DR-KFD: A Differentiable Visual Metric for 3D Shape Reconstruction

Jionghao Jin  
Simon Fraser University

Akshay Gadi Patil  
Simon Fraser University

Hao (Richard) Zhang  
Simon Fraser University

Abstract

We advocate the use of differential visual shape metrics to train deep neural networks for 3D reconstruction. We introduce such a metric which compares two 3D shapes by measuring visual, image-space differences between multi-view images differentiably rendered from the shapes. Furthermore, we develop a differentiable image-space distance based on mean-squared errors defined over HardNet features computed from probabilistic keypoint maps of the compared images. Our differential visual shape metric can be easily plugged into various reconstruction networks, replacing the object-space distortion measures, such as Chamfer or Earth Mover distances, so as to optimize the network weights to produce reconstruction results with better structural fidelity and visual quality. We demonstrate this both objectively, using well-known visual shape metrics for retrieval and classification tasks that are independent from our new metric, and subjectively through a perceptual study.

1. Introduction

Shape distortion metrics play a critical role in geometric deep learning. Most neural networks developed for 3D reconstruction [10, 9, 30, 11, 33] define their loss functions, i.e., reconstruction errors, using a distortion metric. Often, the reconstruction results are also evaluated, against the ground truth, using the same or other metrics. To date, the most frequently used distortion metrics include Chamfer Distance (CD), Earth Mover Distance (EMD), Mean Square Error (MSE), and Intersection over Union (IoU). To serve as a network loss for back propagation, the metric needs to be differentiable. All of CD, EMD, and MSE are differentiable. While the original IoU is not, there have been recent attempts to make it differentiable [37, 17].

What is common about these well-adopted distortion metrics is that they are all defined in object-space, rather than image-space. They all measure geometric relations between shape elements in 3D space, but do not account for how the shapes are viewed by human observers, i.e., they are not visual shape metrics. Some recent works [7, 20, 31, 25, 9, 26] have pointed out the relative insensitivity of these object-space shape metrics to structural errors such as missing parts and topological noise, and visual artifacts such as self intersections and poor surface quality. For example, thickening and elongating/shortening all four legs of a chair may result in a larger CD than removing one of the legs entirely, yet the latter alteration, a structural change, is more visually apparent; see Figure 2. As a result, while some reconstructed 3D shapes do exhibit better visual quality, ratings based on distortion measures such as CD, EMD, MSE, or IoU may not reflect that superiority. This is not entirely surprising — a recent work by Blau and Michaeli [3] even suggests a trade-off between perceptual and distortion measures, albeit for image restoration tasks.

In computer graphics, and in particular, for 3D shape retrieval, one of the best known visual metrics is the light field descriptor (LFD) [6]. LFDs are computed for silhouette images of 3D shapes rendered from multiple viewpoints sampled around the shapes. Both a contour-based (Fourier descriptor) and a region-based (Zernike moment) image-space descriptor are employed, where rotational alignment is resolved via a discrete exhaustive search. However, due to the discrete rasterization during rendering and the use of truncated Fourier descriptors, LFD is non-differentiable.

Figure 1: Single-view 3D reconstruction by AtlasNet [10] and Matryoshka Net [26] were trained using object-space shape distortion losses, CD and IoU, respectively. Replacing them using DR-KFD, our differentiable visual metric, to train these networks produces results of higher visual fidelity in terms of shape structures and surface quality.
In this paper, we advocate the use of differentiable visual shape metrics to train deep neural networks (DNNs) for 3D reconstruction. We introduce such a metric which compares two 3D shapes by measuring visual, image-space differences between multi-view images rendered from the shapes, similar to LFD. However, one key difference is that the rendering process is differentiable, by employing a simplified soft rasterization [17]. In addition, we develop a differentiable image-to-image distance based on MSE defined over probabilistic keypoint maps of the compared images, rather than on RGB values. Furthermore, the MSE is defined over HardNet [21] features, rather than on the original keypoint maps. This choice is motivated in part by the finding in [3] that the perception-distortion trade-off appears less severe for distance between VGG features. Putting all these together, we arrive at a differentiable shape metric, which we call DR-KFD to capture the use of differentiable rendering, keypoint maps, and feature-space image-to-image distances. DR-KFD better matches visual evaluation by humans, compared to the object-space metrics.

