SilNet: Supplementary Material

Olivia Wiles
ow@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

In this supplementary material we include additional details about the SilNet architecture and our two datasets – the blobby object dataset and the sculpture dataset – in sections A and B. We include additional results in section C.

A Additional details about the architecture.

The architecture of the 3D decoder is given in figure 1. It consists of four convolutional transposes, each of which are followed by a ReLU unit except for the last which is followed by a pixel-wise sigmoid layer. This generates a $57 \times 57 \times 57$ volume which is projected with the projection layer $T_{\theta'}$ to generate the silhouette at angle θ'.

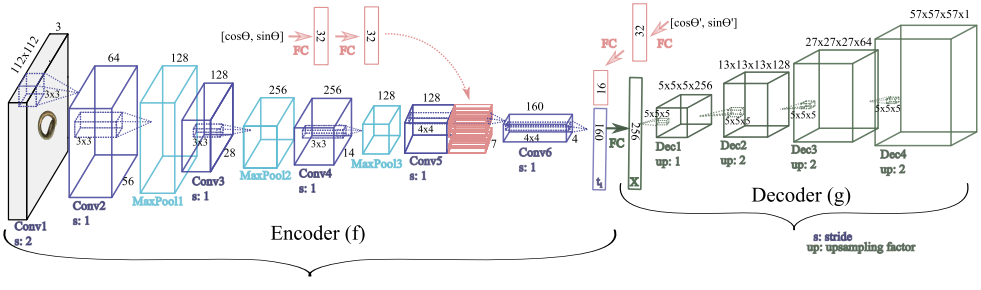


Figure 1: The 3D Decoder.

B Additional details about the datasets.

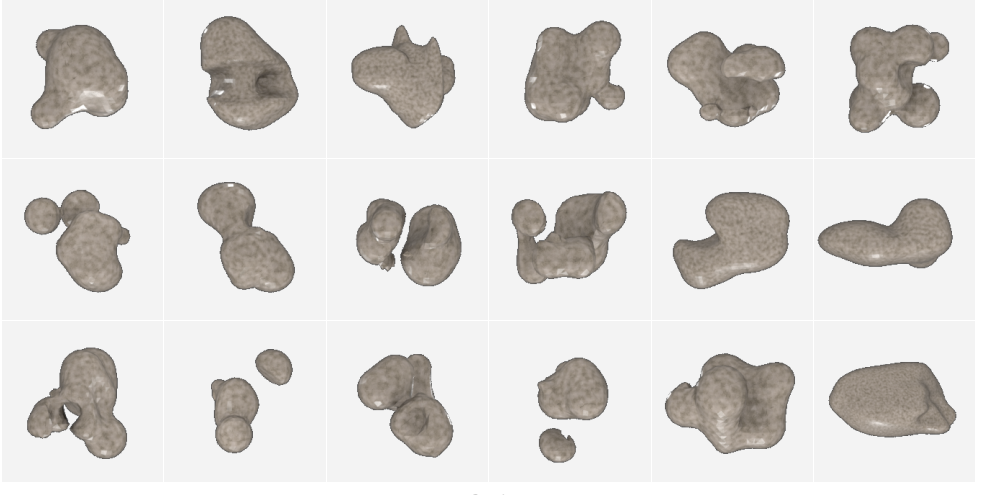
The sculptures were downloaded from SketchFab[1]. In particular the sculptures of two users – *geoffreymarchal* and *britishmuseum* – were used. Additional samples of objects from both the blobby object dataset and the sculpture dataset are provided in figure 2.

C Additional results.

We include additional results for the blobby object dataset, sculpture dataset, and ShapeNet in sections C.1, C.2, and C.3 respectively.

[1] https://sketchfab.com

Sculpture Dataset



Blobby Obj Dataset

Figure 2: A selection of objects from both the blobby object dataset and the sculpture dataset. These examples demonstrate the variety, complexity, and realism of the sculpture dataset. Moreover, they demonstrate the variety of viewpoints and rotations. For example, the two top rightmost sculpture images demonstrate an un-natural rotation of the bust. We have no assumption that the figures will be upright. Indeed in the examples above some sculptures are at 45 degrees, some are horizontal, and some are upside down. Zoom in for details.

C.1 Blobby object dataset.

Additional samples of SilNet's results on the blobby object dataset are provided in figure 3.

