SUPPLEMENTARY MATERIAL OF
SCOPS: Self-Supervised Co-Part Segmentation

Wei-Chih Hung[1], Varun Jampani[2], Sifei Liu[2], Pavlo Molchanov[2], Ming-Hsuan Yang[1], and Jan Kautz[2]

[1]UC Merced [2]NVIDIA

1. Introduction

In this document, we provide more implementation details, additional results and analysis of SCOPS.

2. Optimization Objectives

In the proposed method, we train the part segmentation network and the semantic part basis with several loss functions, including geometric concentration loss $\mathcal{L}_c$, equivariance loss $\mathcal{L}_{eqv}$ and semantic consistency loss $\mathcal{L}_{sc}$, with orthonomal constraint $\mathcal{L}_{ot}$. The final objective function is a linear combination of these loss functions,

$$\mathcal{L}_{all} = \lambda_c \mathcal{L}_c + \lambda_{eqv} \mathcal{L}_{eqv} + \lambda_{sc} \mathcal{L}_{sc} + \lambda_{ot} \mathcal{L}_{ot}. \tag{1}$$

In all our experiments, the weighting coefficients $(\lambda_c, \lambda_{eqv}, \lambda_{sc}, \lambda_{ot})$ are set to $(0.1, 10, 100, 0.1)$. We obtain these weights by conducting a coarse grid search on a subset of the CelebA dataset images [4].

3. Implementation Details of Equivariance Loss

For spatial transform, we apply random $\pm 60°$ rotation, $\pm 20\%$ shifting, $0.3x - 2x$ scaling, as well as TPS transform with 5x5 grid and $\pm 10\%$ shifting for each control point. For color transform, we apply jittering to brightness ($\pm 30\%$), contrast ($\pm 30\%$), saturation ($\pm 20\%$), and hue ($\pm 20\%$). In our experiments, we find that the performance gain is robust with respect to a wide range of perturbation parameters.

4. Additional Experimental Analysis

4.1. On Using Different Part Number K

We evaluate the effects of selected part number K on the unaligned CelebA dataset [4]. We select $K = 2, 4, 6, 8, 10$ and present the landmark estimation errors in Table 1. Note that with higher K the landmarks' center are more robust to face pose variations and thus lead to lower error rate. However, the performance seems to saturate after $K = 8$. The qualitative results in Figure 1 also indicate the similar trend. When observing the part segmentation results for $K = 8$ and $K = 10$, the additional parts seem to be not corresponding to any semantically meaningful area of the face.

Table 1: **Landmark evaluation on CelebA with different K.** Mean L2 distance.

SCOPS	K = 2	K = 4	K = 6	K = 8	K = 10
Error (%)	25.46	21.76	15.92	15.01	14.71

4.2. Feature Visualization of Learned Semantic Part Basis

We also show how the learned part basis improves on training progress in Figure 2. The top images show the part segmentation results while the bottom show the t-SNE visualization [5] on the ImageNet feature of pixels and their corresponding segmentation class i.e., the dot colors correspond to the part segmentation visualization colors, while the black dots represent the background pixels. During the training process, we can see that the part segmentations are improving, while pixels with similar ImageNet features are segmented as same part.

5. Quantitative Results on iCoseg

We present additional evaluation on the iCoseg dataset [1], which is a commonly used co-segmentation dataset. Follow DFF [2], we select 5 image sets and train SCOPS with $K = 4$ on the images. We aggregate the parts and evaluate with foreground segmentation IOU. The results show that SCOPS performs favorably against existing methods.

Table 2: **Evaluation on iCoseg.** Cosegmentation IoU comparing SCOPS to recent techniques on 5 image sets of iCoseg.

Subset	Elephants	Taj-Mahal	Pyramids	Gymnastics1	Statue of Liberty
Rubinstein [6]	63	48	57	**94**	**70**
DFF [2]	65	41	57	43	49
DFF-crf [2]	76	51	70	52	62
SCOPS	78.5	**67.0**	72.7	73.5	63.8
SCOPS-crf	**81.5**	66.0	**75.0**	81.5	65.7

6. Qualitative Results on PASCAL Objects

We present additional part segmentation visual results on different object class images obtained from the PASCAL dataset [3]. In Figures 3-9, the results show that the proposed method works for both rigid (Figures 3-6) and non-rigid (Figures 7-9) objects, where SCOPS produce part segments that are consistent across different instances with variations in appearance, pose and camera viewpoints.

7. Additional Results on Unaligned CelebA and CUB

We provide more visual results on the unaligned CelebA dataset [4] and the CUB dataset [8] in Figures 10-13. The results show that SCOPS is robust to pose and camera variations while having better boundary adherence compared to other techniques.

References

[1] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010.

[2] E. Collins, R. Achanta, and S. Süsstrunk. Deep feature factorization for concept discovery. In *ECCV*, 2018.

[3] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[4] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[5] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[6] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[7] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017.

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

[9] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018.

Figure 1: **Effects of changing part number.** We apply SCOPS on the unaligned CelebA dataset with the part number K ranging from 2 to 10.