To the best of our knowledge, existing 3D reconstruction networks all employ object-space reconstruction losses; see Table 1 for a summary. Our differential visual shape metric can be easily plugged into these networks, replacing the object-space distortion measures so as to optimize the network weights to produce reconstruction results with better structural fidelity and visual quality. We demonstrate this both objectively, using well-known visual shape metrics for retrieval and classification tasks that are independent from DR-KFD, and subjectively through a user study.

Specifically, the 3D reconstruction networks tested for adaptation to DR-KFD training include OGN, AtlasNet [10], and Matryoshka Networks [26], which were picked as representative networks by Tatarchenko et al. [31] in their recent systematic study of single-view 3D reconstruction. In addition, we also consider Pixel2Mesh [33] and 3D-R2N2 [8]. The visual shape metrics we employ for evaluation include Shape Google [4], Multiview CNN or MVCNN [29], normal consistency or NC [20], F-score [31, 26], as well as LFD. It is worth noting that the last three metrics have all been adopted by recent works on 3D reconstruction [20, 31, 26, 7] as visual metrics to complement the object-space metrics.

2. Related work

DNNs for 3D reconstruction. One of the earlier DNNs for 3D reconstruction is 3D-R2N2 [8], which proposes the use of Recurrent Neural Networks (RNNs) to reconstruct a voxelized shape from a single-view image. Another notable work, PointOutNet [9], produces multiple reconstruction candidates for a single-view input image and demonstrates an improvement in the reconstruction quality over 3D-R2N2. OGN [30] generates volumetric 3D outputs in a compute-and-memory efficient manner by using an octree representation. AtlasNet [10] uses a collection of parametric surface elements to represent a 3D shape and to naturally infer a surface representation of the shape. Pixel2Mesh [33] produces 3D triangle meshes from a single image using Graph Convolution Networks (GCN) [15]. Matryoshka Network [26] introduces a novel and efficient 2D encoding scheme for 3D geometry, posing 3D reconstruction as a 2D prediction problem, while also speeding up the process.

Recent works on generative shape modeling using implicit functions have achieved state-of-the-art visual quality. IM-NET [7] and OccNet [20] represent 3D surfaces implicitly in the form of continuous binary decision functions. DeepSDF [12] learns a continuous Signed Distance Function (SDF) representation for a 3D shape. DISN [35] combines implicit SDF descriptors with the local input image features to generate 3D shape surfaces. All these implicit function-based networks can generate 3D models with a smooth surface and good visual quality. Despite the good visual quality of the generated surfaces, they often obtain lower scores on object-space metrics.

Shape metrics for 3D reconstruction. Table 2 lists most of the recent DNNs for 3D reconstruction, along with the shape metrics they employed for training and evaluation. In
principle, a network that is trained on a particular loss, e.g., CD, should be expected to perform better on that metric during testing, than methods that were not trained with the metric. All methods that we are aware of train their networks using object-space distortion metrics. CD and EMD, and to a lesser degree, IoU, are the dominant measures applied for evaluating reconstruction quality. It is interesting to observe that some of the most recent methods, including IM-Net [7], OccNet [20], and Matryoshka Network [26], have opted to evaluate their methods using alternative measures such as LFD, Normal Consistency (NC), and F-score. Not surprisingly, each of these works pointed out the shortcomings of object-space distortion metrics in capturing visual similarity, which motivated their use of other alternatives.

In terms of visual quality of the 3D reconstruction, the current state-of-the-art results are obtained by methods based on learning implicit fields [7, 20, 12, 35].