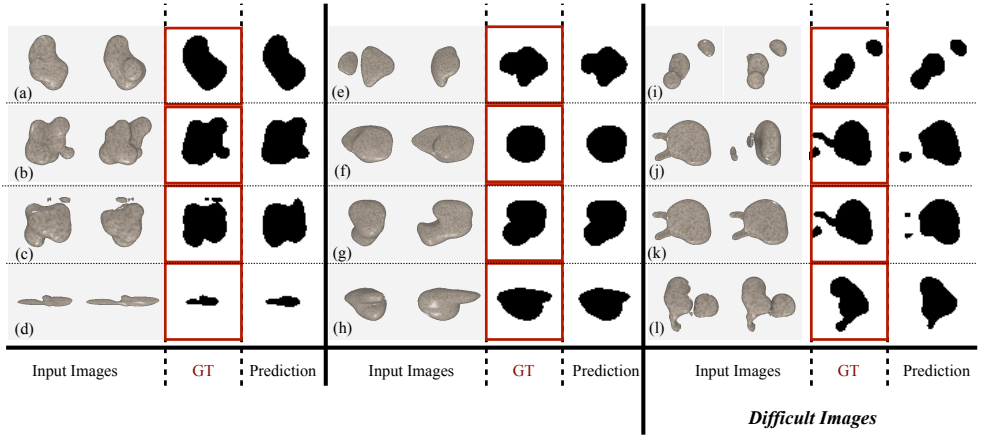


Figure 3: SilNet's predictions on the blobby object dataset. The input images are the left-hand images, the ground truth silhouettes are in the centre boxed in red and SilNet's predictions are to the right. The rightmost column (examples (i)-(l)) exhibits some of the challenging images in this dataset. These cases include: small floating bits which are hard to place accurately; extreme changes in viewpoint – *e.g.* the input images in example (j) and the change in angle between the output view and input views in example (l); and self occluded portions – *e.g.* in example (k) it is unclear from the original images whether the two protrusions are attached to the object or not.

C.2 Sculpture dataset.

Additional samples of SilNet$^{2D/3D}$'s results on the sculpture dataset are provided in figure 4. Additional reconstructions of SilNet^{3D} on the sculpture dataset are provided in figures 5,6. These sets of results demonstrate how increasing the number of input views at runtime improves SilNet's performance.

C.3 ShapeNet results.

Visual results on the ShapeNet database are given in figure 7. These results demonstrate again how increasing the number of input views at runtime improves SilNet's performance. Note that these images are of size 57×57, but we resize these images to 32×32 when comparing to Yan *et al*. [2].

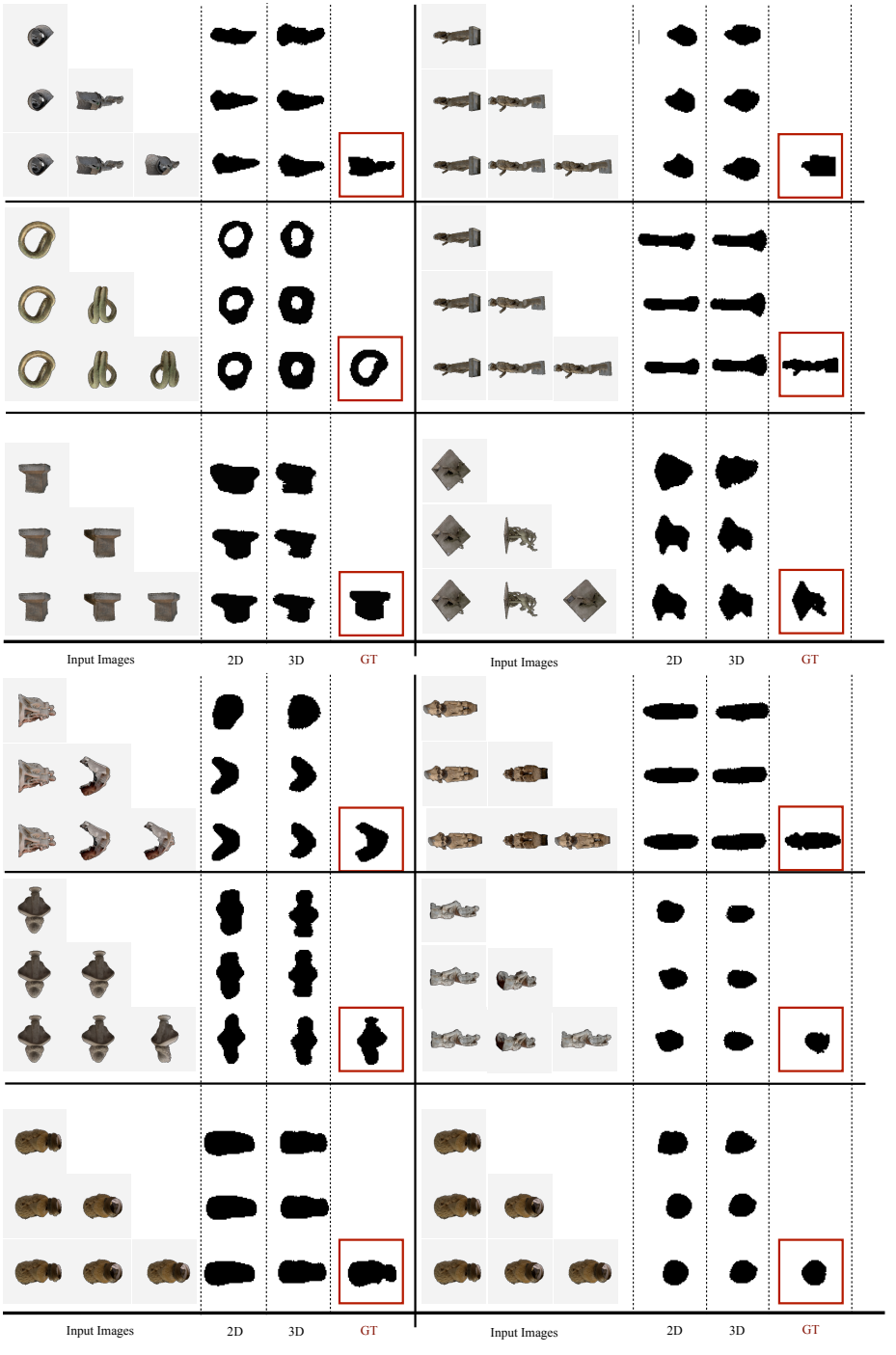
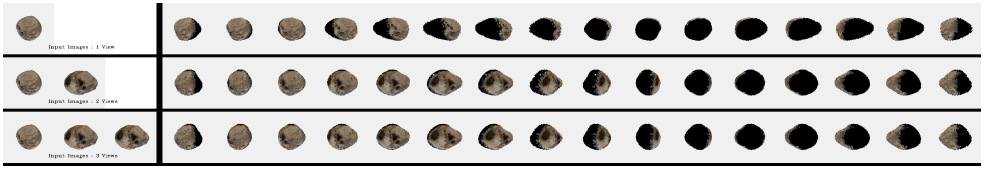
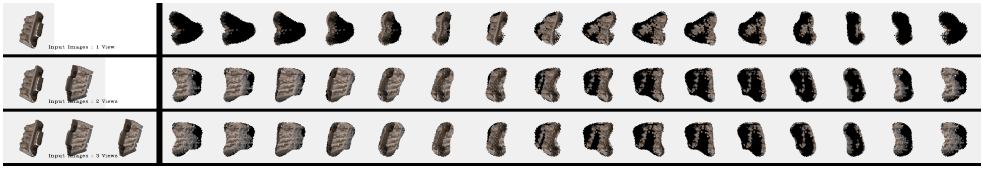


