Supplementary Material
Rethinking Reprojection: Closing the Loop for Pose-aware Shape Reconstruction from a Single Image

Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, Simon Lucey
The Robotics Institute, Carnegie Mellon University
{rz1, hamedk, chaoyanw}@andrew.cmu.edu, slucey@cs.cmu.edu

1. Appendix I.a. Network Architectures

We denote each fully-connected layer *fc(d)* by its output dimension d, and volumetric convolution layer by *conv3D(k, c, s)* representing kernel size of k, strides of s across three spatial axes, and c channels. 2D convolutional layer is represented as *conv2D(k, c, s)*. and the volumetric transpose convolution layer by *deconv3D(k, c, s)*.

Encoder and Generator for Aligned Shapes The variational aligned shape encoder takes as input an $30\times30\times30\times1$ tensor, and consists of 3 convolution layers: *conv3D(4, 32, 2)*, *conv3D(4, 64, 2)*, *conv3D(4, 64, 2)*; two fully-connected layers *fc(200)* and *fc(200)*, regressing from the last convolution feature to the 200-dimensional mean and variance vectors for style embedding, following [2]. The decoder takes in the 200-dimensional style vector, and consists of one fully-connected layer *fc(8192)* to connect the input vector to an $4\times4\times4\times128$ convolutional feature; and 3 transpose convolution layers *deconv3D(4, 64, 2)*, *deconv3D(4, 32, 2)*, *deconv3D(4, 1, 2)* output the reconstructed shape with size $30\times30\times30\times1$. All convolution and transpose convolution layer are batch batch normalized except the first convolution and last transpose convolution layer. LeakyReLU [3][1] is the rectifier for all layers except the output layer which uses *tanh*. This architecture is also used for 3D VAE in Section 4.2.

Image to Style/Pose Regressors The two regressors have identical architecture of convolution layers: *conv2D(11, 64, 4)*, *conv2D(5, 128, 2)*, *conv2D(5, 256, 2)*, *conv2D(5, 512, 2)*, *conv2D(3, 200, 2)*. For the style regressor, an *fc(200)* connects the last convolution layer to the style parameters. For the pose regressor, *fc(5)* is used instead. All but the first convolution layers are batch normalized, and rectified with LeakyReLU.

2. Appendix I.b. Training Details

p-TL Both the aligned shape autoencoder and the style/pose regressors are trained with Adam optimizer at an learning rate of 0.0003 and batch size of 100.

p-3D-VAE-GAN We follow [4] in training the 3D-VAE-GAN. We replace the L2 voxel-wise reconstruction loss described in [5] with the L2 loss between the last layer convolution features in the discriminator. We use RMSProp with a learning rate of 2e-5 and and batch size of 100 in training the 3D-VAE-GAN. The pose regressor is trained in the same routine as in **p-TL**.

3. Appendix II. Fine-tuning Details

In fine-tuning, we fine-tuned all the parameters in style/pose regressors, with an tiny learning rate of 1e-12. Each batch is natural images with silhouette annotations, mixed with rendered image-shape pairs. In this case, the loss is composed of two parts with both weight of 1: reprojection loss for natural images, and loss in shape for natural images (for p-VAE, this part is the euclidean loss in style and pose; for p-3D-VAE-GAN, this part is loss of VAE-GAN).

Fig. 1 gives an evaluation of the test 3D AP of pose-aware shapes over the ratio of natural images in a training batch. We may observe a relatively equivalent portion of rendered and natural samples in a fine-tuning batch returns the best AP, while too few natural images helps little in fine-tuning, and too many natural samples easily lead to over-fitting.

4. Appendix III. Reprojected Silhouettes as Instance Segmentation

In this section we showcase test results for MS COCO dataset where reprojected silhouettes from our pose-aware shape reconstruction could be used as instance segmentation, similar to the practice in [6]. We list 20 samples for each of the three categories: aeroplane, chair, car in Fig. 2 3 4, respectively.

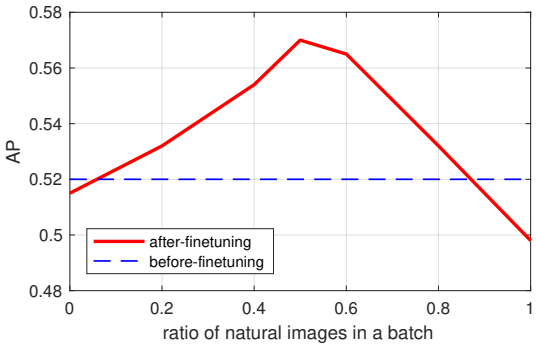

Figure 1: 3D AP as a function of ratio of natural images in a training batch, averaged over both approaches and all categories.