**Visual shape metrics.** In the subfields of shape retrieval and classification, many visual shape similarity measures have been proposed. In this paper, we have chosen a subset of them as representative measures. LFD [6] performs visual shape similarity by extracting local features from one-hundred orthogonal projections of each 3D model. Shape-Google [4] constructs compact and informative shape descriptor using a bag-of-visual-features. MVCNN [29] combines information from multiple views of a 3D shape into a single and compact shape descriptor with the help of CNNs. F-score [31] performs visual shape similarity explicitly by calculating the distance between object surfaces using a harmonic mean of precision and recall scores. And finally, normal consistency (NC) [20, 33] measures how well any 3D reconstruction can capture higher order information by calculating a mean absolute dot product of face-normals of the given 3D models, represented as meshes. However, LFD, ShapeGoogle, F-score are not differentiable, and NC measure can only capture higher-order information.

**Image-space similarity.** Most common quantitative measures of image similarities include pixel-wise MSE or IoU. Such measures may not always capture the visual similarity between a pair of images. Hand-crafted image descriptors such as SIFT [18], SURF [2] and ORB [27] can be used to perform image retrieval to visually correspond closest image matches, but are discrete and non-differentiable in nature. On the other hand, the use of powerful image descriptors obtained from CNNs have been shown to make the image retrieval pipeline differentiable [36, 22, 23, 19, 29], allowing an end-to-end neural network approach for measuring image-similarity.

Generic image similarity computation has to account for rotation, translation, and occlusion of identical entities in the two images, as well as the overall illumination variation. However, the 3D models input to our framework are aligned and consequently, the rendered view-images are aligned as well, and as such do not pose the above challenges. Our inspiration to develop a differentiable and visual image-similarity metric for rendered images of 3D shapes stems from the development of differentiable and visual image-retrieval pipelines based on CNNs.

**Differentiable rendering.** Non-deep learning based renderers employ classical rendering techniques [13, 28] such as discrete rasterization and therefore, are non-differentiable. Advances in deep learning have propelled the evolution of near differentiable rendering frameworks based on CNNs; e.g. 3D-RCNN [16] and RenderNet [24]. These works are not pluggable into other networks because of their computational cost. [14] proposed a neural 3D differentiable mesh renderer that approximates the gradient of discrete raster operation with a linear function.

In our work, we incorporate a simplified version of the recent work of Soft Rasterizer [17], which converts the discrete raster operation to a probabilistic soft raster one, directly solving the indiferentiable raster operation of the

### Table 1: Training loss and evaluation metrics of state-of-the-art 3D reconstruction networks.

| Method                  | Representation        | Training loss        | Evaluation metrics          |
|-------------------------|-----------------------|----------------------|-----------------------------|
| AtlasNet [10]           | Point Cloud+ mesh     | CD                   | CD                          |
| PointOutNet [9]         | Point Cloud           | CD, EMD              | CD, EMD                     |
| OGN [30]                | Octree                | cross-entropy/MSE    | CD, EMD                     |
| HSP [11]                | Voxel                 | cross-entropy/MSE    | CD, IoU                      |
| Pixel2Mesh [33]         | Point Cloud+ mesh     | CD, surface normal loss | CD, EMD                     |
| Matryoshka Networks [26]| Shape layer from Voxel | IoU and cos-similarity | IoU                          |
| IM-Net [7]              | Implicit Field        | Per-point in/output status | CD, MSE, IoU, EMD, LFD     |
| OccNet [20]             | Implicit Field        | Per-point in/output status | CD, IoU, normal consistency |
| DeepSDF [12]            | Implicit Field        | Per-point in/output status | CD, EMD                     |
| DISN [35]               | Implicit Field        | Per-point in/output status | CD, EMD, F-score, IoU      |

CD and EMD are widely adopted in point-based networks, while MSE and cross-entropy loss and differentially implemented Intersection of Union(IoU) loss are more popular in voxel based reconstruction. Implicit decoders all adopted per-point reconstruction loss.
traditional renderer. The modified adoption of soft rasterization to render view-images of the given 3D model is far more efficient than other CNN-based methods that can produce good renderings.

