SurfNet: Generating 3D shape surfaces using deep residual networks-Supplementary Material

Ayan Sinha
MIT
sinhayan@mit.edu

Asim Unmesh
IIT Kanpur
a.unmesh@gmail.com

Qixing Huang
UT Austin
huangqx@cs.utexas.edu

Karthik Ramani
Purdue
ramani@purdue.edu

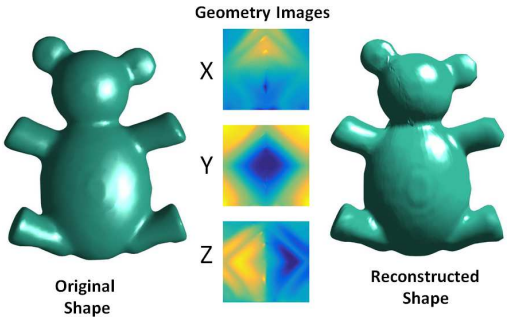

Figure 1. The x, y, z surface coordinates on the teddy model (left) are represented using 3 feature channels of a geometry image (center) and the surface reconstructed from the geometry image is shown to the right [3].

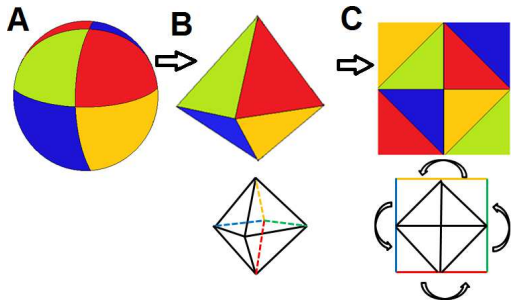

Figure 2. Creation of geometry image: (A) The mesh is first authalically parameterized on a sphere, then (B) projected onto an octahedron using area sampling to preserve the spherical triangular areas on the sphere, and finally (C) cut and unfolded along 4 edges as shown in the line plots below to create a flat and regular geometry image.

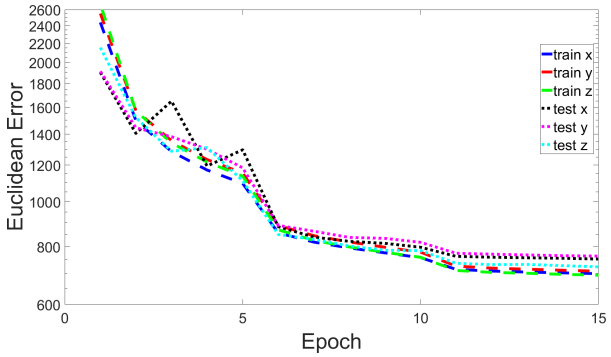

Figure 3. Error of x, y, z geometry image created by the deep residual network from a single depth image over training and test datasets for hands.

1. Geometry Images

As mentioned in the main manuscript, geometry images are a specific kind of surface parametrization wherein the geometry is resampled into a regular 2D grid akin to an image. As discussed in [3], geometry images reduce memory and complexity for learning shapes using CNNs over free boundary or disc parameterizations as every pixel encodes desired shape information. This is shown in figure 1 wherein the x, y, z coordinates of the teddy mesh model are encoded in a separate geometry image. The 3D surface reconstructed from this representation closely resembles the original mesh and preserves its prominent features. We follow the approach of [3] to create a geometry image which consists of authalically parameterizing a surface mesh on a spherical domain, then projecting it onto an octahedron and cutting to convert the original 3D shape into a flat and regular geometry image (see 2 and [2]). The geometry image representation possess symmetry beneficial during learning.