Figure 4: We exhibit some more examples of the results on the sculpture dataset for SilNet$^{2D/3D}$. The input images are the left-hand images, the predictions are in the centre and the ground truth silhouettes are the rightmost images boxed in red. Moving down the row corresponds to adding an additional input view, which improves SilNet's performance.

(a)



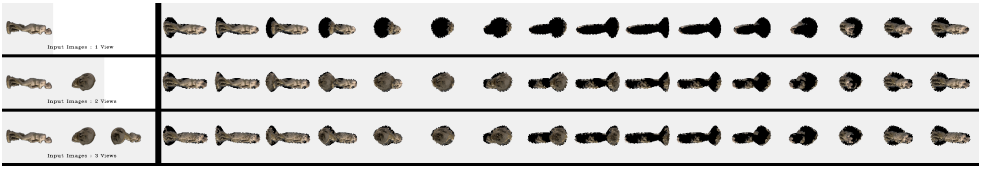
(b)



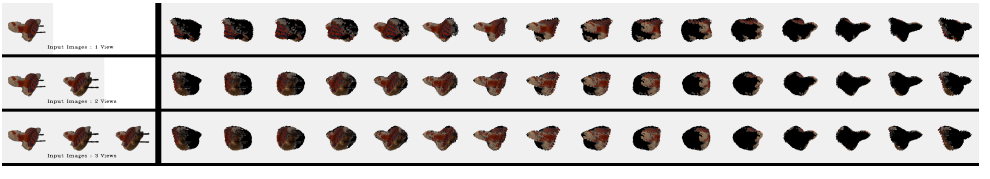
(c)

Figure 5: Some more examples of the results on the sculpture dataset for SilNet^{3D}. SilNet's generated 3D volume is rendered using view depending texturing [1]. The black pixels indicate that this part of the sculpture is not visible in the original images. The leftmost column of images contains the input images. The rightmost column contains SilNet's predictions for $\theta' \in [0°, 360°]$. The first six values are *interpolated* and the last ten values *extrapolated*, as we generate our dataset using θ' from the range $[0°, 120°]$. The number of input images increases from one to three down the rows. The interesting things to note are how SilNet interpolates and extrapolates the unseen angles based on the input images and how the reconstructions improve given more views. For example the pawn (a) tapers at the end in the first view but does not given two/three views. The skull (b) rotates the wrong way given one view but rotates correctly given two views. Zoom in for details.

(a)



(b)



(c)

Figure 6: Some more examples of the results on the sculpture dataset for SilNet^{3D}. Please refer to the caption of figure 5 for an explanation of this figure.

(a)

(b)

(c)

(d)

Figure 7: Performance of SilNet^{2D} on the ShapeNet chair test set. The input images (and angles) are kept constant while the angle corresponding to the output view θ' is rotated between $[0^\circ, 360^\circ]$. The input images are the coloured images to the left and each row corresponds to the addition of another view. These images demonstrate that SilNet gives visually compelling results and that increasing the number of views visually improves the results. For example, the silhouette predictions in (b) are more realistic given two images and those in (d) are more clearly delineated given two views.

References

[1] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image- based approach. In *Proc. ACM SIGGRAPH*, pages 11–20, 1996.

[2] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision. In *NIPS*, 2016.