3. Differential visual metric for 3D shapes

Objective visual metrics for assessing 3D shape similarity are mostly LFD [6] dependent, which in-turn, is based on the similarity of rendered view-images of the 3D models from all viewing angles. We follow this theme of rendered-and-match-images to determine the visual similarity of two 3D models and also make our framework differentiable. An overall pipeline of our approach is shown in Figure 4. Given two 3D models, one reconstructed from a single-view input image using an existing approach such as AtlasNet [10], Matryoshka Network [26], Pixel2Mesh [33] or OGN [30], and the other being the corresponding ground truth (GT) model, we first align them and render the models from twenty-five viewing angles, using the Soft Rasterizer [17] renderer, which is differentiable. For each view-image, a Point of Interest (PoI) map is obtained using a Keypoint detection network. Visual similarity of the 3D models is determined using DR-KFD feature matching (see Figure 4) obtained by extracting local features from PoI maps.

3.1. Differentiable Renderer

To accurately determine 3D shape similarity using LFD, view-images from all angles and at every level-of-detail are required. However, due to hardware constraints, we only use twenty-five viewpoints, sampled on a semi-sphere as an approximation to the ideal case. To achieve differentiability in the rendering process, we replace rasterization and z-buffering operations used in conventional rendering pipelines with soft rasterization and probabilistic map aggregation [17]. No color information is used in DR-KFD renderer as many reconstruction networks are incompatible with texture projections on the reconstructed model surfaces. A fixed position light source is aptly placed to provide information about shape surface quality when rendering the view-images. We make use of the rendered view-images to determine the similarity of given 3D models, as explained below.

3.1.1 Image Matching

Our differentiable and visually sensitive metric is based on matching rendered view-images by extracting discriminative local features over Point of Interest (PoI) maps, using Convolutional Neural Networks (CNNs).

PoI map. A PoI map is a probabilistic map that gives a score for every pixel being a keypoint. It is generated by a CNN-based keypoint detection network corresponding to every view-image. To this end, we borrow the LIFT detector introduced in [36], which uses piece-wise linear activation functions [32] in convolution layers to get a PoI map for an input image. It is formulated as:

$$PoI_{map} = f_{Net}(I) = \sum_{n}^{N} \delta_{n} \max_{m}(W_{mn} \odot I + b_{mn}) \quad (1)$$

where $f_{Net}(I)$ is a non-linear function of the rendered view-image $I$, using a neural network $Net$, which is nothing but a CNN-based keypoint detector. $N, M$ are hyperparameters controlling the complexity of the piece-wise linear activation function. $\delta$ is +1 if $n$ is odd, and -1 otherwise. The parameters of the network $Net$ to be learned are the convolution filter weights $W_{mn}$ and biases $b_{mn}$, and $\odot$ denotes the convolution operation. A detailed description of LIFT based keypoint detection can be found in [36].

Continuous patch features. After generating PoI maps for each of the rendered images, local patch features are extracted from them using a sliding window of size 32 $\times$ 32 with a stride of 16, resulting in 225 (15 $\times$ 15) patches per image $I$. These patches span the entire image and there exists an overlap between adjacent patches which provides continuity over the feature space. Note that in existing image retrieval works, such as in [36, 32], detected keypoints are projected back onto the original image and image features are extracted locally around these keypoints. Image retrieval is performed by matching these keypoint-based local features. This is a discrete process over the image space. As a result, this approach of matching images based on discrete keypoints, if employed, introduces non-differentiability into the latter parts. Unlike such methods, we avail continuous local patch features from the PoI maps by employing a sliding window technique spanning the entire map. Our framework is inspired by the recent work of [3] which shows that local patch features help alleviate the problem of “human perception vs image distortion” trade-off. A sliding window on a PoI map, followed by the HardNet[21] feature extractor outputs a 128-D feature descriptor per patch (see Figure 3). The HardNet [21] module we use is pretrained on UBC PhotoTour[34], a standard dataset used to extract local patch features. We coin these features as the DR-KFD feature descriptors. The continuous patch features, coupled with the early-stage render (Soft Rasterizer) allow our pipeline to be entirely differentiable, demonstrated further by our end-to-end learning framework in Figure 4.