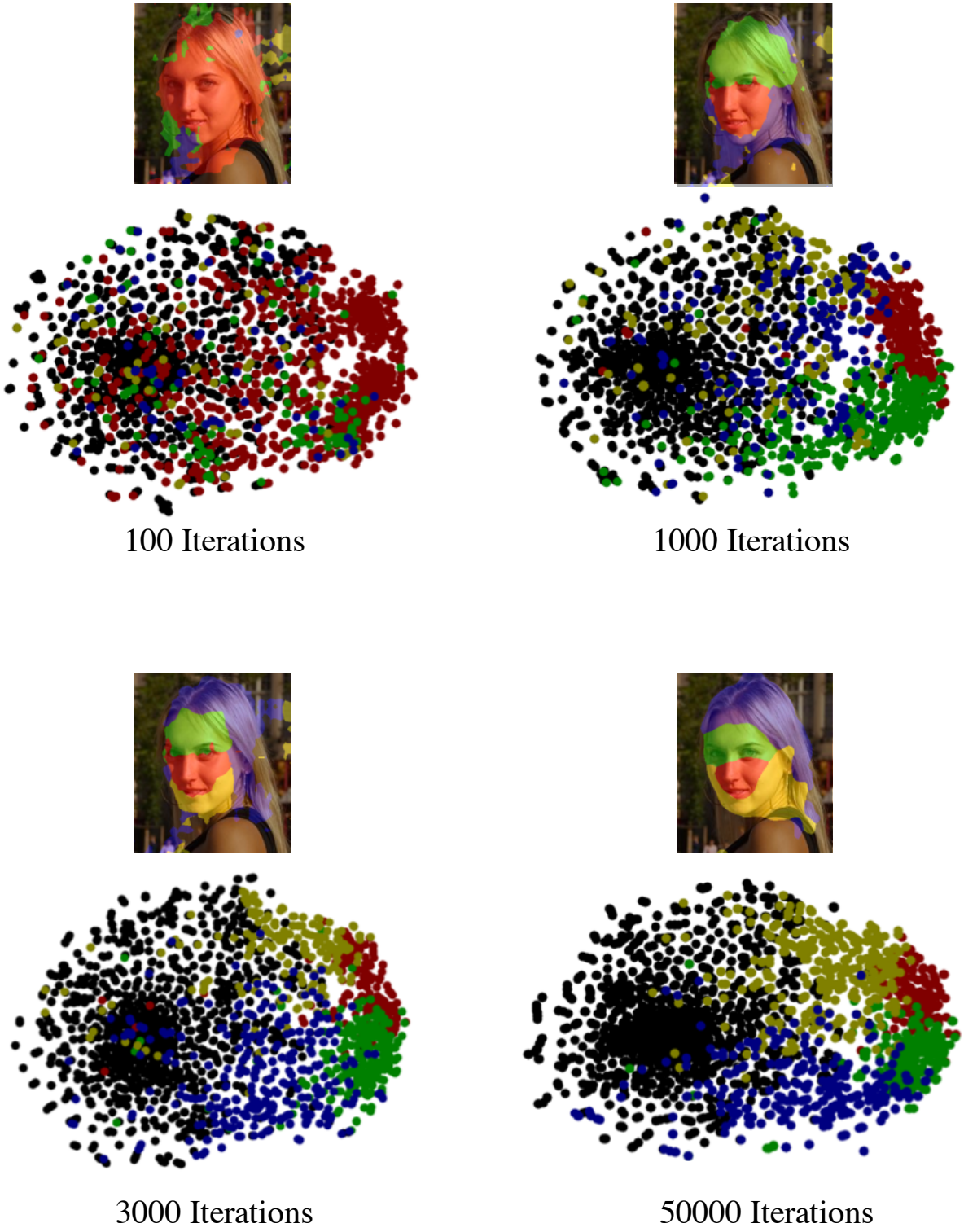

Figure 2: **Training progression.** The top images show the part segmentation results while the bottom show the TSNE visualization on the imagenet feature of pixels and their corresponding segmentation class. Black dots are the background pixels. During the training process, we can see that the part segmentations are improving, while pixels with similar imagenet features are segmented as same part.

Figure 3: SCOPS visual results on **car** class images in the PASCAL dataset.

Figure 4: SCOPS visual results on **bus** class images in the PASCAL dataset.

Figure 5: SCOPS visual results on **aeroplane** class images in the PASCAL dataset.

Figure 6: SCOPS visual results on **motor** class images in the PASCAL dataset.

Figure 7: SCOPS visual results on **sheep** class images in the PASCAL dataset.

Figure 8: SCOPS visual results on **horse** class images in the PASCAL dataset.

Figure 9: SCOPS visual results on **cow** class images in the PASCAL dataset.

Figure 10: **Additional visual results on CelebA face images.** SCOPS produce consistent part segments compared to existing techniques. Also shown is the effect of different loss constraints.

Figure 11: **Additional visual results on CelebA face images.** SCOPS produce consistent part segments compared to existing techniques. Also shown is the effect of different loss constraints.

Figure 12: **Additional visual results on CUB bird images.** SCOPS is robust to pose and camera variations while having better boundary adherence compared to other techniques.

Figure 13: **Additional visual results on CUB bird images.** SCOPS is robust to pose and camera variations while having better boundary adherence compared to other techniques.