Supplemental Materials for
Network-free, unsupervised semantic segmentation with synthetic images

Contents

1. Parameter Selection

The key parameters of the proposed algorithm are number of clusters k, style mixing cutoff c, foreground approach, and color space. Tab. S1 specifies the parameter values used in our experiments.

Data / generators	k	c	fg	color space
FFHQ (human faces)	3	8	saliency	LAB
AFHQ-wild	3	7	saliency	LAB
AFHQ-cat	2	5	corner	LAB
CelebAHQ (real, inverted)	3	8	saliency	LAB
LSUN-Horse	2	5	corner	LAB
DeepRooms	2	8	corner	RGB

Table S1. Parameter selection in our experiments. fg: foreground approach.

2. On Computational and Data Demands

Compared to other unsupervised segmentation algorithms, our method requires *no training setup* (*i.e.* training time = 0), does not need original training data (unlike L4F), and uses less/comparable memory during inference, although the inference speed is slightly lower than our competing methods. Tab. S2 provides comparative numbers.

Methods	data (#img)	training VRAM (GB)	training time (hr)	Inference VRAM (GB)	Inference speed (s/img)
DsetGAN	16	10*	3*	18*	0.2*
L4F	10k	13	0.32	7	0.6
SiS	15k+50	27	22	3	0.2
Ours	0	0	0	4.8	0.9

Table S2. Comparison on computational resources needed for each method during training and inference stage. DsetGAN: Dataset-GAN [3]. L4F: Labels4Free [1]. SiS: Semgnetation in Style [2]. *: with 512×512 due to OOM in 1024×1024.

3. Saliency map

Here we provide visualization on the pre-defined saliency map used in our experiment, Fig. S1. This map is applicable to human faces, animal faces, and other objects with a convex, blob shape. We recommend trying this first and modify its mean and covariance only where necessary.

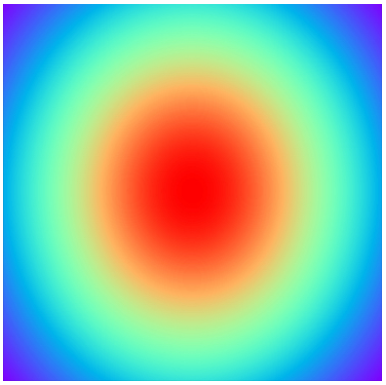

Figure S1. Gaussian saliency map used in our experiments. When used with bounding boxes, the map is resized to the shape of the bounding boxes.

4. Cluster Map Visualization

Fig. S2 shows cluster assignment maps $\mathbf{Y}'$ across different values of k. The clustering is performed on the entire image.

Figure S2. Raw cluster map $\mathbf{Y}'$ for $k = 2, 4, 10$ on FFHQ, AFHQ-Cat/Wild and Horse samples.

5. Additional Results

In this section, we provide additional quantitative results between our methods and baselines on DeepRooms