2. Non-rigid shapes

We first discuss the reconstruction error of our deep residual networks, and then provide additional qualitative results to validate our approach. Figure 3 shows the Euclidean distance error between the ground truth geometry image and the geometry image created by the deep neural network over epochs when shown a depth image. We do this for each of the three feature channels of the geometry image, *i.e.* the x, y, z coordinates and for both the training and test datasets. The hand is enclosed in approximately a $20 \times 20 \times 20$ bounding box centered at the origin. The ver-

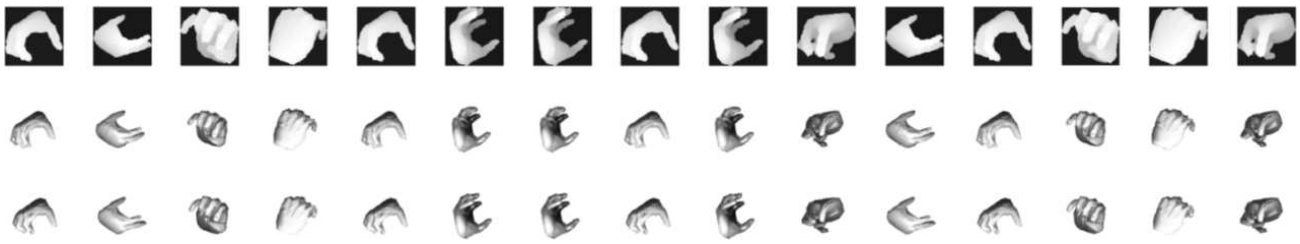

Figure 4. Results on test dataset for reconstructing the 3D shape surface of the hand from a single depth image. The first row is the depth image, the second row is the ground truth and the third row is our reconstruction.

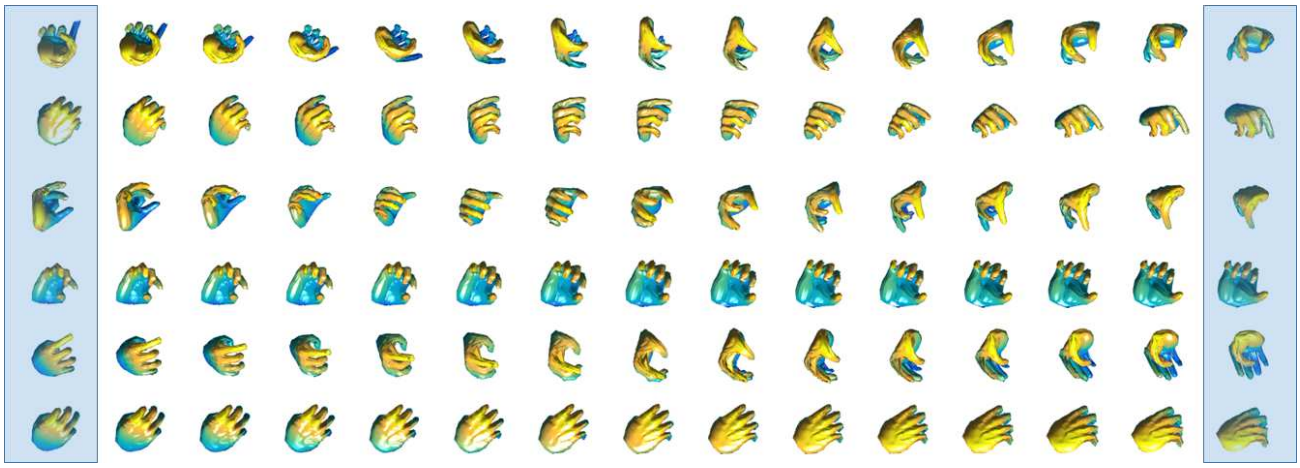

Figure 5. Each row shows the 3D surface plots of geometry images created by our neural network by inputting uniformly spaced parametric joint angle vectors. The highlighted shapes are in the training dataset.

tical axis shows the total sum of reconstruction error over all 4096 pixels in the 64×64 geometry image. We see that the errors decrease over epochs and the training set error is lower than the test set error, as is natural. The reconstruction errors for the x, y, z geometry image are approximately equal because the hand undergoes full articulation. We can evaluate the per pixel reconstruction error for a geometry image to be approximately 0.2 units, *i.e*. approximately 1% of the length of the bounding box. This suggest that our reconstruction from a depth image is very accurate.