Feature Matching. After extracting the DR-KFD feature descriptors for the two 3D models, we perform feature matching. For every DR-KFD feature descriptor of the reconstructed 3D model, we find the best DR-KFD feature descriptor of the GT 3D model, in terms of minimum mean squared error (MSE) loss. We do not correspond patches merely based on their position in the PoI maps. This can be thought of as a feature matching function. Symbolically, this function finds the best match for every PoI patch of the
Figure 3: An overall framework of obtaining DR-KFD feature descriptors via continuous patch features (Section 3.1.1). In the first step is the differentiable renderer, Soft Rasterizer, used to render the view-images. This renderer, coupled with the process of continuous patch feature extraction over the PoI maps, allows our framework to be entirely differentiable.

Figure 4: An end-to-end pipeline of our visual, differentiable shape-similarity metric framework for single view 3D reconstruction. Feature descriptors from Figure 3 are used to calculate the overall loss, which is backpropagated to the 3D reconstruction network (which reconstructs a 3D model, shown on the left) via the associated DR-KFD module, improving the reconstruction quality.

4. Results and Evaluation

We assess the impact of our differentiable visual metric, DR-KFD, both qualitatively and quantitatively, on state-of-the-art single-view 3D reconstruction networks. A gallery of qualitative comparison on results generated when such networks are trained using DR-KFD is shown in Figure 6.

4.1. Quantitative evaluation

We consider existing single-view 3D reconstruction networks, including 3D-R2N2 [8], OGN [30], AtlasNet [10], Matryoshka Network [26], and Pixel2Mesh [33], to quantify both (object-space) shape distortion (CD and IoU) and the visual quality of the reconstructed models. The networks are trained either using their original losses or our metric DR-KFD. For visual quality measures, we choose representative tools including Shape Google, Normal Consistency, F-score, LFD, and MVCNN.

Dataset. All the tested networks are trained and evaluated on the large-scale 3D CAD model dataset, ShapeNet [5]. We use four shape categories to train and test all the net-
Table 2: Quantitative comparison results for single-view 3D reconstruction networks when trained using their original (object-space) losses vs. using DR-KFD, our new differentiable visual metric. The evaluation is done on both object-space (CD and IoU) and visual shape categories. Numbers reported represent average performance over all four object categories.

Comparison results. Table 2 shows the comparison results, where the numbers report average performance over all four object categories. Note that IM-NET employs per-point reconstruction losses and it is unclear how we could have trained it using our new metric. However, we report the IM-NET numbers for reference purposes.

Our first observation, perhaps a sanity check, is that when a network is trained using the new metric DR-KFD, it always performed better in DR-KFD, compared with the original network. Second, when trained using DR-KFD, some of the networks even outperformed their counterparts trained with the original (object-space) losses, when the results are evaluated using CD and IoU — this is the case 4 out of 10 or 40% of the time. In particular, Matryoshka was trained using IoU, but training with DR-KFD outperformed the original network when evaluated on IoU. Next, perhaps most importantly, we see that in majority of the cases (24 out of 30 or 80% of the time), replacing the original object-space losses using DR-KFD, the networks improved their performance in terms of visual shape metrics.

4.2. Qualitative evaluation

To understand the visual quality of the reconstructed 3D models trained using DR-KFD as the network loss, we conduct two kinds of perceptual studies (PS) on Amazon Mechanical Turk (AMT), as described below.

PS-1. In this study, we present 50 questions, each containing a single-view image and a pair of reconstructed models, one using the original loss and one using DR-KFD. Each model is rendered along two views, with one of the views aligned to the input image. We use the reconstructed models from AtlasNet [10] and Matryoshka Net [26]. Models are selected at random from the entire reconstructed set. Turkers are presented with 50 questions in a random order, with randomized orderings of the reconstructed models as well as their view-images.

PS-2. We essentially repeat PS1, but instead of selecting the models from the entire reconstructed set, we select them randomly from the top-20% of the reconstructed set, filtered based on the scores obtained using three visual measures: LFD, ShapeGoogle and MvCNN.

