Labels4Free: Unsupervised Segmentation using StyleGAN

Rameen Abdal[1] Peihao Zhu[1] Niloy J. Mitra[2] Peter Wonka[1]

[1]KAUST [2]UCL and Adobe Research

Figure 1: Complex segmentations using Labels4Free (zoom in for details).

1. Appendix A

1.1. Ablation Study

In order to validate the importance of the in-domain backgrounds for the training of the unsupervised network, we train our framework with the backgrounds taken from the MIT places [6] dataset. In order to do so, we replace the background generator with a random selection of an image from MIT places. During training, we not only train the alpha network but also the discriminator (as in the main method described in the paper). As discussed in the main paper, the discriminator is very good in identifying the out of the distribution images. In Table 1, we show the scores compared with BiSeNet and Detectron 2 (see Table 1 and Table 2 in the main paper) when using the MIT places for the backgrounds. Note that the scores decrease drastically.

As a second ablation study, we try to learn the Alpha mask only from features of the last layer before the output. This straightforward extension does not work well as seen in Table 2. In summary, using features from multiple layers in the generator is important to achieve higher fidelity.

2. Appendix B

2.1. Visualization of tRGB layers

In Fig. 3, 4 and 5, we visualize the tRGB layers at different resolutions as discussed in the experiment in Section 3.1 of the main paper. We select all the resolution tensors including and after the highlighted resolution. Here, we first normalize the tensors by $\frac{x-\min(x)}{\max(x)-\min(x)}$, where x represents the tRGB tensor at a given resolution. Notice that for the face visualization in Fig. 3, the face structure is clearly noticeable at the 32×32 resolution corresponding to the 4th layer of StyleGAN2. Other efforts in the StyleGAN-based local editing [1, 5] also selects early layers for the semantic manipulation of the images. These tests support that even features from earlier layers are beneficial for segmentation.

2.2. Complex background and multiple object segmentation

In Fig. 1, we show some more examples of segmentations possible with our method. Notice that the backgrounds in all the cases are complex and difficult to handle. Apart from the main object, there can be multiple instances of the same object or multiple objects in the scene in correlation

Table 1: Quantitative results of using the MIT places dataset for the backgrounds.

Dataset	mIOU	F1	Prec	Rec	Acc
FFHQ	0.34	0.40	0.34	0.50	0.68
LSUN-Horse	0.14	0.22	0.14	0.5	0.28
LSUN-Cat	0.20	0.28	0.69	0.50	0.39
LSUN-Car	0.34	0.51	0.63	0.63	0.51

Table 2: Quantitative results of using only the last layer of the StyleGAN2 for the construction of the Alpha Network.

Dataset	mIOU	F1	Prec	Rec	Acc
FFHQ	0.34	0.41	0.57	0.50	0.69
LSUN-Horse	0.33	0.42	0.41	0.44	0.63
LSUN-Cat	0.31	0.41	0.48	0.50	0.58
LSUN-Car	0.42	0.57	0.57	0.58	0.63

Figure 2: Multiple object segmentation.

with the foreground object. Fig. 2 shows that our method is able to handle such cases.

3. Appendix C

3.1. Custom dataset

We curated a custom dataset to evaluate the performance of our method with the supervised frameworks (BiSeNet and Detectron 2). We collected 10 images per class. These images were sampled by using the pretrained StyleGAN2 at different truncation levels and on different datasets i.e., FFHQ, LSUN-Horse, LSUN-Cat and LSUN-Car. Note that we collected a diverse set of images, *e.g.*, diverse poses, lighting, and background instances (see Fig. 6). The images were annotated using the LabelBox tool [4].

3.2. Comparison with ReDo [2] and Editing in Style [3] Methods

We further compare our method with ReDo [2] and Editing in Style [3]. In Table 4 and 5, we show the results of the ReDo method using the CelebA-Mask dataset and our custom dataset respectively (see the main paper). Notice that the ReDo method fails on the diverse LSUN-object datasets. We also show the results of Editing in Style (see Table 3) method on FFHQ trained StyleGAN with no truncation. We use 64×64 tensors of StyleGAN to predict the semantic regions. Clearly our results are better.