Figure 4 shows additional 3D surface plots of the generated geometry image by our neural networks on the test depth images. We see that it is able to recover the full articulation of the hand very close to the ground truth even in the presence of occlusion. Next, we perform additional experiments on generative modeling of non-rigid shape surfaces from a parametric representation. We also create two random 18 dimensional vectors, and uniformly sampled from the linearly interpolated joint-angle values from the first to the second vector. The rows of figure 5 shows the output 3D surface plots reconstructed from the x, y, z geometry image feature channels. We see natural transition from the first to the second pose.

3. Rigid shapes

We now discuss the reconstruction error of the deep convolutional neural networks for rigid shape surface creation. Figures 6 and 7 show the Euclidean distance error between the ground truth geometry image and the geometry image created by the deep neural network over epochs when shown a single RGB image for the dataset of cars and airplanes respectively. We do this for each of the three feature channels of the geometry image, *i.e*. the x, y, z coordinates and for both the training and test datasets. The rigid models are enclosed in a $128 \times 128 \times 128$ bounding box centered at the origin. The vertical axis shows the total sum of reconstruction error over all 4096 pixels in the 64×64 geometry image. We see that the errors decrease over epochs and the training set errors are generally lower than the test set errors. Furthermore, we see that the error along the z dimension is lower than the other two dimensions. This is because we consider a limited range of elevation angle between 0 and 45 degrees. A Euclidean distance error of 10000 over 4096 pixels indicates that the per-pixel reconstruction error is about 2.5 units, *i.e*. about 2% of the length of the bounding box. This is a reasonable accuracy consid-

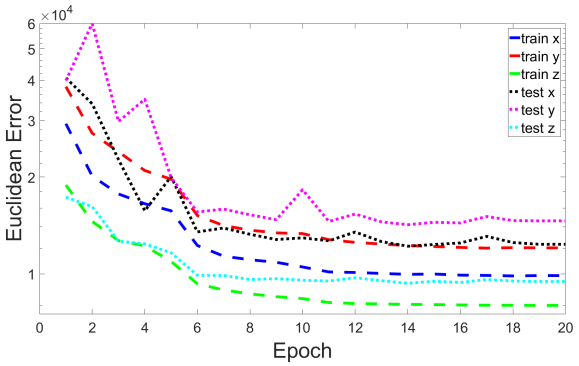

Figure 6. Error of x, y, z geometry image created by the deep residual network from a single depth image over training and test datasets for cars.

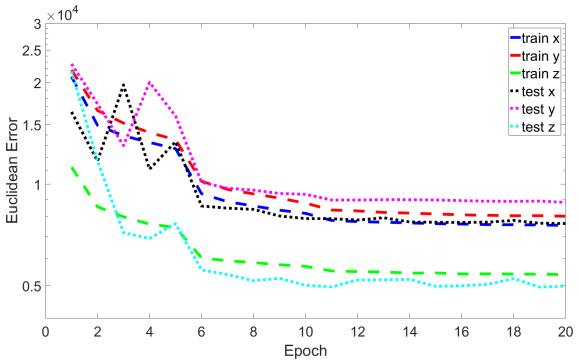

Figure 7. Error of x, y, z geometry image created by the deep residual network from a single depth image over training and test datasets for airplanes

ering both these categories of shapes have high intra-class variation. Our experiments indicate that a few samples for which the reconstruction fail dominate the contribution to this error. The failure cases are discussed later.

Figure 8 shows the total reconstruction error due to a one-hot encoded and view parameters fed into a deep neural network for two cases (1) A deep residual network that learns the geometry image feature channels directly in blue, and (2) A deep residual network that learns the residual of the the geometry image and the shape surface is constructed by adding the base geometry image to the residual geometry image in red. We see that the total reconstruction error, *i.e.*, the total sum over 4096 pixels each encoding the x, y, z coordinate of the surface model, is a lot lower over epochs for the neural network learning the residual geometry image compared to the neural network learning the geometry image directly. This highlights the benefit of learning the residual geometry image whenever possible in the spirit of deep residual networks [1]. Figure 9 shows the total reconstruction error over epochs for all models in the dataset when trained with different number of azimuth angles. The