Both PS1 and PS2 are forced-choice responses involving 80 different participants per study. For PS-1, DR-KFD metric received 71% of the total votes (4000) and for PS-2, the percentage share shot up to 86.075%, indicating a positive trend when high quality visual samples are considered for perceptual evaluation. Using the Wilson score confidence interval for Bernoulli trials [1], there is a 95% chance that users will select our metric at least 69.57% and 84.95% of the times, for PS-1, PS-2, respectively.

PS-3. The theme of PS-3 is a bit different from the previous two. As IM-NET [7] wins 50% of the times (see Table 2) over the six visual similarity measures (ShapeGoogle, F-Score, NC, LFD, MVCNN and DR-KFD) in terms of single-view 3D reconstruction, we perform a perceptual study against IM-NET. Each of the 50 questions presented consist of reconstructed 3D models from IM-NET, AtlasNet/Matryoshka Net with the original loss and AtlasNet/Matryoshka Net with DR-KFD loss. Each of the models is rendered in two view-angles. Turkers are asked to rank the three models in terms of their visual appeal (close-ness to the input image and the reconstruction quality). A
Figure 5: Vertical positional shifts, and deletion of the leg bars. Binary images are shown on top and their corresponding PoI maps are on the bottom. Image similarity scores w.r.t. the reference image for each operation using different image-level features are tabulated in Table 3.

total of 80 participants took the study. The results are tabulated in Table 4. IM-NET wins over our metric with a slight margin for the first rank, indicating that reconstruction networks inferior to IM-NET can provide visually comparable reconstruction results when trained using our metric as the loss function. Models reconstructed using the original loss are the least preferred.

4.3. What do DR-KFD features capture?

Given a four-legged chair with a bar between adjacent pair of legs, humans can still relate to it even after vertical shifts of the bars. However, if all the bars are removed, one cannot firmly associate the new chair to the old one (see Figure 5). It is not clear as to which kind of image-level features capture such object space manipulation well. We investigate this problem by calculating the image distance (MSE loss) using different features: raw image pixels, image features based on LIFT descriptors [36], PoI map and our DR-KFD features. The input image and its corresponding PoI map are shown in Figure 5. In our experiment, vertical shifts of the bars are obtained by moving them five pixels up/down. From Table 3, We observe that DR-KFD is more sensitive towards the deletion operation compared to others. Moreover, image matching using DR-KFD and LIFT based local image features seem to be tolerant to small part shifts, while pixel-based MSE is quite sensitive to such changes.

5. Conclusion

We make a first step towards improving the visual quality of 3D reconstruction networks by making their training/reconstruction loss functions visual. Overall, the new differentiable visual metric we develop, DR-KFD, is shown to improve the reconstruction quality for all the tested networks (AtlasNet, Matryoshka, Pixel2Mesh, OGN, and 3D-R2N2), as judged by a variety of visual similarity measures (LFD, MVCNN, Shape Google, etc.). This is clearly a positive trend to motivate further investigation. However, the demonstrated advantage of DR-KFD is not yet so lopsided. Indeed, our metric is still rather primitive as it does not utilize the most advanced and up-to-date tools that are available for multi-view rendering, feature extraction and learning, or image-space/perceptual assessment.

In general, the question of what the best 3D reconstruction is should depend on the application. For example, if the goal is to recognize or classify the shape, then visual similarity is more important. As well, functional understanding of an acquired shape hinges on accurate recovery of shape structures, which would support the use of visual metrics as training losses. On the other hand, if the reconstructed 3D shape is to reflect accurate physical measures, then object-space metrics should play a more prominent role.

