforms all other methods on this dataset, *i.e.*, the test data can be explained by simply retrieving models from the training set. This indicates a critical problem in using ShapeNet to evaluate 3D reconstruction: for a typical shape in the test set, there is a very similar shape in the training set. In effect, the train/test split is contaminated, because so many shapes within a class are similar. A reconstruction model evaluated on ShapeNet does not need to actually perform reconstruction: it merely needs to retrieve a similar shape from the training set.

6. Conclusion

In this paper, we reasoned about the spectrum of approaches to single-view 3D reconstruction, spanned by reconstruction and recognition. We introduced two baselines, classification and retrieval, which leverage only recognition. We showed that the simple retrieval baseline outperforms recent state-of-the-art methods. Our analysis indicates that state-of-the-art approaches to single-view 3D reconstruction primarily perform recognition rather than reconstruction. We identify aspects of common experimental procedures that elicit this behavior and make a number of recommendations, including the use of a viewer-centered coordinate frame and a robust and informative evaluation measure (the F-score). Another critical problem, the dataset composition, is identified but left unaddressed. We are working towards remedying this in a subsequent work.

Acknowledgements

We thank Jaesik Park for his help with F-score evaluation. We also thank Max Argus and Estibaliz Gómez for valuable discussions and suggestions. This project used the Open3D library [58].

References

[1] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An information-rich 3D model repository. *CoRR*, abs/1512.03012, 2015. 2, 3

[2] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3D model retrieval. *Comput. Graph. Forum*, 22(3):223–232, 2003. 4

[3] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NIPS*, 2016. 2

[4] Christopher Bongsoo Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, 2016. 2, 3, 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[6] Jean-Denis Durou, Maurizio Falcone, and Manuela Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *CVIU*, 109(1):22–43, 2008. 2

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 2

[8] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 2, 3, 4, 7

[9] P. Favaro and S. Soatto. A geometric approach to shape from defocus. *TPAMI*, 27(3):406–417, 2005. 2

[10] Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2

[12] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A papier-mâché approach to learning 3D surface generation. In *CVPR*, 2018. 1, 2, 3, 4, 7

[13] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3D object reconstruction. In *3DV*, 2017. 2, 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[15] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Automatic photo pop-up. *ACM Trans. Graph.*, 24(3):577–584, 2005. 2

[16] Berthold K.P. Horn. *Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1970. 2

[17] Adrian Johnston, Ravi Garg, Gustavo Carneiro, and Ian D. Reid. Scaling CNNs for high resolution volumetric reconstruction from a single image. In *ICCV Workshops*, 2017. 2

[18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2

[19] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *TPAMI*, 36(11):2144–2158, 2014. 2

[20] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *CVPR*, 2018. 2

[21] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4):78:1–78:13, 2017. 7

[22] Chen Kong, Chen-Hsuan Lin, and Simon Lucey. Using locally corresponding CAD models for dense 3D reconstructions from a single image. In *CVPR*, 2017. 2

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 4

[24] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964. 4

[25] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J. Guibas. Joint embeddings of shapes and images via CNN image purification. *ACM Trans. Graph.*, 34(6):234:1–234:12, 2015. 4

[26] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 2

[27] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3D object reconstruction. In *AAAI*, 2018. 2

[28] Angeline M. Loh. *The recovery of 3-D structure using visual texture patterns*. PhD thesis, University of Western Australia, 2006. 2

[29] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *SIGGRAPH*, 1987. 3

[30] Zhaoliang Lun, Matheus Gadelha, Evangelos Kalogerakis, Subhransu Maji, and Rui Wang. 3D shape reconstruction from sketches via multi-view convolutional networks. In *3DV*, 2017. 2

[31] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1967. 3

[32] F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. 5

[33] Chengjie Niu, Jun Li, and Kai Xu. Im2Struct: Recovering 3D shape structure from a single rgb image. In *CVPR*, 2018. 2

[34] Martin R. Oswald, Eno Töppe, and Daniel Cremers. Fast and globally optimal single view reconstruction of curved objects. In *CVPR*, 2012. 2

[35] Jhony Kaesemodel Pontes, Chen Kong, Anders Eriksson, Clinton Fookes, Sridha Sridharan, and Simon Lucey. Compact model representation for 3D reconstruction. In *3DV*, 2017. 2

[36] Danilo Jimenez Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3D structure from images. In *NIPS*, 2016. 2

[37] Stephan R. Richter and Stefan Roth. Matryoshka networks: Predicting 3D geometry via nested shape layers. In *CVPR*, 2018. 1, 2, 3, 4

[38] Gernot Riegler, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger. OctNetFusion: Learning depth fusion from data. In *3DV*, 2017. 2

[39] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Comp.*, 18(5):401–409, 1969. 4

[40] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. Learning depth from single monocular images. In *NIPS*, 2005. 2

[41] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Learning 3-D scene structure from a single still image. In *ICCV*, 2007. 3

[42] Daeyun Shin, Charless Fowlkes, and Derek Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3D object shape prediction. In *CVPR*, 2018. 3, 5

[43] Amir Arsalan Soltani, Haibin Huang, Jiajun Wu, Tejas D Kulkarni, and Joshua B Tenenbaum. Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In *CVPR*, 2017. 2

[44] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3D: Dataset and methods for single-image 3D shape modeling. In *CVPR*, 2018. 2, 7

[45] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3D models from single images with a convolutional network. In *ECCV*, 2016. 2

[46] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *ICCV*, 2017. 1, 2, 3, 4

[47] Shubham Tulsiani, Abhishek Kar, João Carreira, and Jitendra Malik. Learning category-specific deformable 3D models for object reconstruction. *TPAMI*, 39(4):719–731, 2016. 2

[48] Shubham Tulsiani, Hao Su, Leonidas J. Guibas, Alexei A. Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *CVPR*, 2017. 2

[49] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 2

[50] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *ECCV*, 2018. 2

[51] O. Wiles and A. Zisserman. SilNet: Single- and multi-view reconstruction by learning from silhouettes. In *BMVC*, 2017. 2

[52] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *NIPS*, 2017. 2

[53] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *NIPS*, 2016. 2

[54] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *CVPR*, 2018. 2

[55] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision. In *NIPS*, 2016. 2, 3, 4

[56] Li Yi, Lin Shao, Manolis Savva, et al. Large-scale 3D shape reconstruction and segmentation from ShapeNet Core55. *CoRR*, abs/1710.06104, 2017. 3

[57] Ruo Zhang, Ping sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape from shading: A survey. *TPAMI*, 21:690–706, 1999. 2

[58] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 8

[59] Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, and Simon Lucey. Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In *ICCV*, 2017. 2