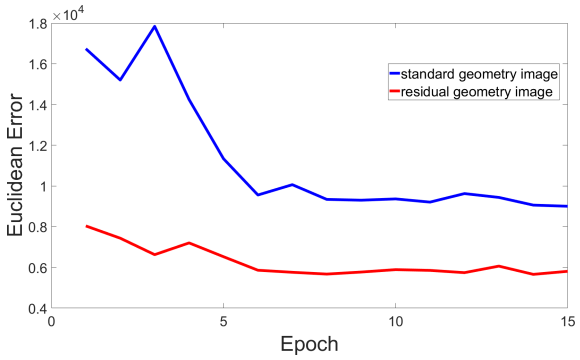

Figure 8. Error of 3D surface generation over epochs for residual geometry images and normal geometry images.

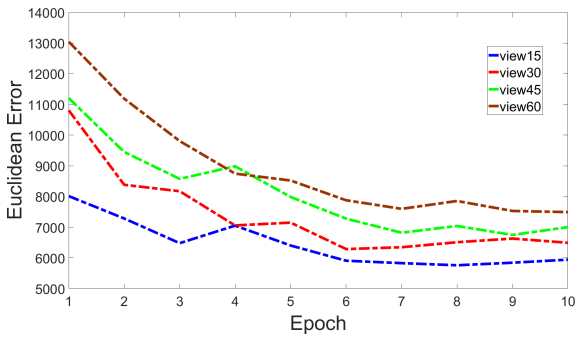

Figure 9. Error of viewpoint interpolation for different sizes of training set as determined by the interval between azimuth angles. The elevation angles are fixed at 4 values of [0, 15, 30, 45].

models in the dataset are in intervals of 15 degrees for the azimuth angle and this consists the base case shown in blue with naturally the lowest reconstruction error. The red line shows the total reconstruction error over epochs for all models in the dataset when trained with models at intervals of 30 degrees for the azimuth angle. In a similar manner, the green and the magenta plots show the total reconstruction error over epochs for all models in the dataset when trained with models at intervals of 45 degrees and 60 degrees respectively. We see that the error increases as we train the neural network with models at larger intervals of the azimuth angle.

Finally in figure 10 we qualitatively show the quality of our reconstructions for a single RGB image on the test car and airplane datasets. We see that the reconstruction is sensitive to the features on the airplane and the car, and often times our approach provides better reconstruction that even the ground truth model. See for example the seventh and eight examples for the airplane models. The ground truths are noisy because of poor correspondence identification by blended intrinsic maps. However, as our method learns a internal representation of a category of shapes, we are able to reconstruct the 3D shape surface with high fidelity.