One of the most obvious limitations of any visual metric (i.e., one judged from rendered views), DR-KFD included, is that it is oblivious to errors that are hidden from the viewers due to occlusion. Such errors are not visible on the projected images, but they can be captured by object-space shape distances. One possibility to explore is to combine differentiable object- and image-space shape metrics as the training loss, e.g., as a weighted sum which would preserve the differentiability. This could be a good strategy to address the distortion-perception trade-off [3], but leaves the question of how to choose the weight.
Figure 6: A gallery of reconstructed 3D models obtained from AtlasNet (AN) [10], Matryoshka Net (MN) [26], Pixel2Mesh (P2M) [33] and 3D-R2N2 [8], trained using the metrics as adopted in the respective works and by using our DR-KFD metric. Given an input image, replacing the original training loss with our DR-KFD loss results in an improvement in the visual quality of the reconstructed 3D models as shown above, and also supported by the numbers in Table 2.
References

[1] Alan Agresti and Brent A Coull. Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician, 52(2):119–126, 1998.
[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In European conference on computer vision, pages 404–417. Springer, 2006.
[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6228–6237, 2018.
[4] Alexander M. Bronstein, Michael M. Bronstein, Leonidas J. Guibas, and Maks Ovsjanikov. Shape google: Geometric words and expressions for invariance shape retrieval. Trans. Graph., 30(1):1:1–1:20, Feb. 2011.
[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012, 2015.
[6] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouyoung. On visual similarity based 3d model retrieval. In Computer graphics forum, volume 22, pages 223–232. Wiley Online Library, 2003.
[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In CVPR, 2019.
[8] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision, pages 628–644. Springer, 2016.
[9] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 605–613, 2017.
[10] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. Atlasnet: A papier-maché approach to learning 3d surface generation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[11] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3d object reconstruction. In 3DV, 2017.
[12] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In CVPR, 2019.
[13] James T Kajiya. The rendering equation. In ACM SIGGRAPH computer graphics, volume 20, pages 143–150. ACM, 1986.
[14] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3907–3916, 2018.
[15] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.
[16] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3559–3568, 2018.
[17] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In ICCV, 2019.
[18] David G Lowe et al. Object recognition from local scale-invariant features. In iccv, volume 99, pages 1150–1157, 1999.
[19] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2527–2536, 2019.
[20] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In CVPR, 2019.
[21] Anastasia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor’s margins: Local descriptor learning loss. In Advances in Neural Information Processing Systems, pages 4826–4837, 2017.
[22] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In Proceedings of the European Conference on Computer Vision (ECCV), pages 284–300, 2018.
[23] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In Proceedings of the European Conference on Computer Vision (ECCV), pages 284–300, 2018.
[24] Thu H Nguyen-Phuoc, Chuan Li, Stephen Balaban, and Yongliang Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. In Advances in Neural Information Processing Systems, pages 7891–7901, 2018.
[25] Jhony K Pontes, Chen Kong, Sridha Sridharan, Simon Lucey, Anders Eriksson, and Clinton Fookes. Image2mesh: A learning framework for single image 3d reconstruction. In Asian Conference on Computer Vision, pages 365–381. Springer, 2018.
[26] Stephan R. Richter and Stefan Roth. Matryoshka networks: Predicting 3d geometry via nested shape layers. In CVPR, 2018.
[27] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. Orb: An efficient alternative to sift or surf. In ICCV, volume 11, page 2. Citeseer, 2011.
[28] Peter-Pike Sloan, Jan Kautz, and John Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In ACM Transactions on Graphics (TOG), volume 21, pages 527–536. ACM, 2002.
[29] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In Proc. ICCV, 2015.
[30] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional archi-
tectures for high-resolution 3d outputs. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[31] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3D reconstruction networks learn? In CVPR, 2019.

[32] Yannick Verdie, Kwang Yi, Pascal Fua, and Vincent Lepetit. Tilde: a temporally invariant learned detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5279–5288, 2015.

[33] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[34] Simon AJ Winder and Matthew Brown. Learning local image descriptors. In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007.

[35] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. DISN: deep implicit surface network for high-quality single-view 3d reconstruction. CoRR, abs/1905.10711, 2019.

[36] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In European Conference on Computer Vision, pages 467–483. Springer, 2016.

[37] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, MM ’16, pages 516–520, 2016.