References

[1] B. Graham. Spatially-sparse convolutional neural networks. *arXiv preprint arXiv:1409.6070*, 2014. 1

[2] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[3] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013. 1

[4] T. Sainburg. Tensorflow Multi-GPU VAE-GAN implementation. https://github.com/timsainb/Tensorflow-MultiGPU-VAE-GAN/, 2016. [Online; accessed 2-January-2017]. 1

[5] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016. 1

[6] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

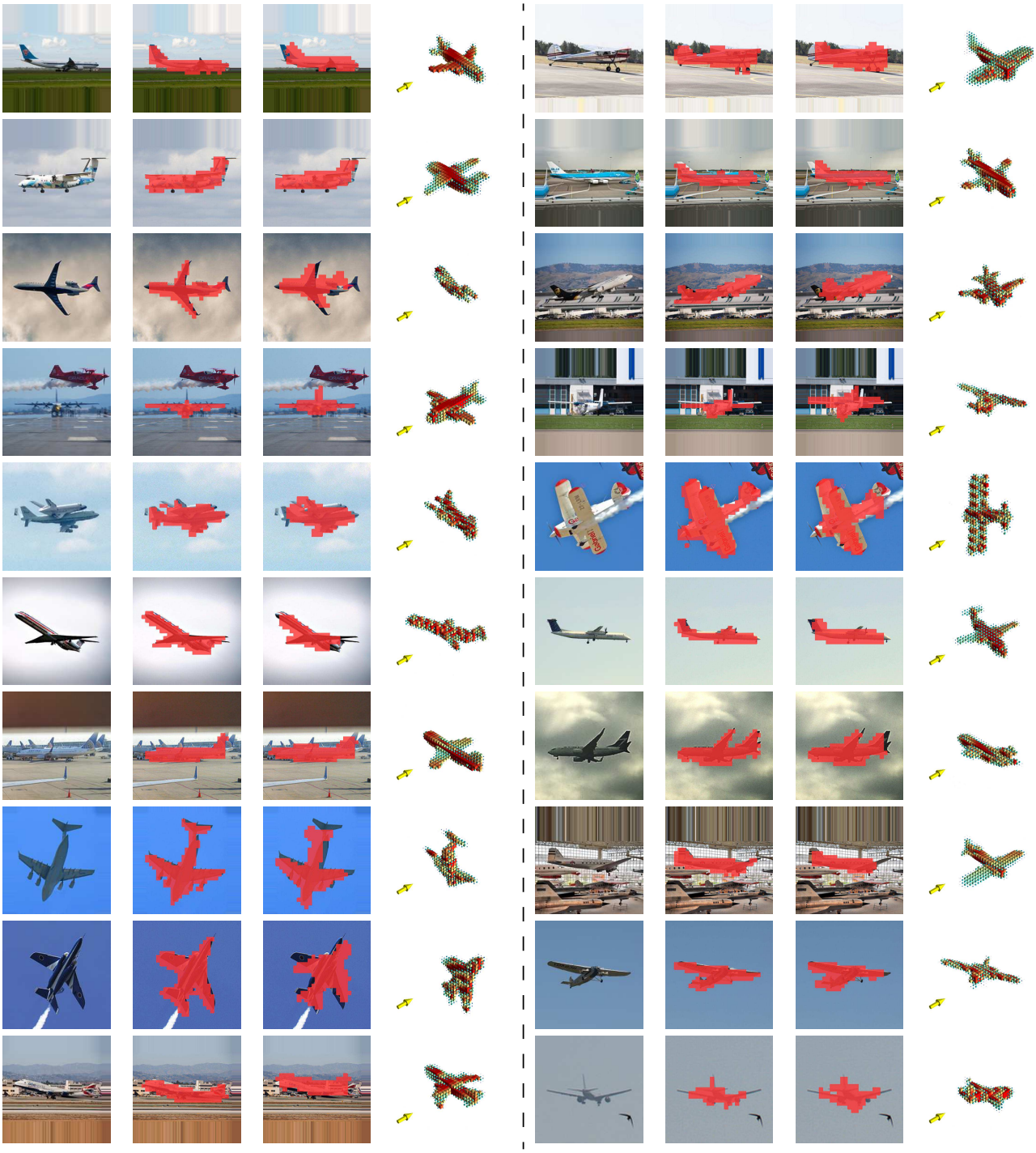

Figure 2: **Reprojected silhouette as instance segmentation (aeroplane)**. For each sample we show the input image (left), ground truth segmentation (middle left), reprojected silhouette(middle right), and reconstructed pose-aware shape(right).

Figure 3: **Reprojected silhouette as instance segmentation (chair)**. For each sample we show the input image (left), ground truth segmentation (middle left), reprojected silhouette(middle right), and reconstructed pose-aware shape(right).

Figure 4: **Reprojected silhouette as instance segmentation (car)**. For each sample we show the input image (left), ground truth segmentation (middle left), reprojected silhouette(middle right), and reconstructed pose-aware shape(right).