Fine-Grained Recognition without Part Annotations: Supplementary Material

Jonathan Krause[1] Hailin Jin[2] Jianchao Yang[2] Li Fei-Fei[1]
[1]Stanford University [2]Adobe Research
{jkrause,feifeili}@cs.stanford.edu {hljin,jiayang}@adobe.com

1. Network Architecture Comparison on cars-196

In the main text we showed that large gains from using a VGGNet [5] architecture on the CUB-2011 [6] dataset. We show a similar comparison on the cars-196 [3] dataset in Tab. 1. As before, using a VGGNet architecture leads to large gains. Particularly striking is the gain from fine-tuning a VGGNet on cars-196 – a basic R-CNN goes from 57.4% to 88.4% accuracy only by fine-tuning, much larger than the already sizeable gain from fine-tuning a CaffeNet [2].

Method	CNN Used	
	[2]	[5]
R-CNN [1]	51.0	57.4
R-CNN+ft	73.5	88.4
CNN+GT BBox	53.9	59.9
CNN+GT BBox+ft	75.4	89.0
PD+DCoP+flip	65.8	75.9
PD+DCoP+flip+ft	81.3	92.6
PD+DCoP+flip+GT BBox+ft	81.8	92.8

Table 1. Analysis of variations of our method on cars-196, comparing performance when using a CaffeNet [2] versus a CNN with a VGGNet architecture [5]. Performance is measured in 196-way accuracy.

2. Additional Visualizations

The visualizations in this section are expanded versions of figures from the main text.

2.1. Pose Nearest Neighbors

In Fig. 1 we show more examples of nearest neighbors using conv_4 features, which is our heuristic for measuring the difference in pose between different images (cf. Fig. 4 of the main text). In most cases the nearest neighbors of an image come from a variety of fine-grained classes and tend to have similar poses, justifying their use as a heuristic. In cases where there are potentially many instances with similar poses (*e.g.* first row, third column, or fifth row, first column), the nearest neighbors may share more than just pose. This heuristic still works reasonably when the pose is relatively unusual (third row, first column, and fourth row, third column), although occasionally small pose differences persist (direction of the head in the third row, third column).

2.2. Foreground Refinement

Additional examples of images where the foreground refinement (cf. Sec. 3.1 and Fig. 3 of the main text) changes the segmentation are given in Fig. 2. Most errors in a GrabCut[4]+class model which can be corrected by a foreground refinement are undersegmentations. In the most extreme case, these undersegmentations can actually be empty, which the foreground refinement fixes. In all cases the segmentation after refinement is better than the segmentation before refinement, though the final segmentation may still have imperfections.

2.3. Co-segmentation

We show additional qualitative co-segmentation results in Fig. 3 to supplement the results in Fig. 6 of the main text. In general, co-segmentation works quite well, but in cases where part of the background is sufficiently different from the rest of the background the segmentation quality can suffer. Segmentation is also difficult at certain car parts, *e.g.* the wheels, since they look very different from the rest of the car. It is also difficult to properly segment the bottom of many cars, since the shadow of the car often looks similar to the foreground.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014. 1

[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 1

[3] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision Workshops (ICCVW)*, pages 554–561. IEEE, 2013. 1

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

Figure 1. Additional visualizations for nearest neighbors with conv_4 features, which tend to preserve pose.