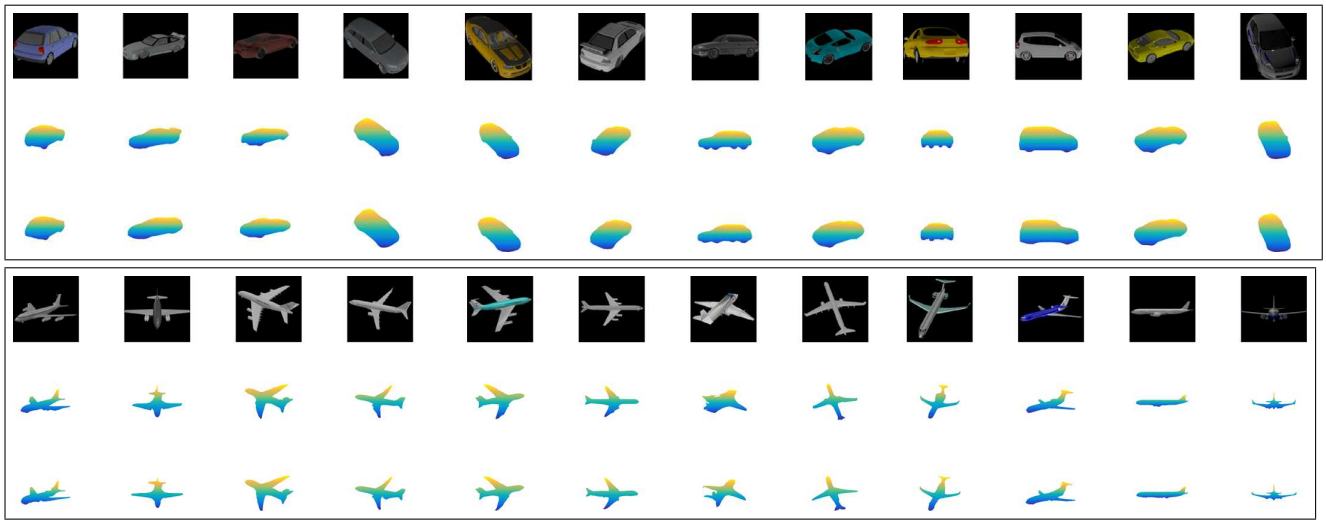

Figure 10. Qualitative evaluation of 3D surface reconstruction from a single image on car (top) and airplane (bottom) dataset.

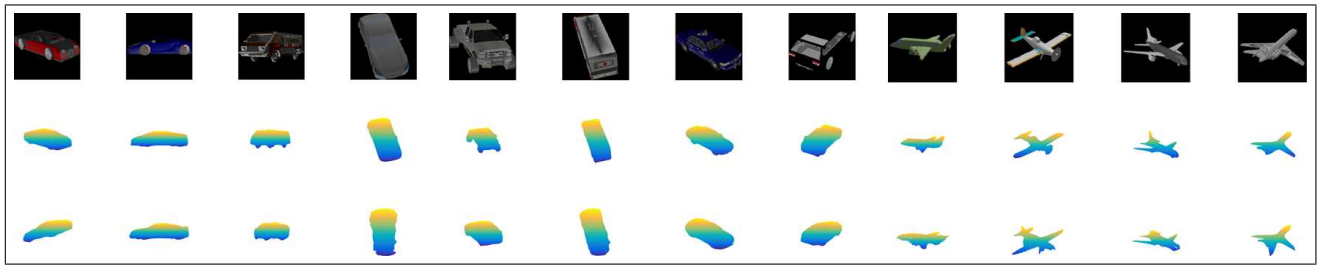

Figure 11. Failure cases for 3D surface reconstruction from a single image.

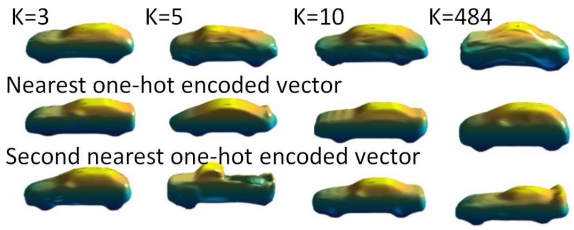

Figure 12. *Top row shows shapes created for different K's for K-hot encodings.*

Figure 11 shows some failure cases of our method. The surface reconstruction fails when the image is of low contrast with a lot of black pixels blending it with the background or insufficient training examples in the dataset for a particular subclass of shapes such as buses. Also, sometimes the poses are identified incorrectly, albeit rarely. The thin and elongated structure of the airplanes also results in some outlier points among some feature channels in a geometry image.

We can create more diverse content by inputting random values in K-hot encodings with sum of vector equal to 1. Figure 12 shows the cars created for values of $K = 3, 5, 10, 485$ respectively. We see the variability of new shape surfaces created by the network increases with K. As a final note, the total mean reconstruction error per shape at elevation 45° for a car network trained only on [0,15,30]° is 8143 compared to 6876 for a network trained on [0,15,30,45]° showing that our network can reasonably extrapolate. Also, the total reconstruction errors for cars with and without curvature weighted loss function are 1.76e4 and 2.77e4, respectively, indicating that high frequency features are better preserved. This relative improvement is more meaningful, as the absolute values are hard to gauge.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016. 3

[2] E. Praun and H. Hoppe. Spherical parametrization and remeshing. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 340–349. ACM, 2003. 1

[3] A. Sinha, J. Bai, and K. Ramani. Deep learning 3d shape surfaces using geometry images. In B. Leibe, J. Matas, N. Sebe,

and M. Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 223–240, Cham, 2016. Springer International Publishing. 1