Supplementary Materials for FurryGAN: High Quality Foreground-aware Image Synthesis

Anonymous ECCV submission

Paper ID 408

We provide the following supplementary materials:

A Uncurated visual comparison of Labels4Free and ours
B Choice of pseudo ground truth masks
C Quantitative results with alternative pseudo ground truth masks
D Examples of style mixing
E Additional details and visualization of the mask
F Results on unaligned datasets (LSUN-Church, LSUN-Horse, and CUB)
G User study on mask quality (between Lables4Free and ours)

A Uncurated comparison

Fig. S7-S8 (located at the end for clear spacing) present uncurated visual comparisons between Labels4Free and ours on FFHQ and AFHQv2-Cat. The columns represent foreground images, alpha masks, and composite images with generated backgrounds. While Labels4Free often misses clothes and whiskers, our method produces more accurate and detailed masks, especially on hair, fur, and whiskers. Consistency between the generated masks and the actual foreground region in the composite image also demonstrates the superiority of our method.

B Choice of pseudo ground truth masks

In this section, we provide the grounds for choosing TRACER (TE7) [19] to prepare pseudo ground truth masks over BiSeNet [38] (in Labels4Free [1]) and Mask R-CNN[1] (in PSeg [5]). As FFHQ does not have ground truth masks, we manually annotate ten images for the evaluation. The images are broadly chosen to cover various ages, genders, ethnic groups, and accessories. Fig. S1 shows the chosen images, annotated ground truths, and the pseudo ground truths from the methods. The quantitative comparison also reveals that TRACER achieves the best performance. Note that CelebAMask-HQ does not suffice to serve as the benchmark because BiSeNet is trained on it.

Fig. S2 further contrast the performance of the methods. On FFHQ, TRACER captures even hair while BiSeNet struggles. On AFHQv2-Cat, TRACER precisely captures even long fur on the ears and the top of the heads.

[1] https://github.com/facebookresearch/maskrcnn-benchmark

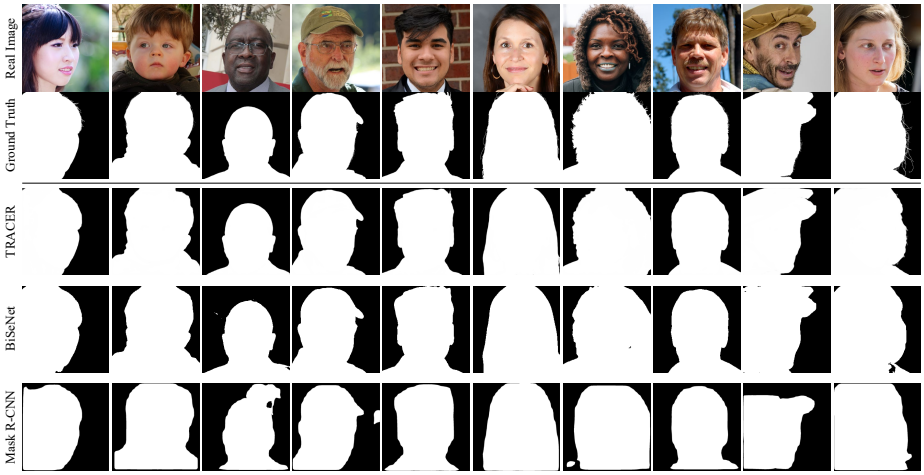


Fig. S1: **Qualitative comparison of masks.** We manually annotated ground truth masks (second row). TRACER produces masks very similar to the ground truth. BiSeNet also shows acceptable performance, but it often misclassifies the background as a foreground (3rd, 9th column) and vice versa (10th column). Mask R-CNN is relatively poor in quality, especially near the borders of the mask.