[4] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004. 1

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

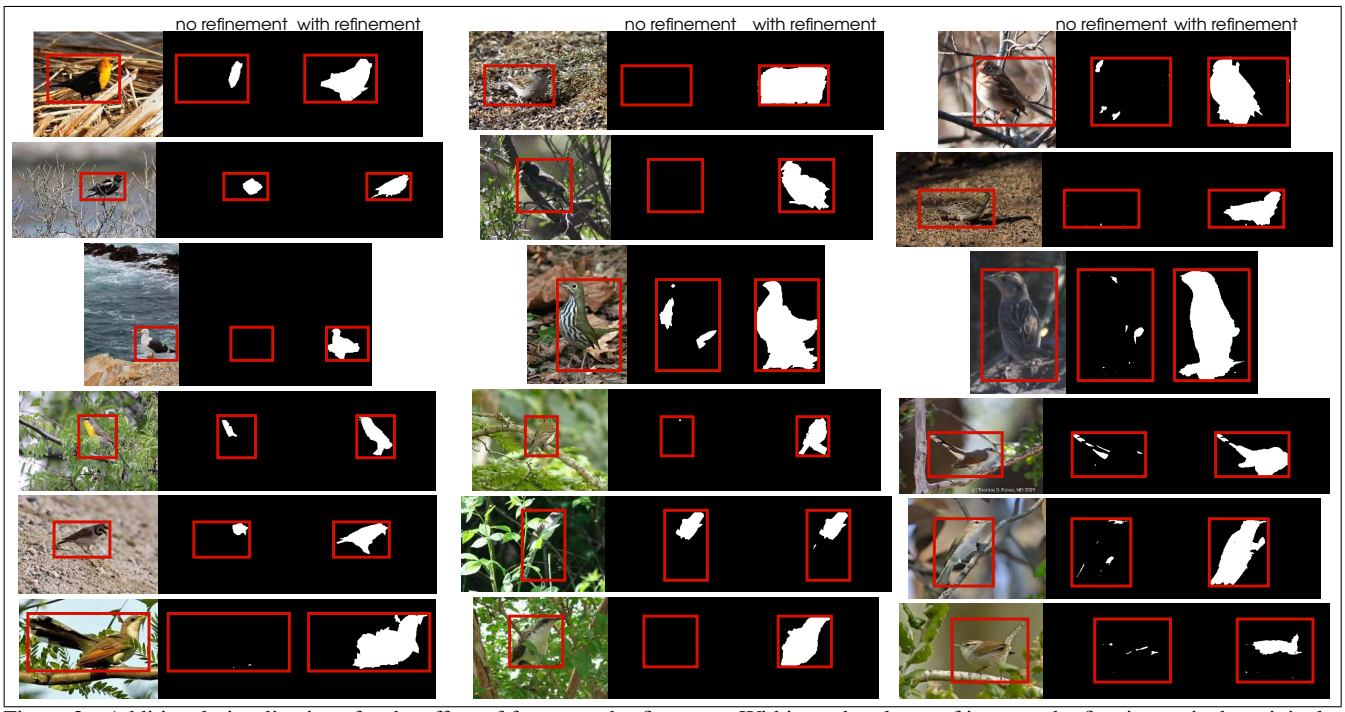

Figure 2. Additional visualizations for the effect of foreground refinement. Within each column of images, the first image is the original image, the second is the GrabCut+class model, and the third is GrabCut+class+refine.

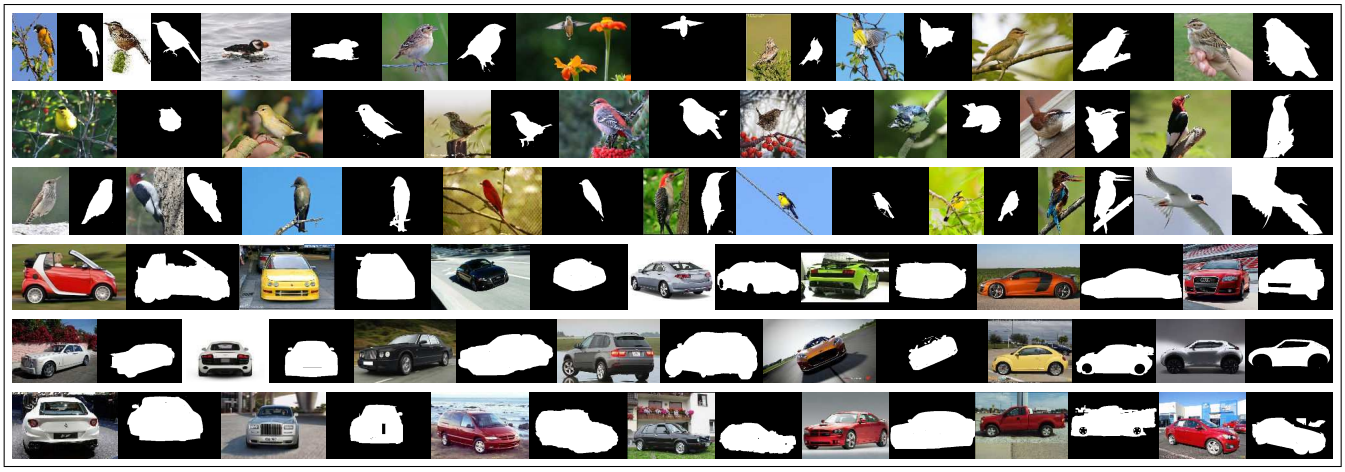

Figure 3. Additional visualizations of co-segmentation results. The last results in each row are failure cases.