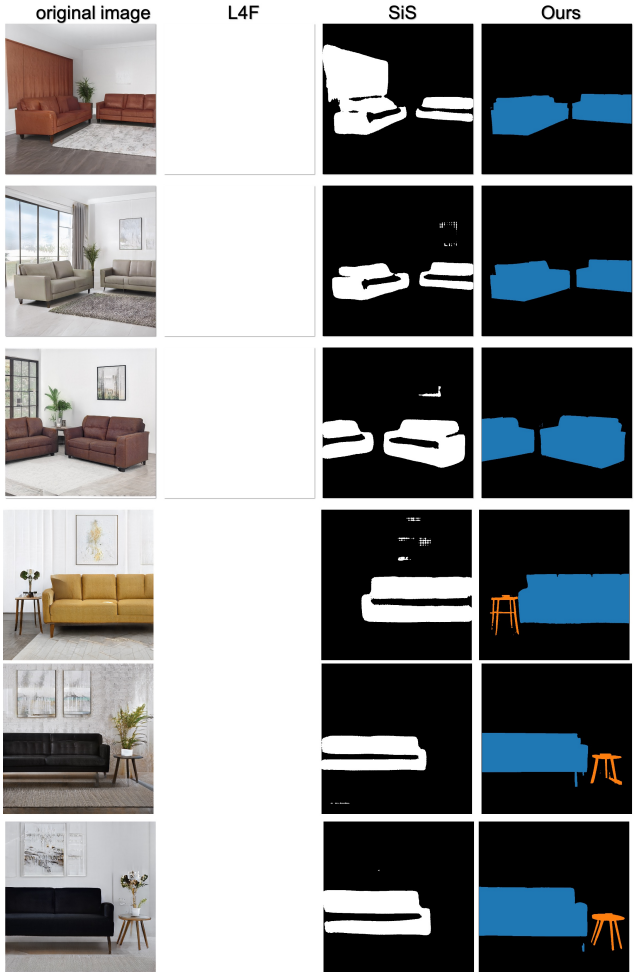

Figure S3. Segmentation comparison on DeepRoom images. L4F failed to produce meaningful mask even with our attempts to re-train and re-configure.

dataset	metric	L4F	SiS	Ours
AFHQ-cat	IOU (fg)	0.897	0.938	**0.944**
	IOU (bg)	0.691	0.844	**0.872**
	mIOU	0.794	0.891	**0.908**
DeepRooms	IOU(sofa-fg)	0.198	0.597	**0.880**
	IOU(sofa-bg)	0.000	0.901	**0.974**
	mIOU(sofa)	0.099	0.749	**0.927**
	IOU(table-fg)	0.007	0.011	**0.141**
	IOU(table-bg)	0.000	0.840	**0.963**
	mIOU(table)	0.003	0.426	**0.552**

Table S3. Additional quantitative results on AFHQ-Cat and DeepRooms dataset. For L4F and SiS, we re-trained their method on the DeepRooms generator and calculate the metric within the same detected bounding boxes used in our method. Note that L4F and SiS are not designed to handle complex rooms with no obvious foreground definition.

and AFHQ-Cat datasets.

As mentioned in the main paper, it is not straightforward to generalize foreground segmentation methods like L4F and SiS to complex scene datasets like DeepRooms as the concept of foreground is not well-defined in these scenes. This is especially the case for L4F as it assumes a set of background patterns that might not generalize to complex scenes. We further support this claim with quantitative and qualitative results below in Tab. S3 and Fig. S3.

6. More visuals

We now provide additional visual results on DeepRooms, animal faces, human faces, and horses. The baselines are L4F and SiS. As we showed in Fig. S3, both L4F and SiS struggles to synthesize masks that do not have a clear-cut foreground in a complex scenes, whereas our method can accurately segment sofas and coffee tables. Additionally, randomly selected synthetic images overlayed with masks for animal faces, human faces, and horses are provided in Figs. S4-S6.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13970–13979, October 2021. 1

[2] Daniil Pakhomov, Sanchit Hira, Narayani Wagle, Kemar E Green, and Nassir Navab. Segmentation in style: Unsupervised semantic image segmentation with stylegan and clip. *arXiv preprint arXiv:2107.12518*, 2021. 1

[3] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 1

Figure S4. Additional FFHQ results generated from seed 0 to 42 with truncation= 0.7. The number on the top-left is the seed index to reproduce the image given the StyleGAN2 pre-trained generator. In the cases of missing seeds, the foreground heuristic reports no identifiable or confident foreground.

Figure S5. Additional AFHQ-cat results generated from seed 0 to 40 with truncation= 0.7. The number on the top-left is the seed index to reproduce the image given the StyleGAN2 pre-trained generator. In the cases of missing seeds, the foreground heuristic reports no identifiable or confident foreground.