Method	IOU fg/bg	mIOU	F1	Prec	Rec	Acc
EditStyle	0.47/0.56	0.52	0.69	0.73	0.78	0.69
Ours	**0.75/0.89**	**0.82**	**0.90**	**0.92**	**0.89**	**0.92**

Table 3: Comparison with Editing in Style [3].

Table 4: Quantitative results of the ReDo method using the custom dataset.

Dataset	mIOU	F1	Prec	Rec	Acc
FFHQ	0.61	0.76	0.78	0.82	0.76
LSUN-Horse	0.39	0.56	0.61	0.60	0.57
LSUN-Cat	0.26	0.40	0.49	0.49	0.45
LSUN-Car	0.11	0.19	0.44	0.50	0.23

Table 5: Quantitative results of the ReDo method using the CelebA-Mask annotations.

Method	IOU fg/bg	mIOU	F1	Prec	Rec	Acc
ReDo	0.56/0.69	0.63	0.77	0.77	0.82	0.78

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 1

[2] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[3] Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. Editing in style: Uncovering the local semantics of gans, 2020. 2

[4] LabelBox. Labelbox tool. https://labelbox.com/. 2

[5] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. *arXiv preprint arXiv:2011.12799*, 2020. 1

[6] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene

Figure 3: Visualization of the tRGB layers in the StyleGAN2 trained on FFHQ dataset. Note that the maps produce a prominent face structure at 32×32 resolution corresponding to layer 4 of StyleGAN2.

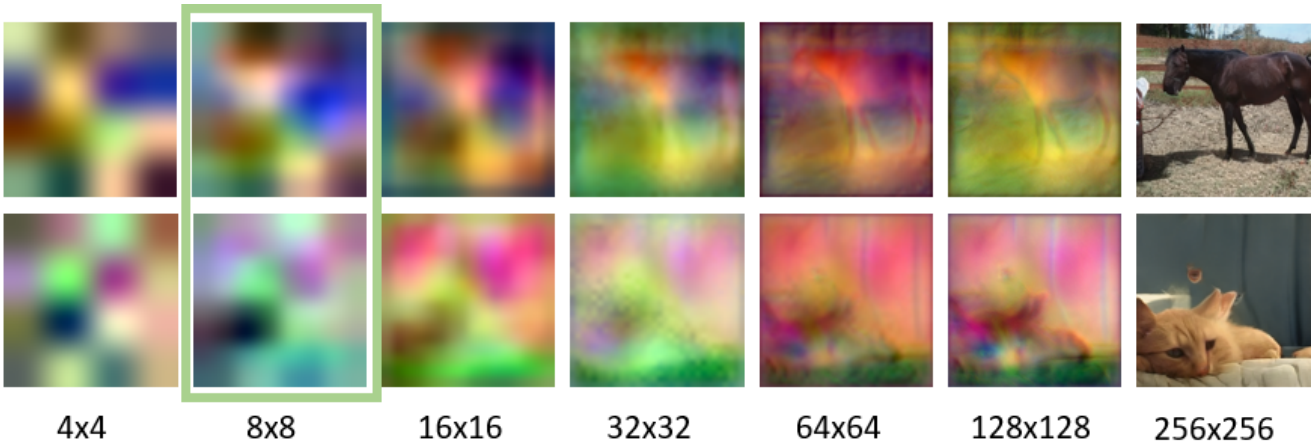

Figure 4: Visualization of the tRGB layers in the StyleGAN2 trained on LSUN-Horse and LSUN-Cat datasets.

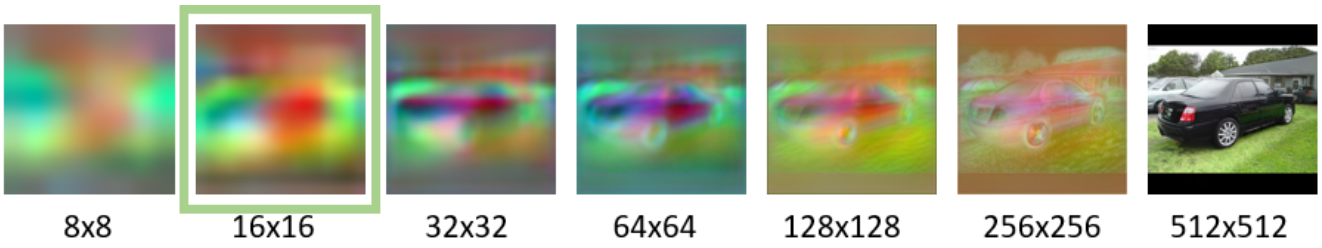

Figure 5: Visualization of the tRGB layers in the StyleGAN2 trained on LSUN-Car dataset.

recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1

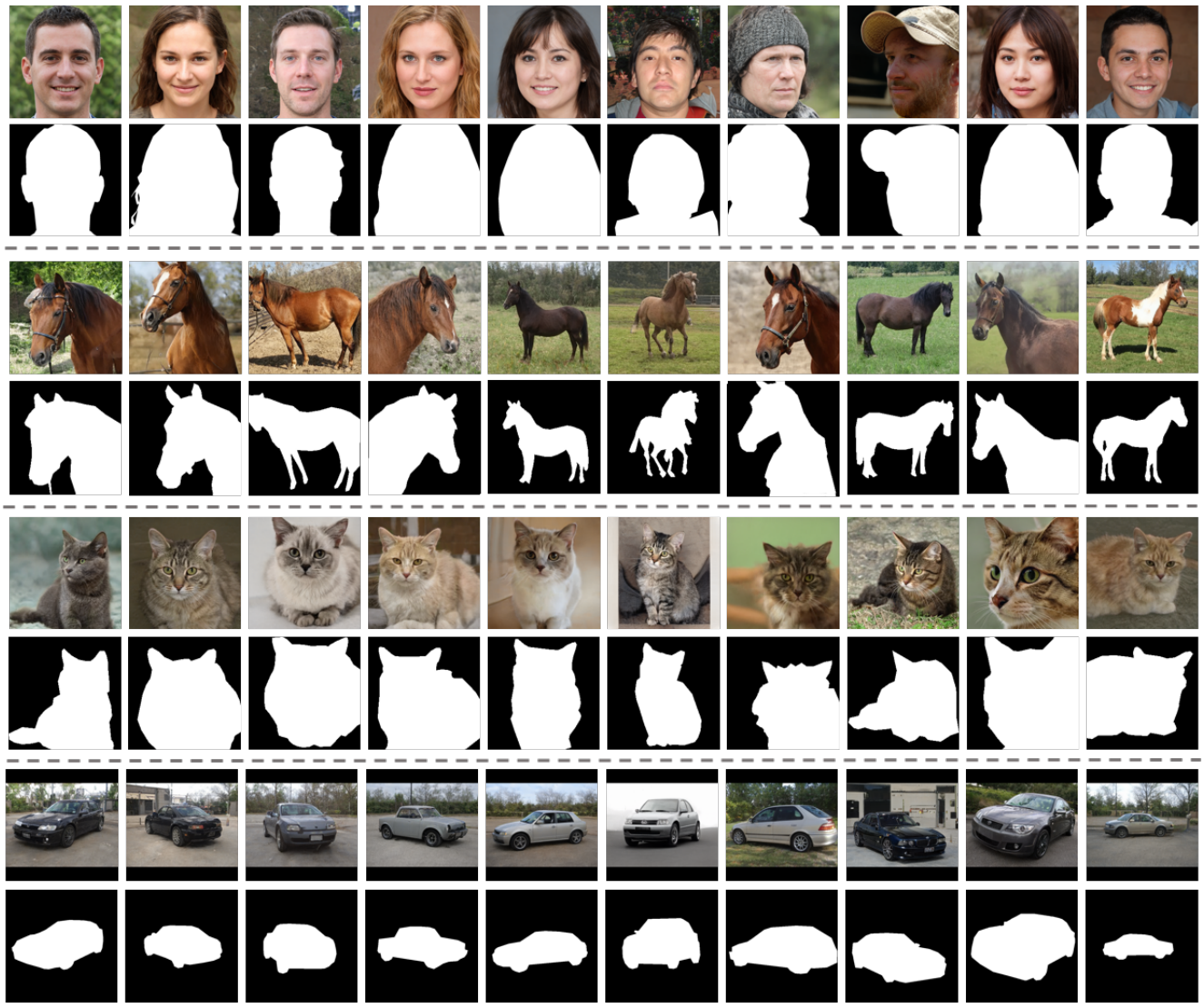

Figure 6: Our custom dataset with the corresponding ground truth labels.