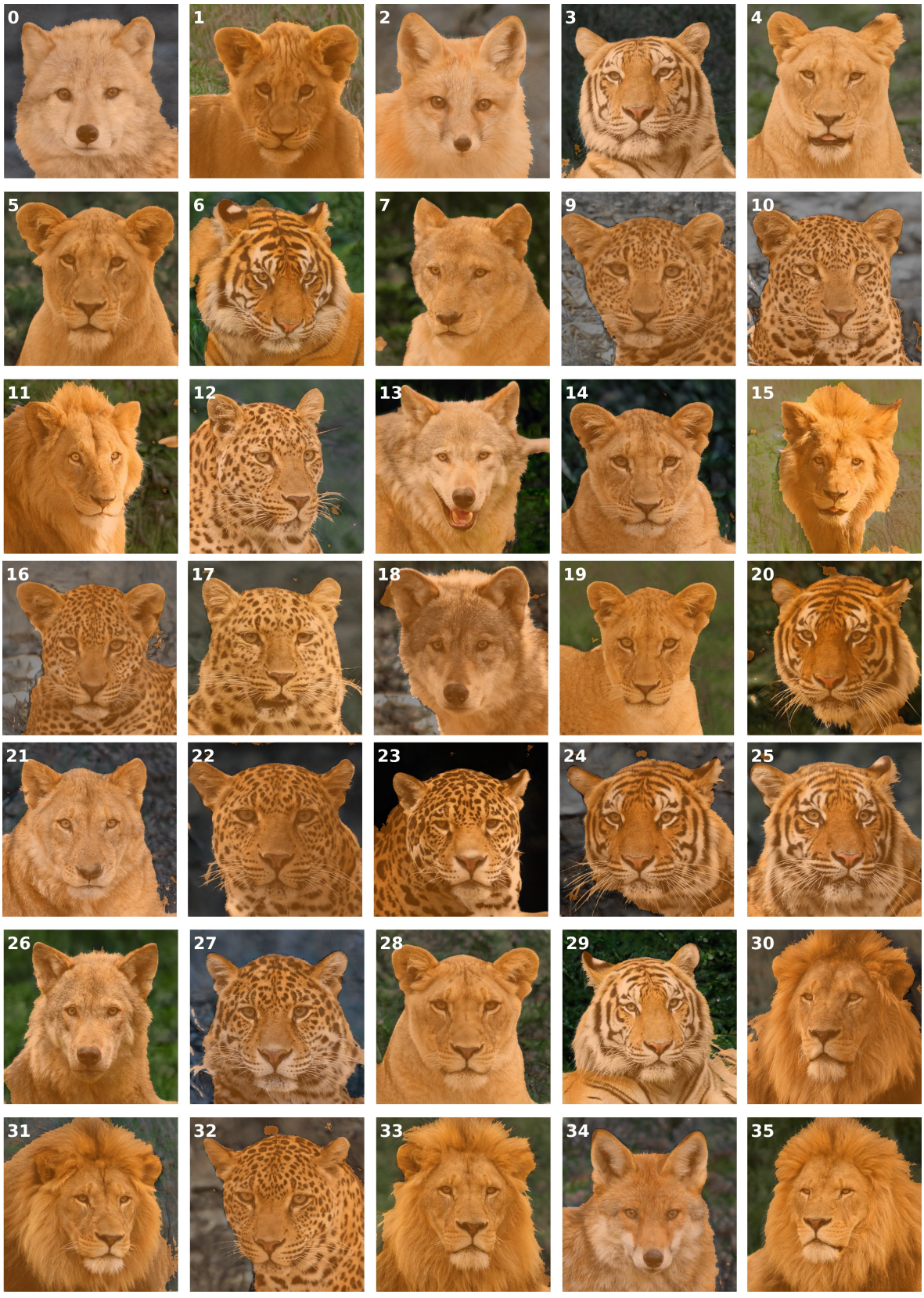

Figure S6. Additional AFHQ-wild results generated from seed 0 to 35 with truncation= 0.7. The number on the top-left is the seed index to reproduce the image given the StyleGAN2 pre-trained generator. In the cases of missing seeds, the foreground heuristic reports no identifiable or confident foreground.