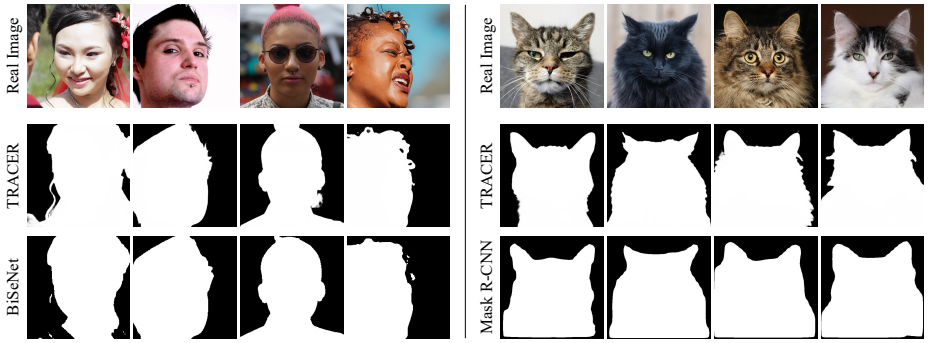
method	IoU(fg/bg)	mIoU	recall	precision	F1	Accuracy
Mask R-CNN	0.92/0.85	0.88	0.97	0.94	0.96	0.92
BiSeNet	0.98/0.96	0.97	0.99	**0.99**	**0.99**	0.98
TRACER	**0.99/0.97**	**0.98**	**1.00**	**0.99**	**0.99**	**0.99**

Table S1: **Quantitative comparison of predicted masks on the ten selected FFHQ images.** We evaluate the performance of the models with ten manually annotated ground truth masks.

C Quantitative evaluation with alternative pseudo ground truth masks

In this section, we report quantitative results with other choices of generating pseudo ground truth masks: BiSeNet for FFHQ and Mask R-CNN for AFHQv2-Cat following Labels4Free[2]. Table S2 confirms the same rankings as the ones with TRACER; our method consistently outperforms the competitors in all settings.

[2] Labels4Free uses Mask R-CNN for LSUN-Cat.



(a) Examples of predicted mask on FFHQ (b) Examples of predicted mask on AFHQv2-Cat

Fig. S2: **Further comparison of TRACER and other methods.** We evaluate each model on real images from FFHQ and AFHQv2-Cat datasets.

	ψ	method	IoU(fg/bg)	mIoU	recall	precision	F1	Accuracy
FFHQ (BiSeNet)	1.0	PSeg	0.05/0.24	0.14	0.05	0.16	0.07	0.05
		L4F	0.86/0.70	0.78	0.93	0.92	0.92	0.86
		Ours	**0.92/0.80**	**0.86**	**0.95**	**0.96**	**0.95**	**0.92**
	0.7	PSeg	0.01/0.23	0.12	0.01	0.04	0.01	0.01
		L4F	0.94/0.87	0.91	**0.96**	**0.99**	**0.97**	0.94
		Ours	**0.95/0.89**	**0.92**	**0.96**	**0.99**	**0.97**	**0.95**
AFHQv2-Cat (Mask R-CNN)	1.0	PSeg	0.06/0.21	0.13	0.06	0.17	0.07	0.06
		L4F	0.88/**0.72**	0.80	0.91	**0.97**	0.94	0.88
		Ours	**0.91/0.72**	**0.81**	**0.95**	0.95	**0.95**	**0.91**
	0.7	PSeg	0.01/0.17	0.09	0.01	0.13	0.01	0.01
		L4F	0.91/**0.77**	**0.84**	0.92	**0.98**	0.95	0.91
		Ours	**0.92/0.77**	**0.84**	**0.95**	0.96	**0.96**	**0.92**

Table S2: **Quantitative comparison of alpha masks on FFHQ and AFHQv2-Cat.** We use results of BiSeNet trained on CelebAMask-HQ as ground truth for FFHQ and results of Facebook's Detectron2 Mask R-CNN Model (R101-FPN) as ground truth for AFHQv2-Cat. We report the result with/without truncation trick (ψ=0.7, 1.0). The threshold for the alpha mask is 0.5 in ours and PSeg, and 0.9 in Labels4Free.

D Style mixing

Our generator supports style mixing since it is based on StyleGAN2. As coarse style affects shape in StyleGAN2, the masks of the coarse source determine the masks of the mixed results in our generator (Fig. S3). Note that we do not use mixing regularization during the training.

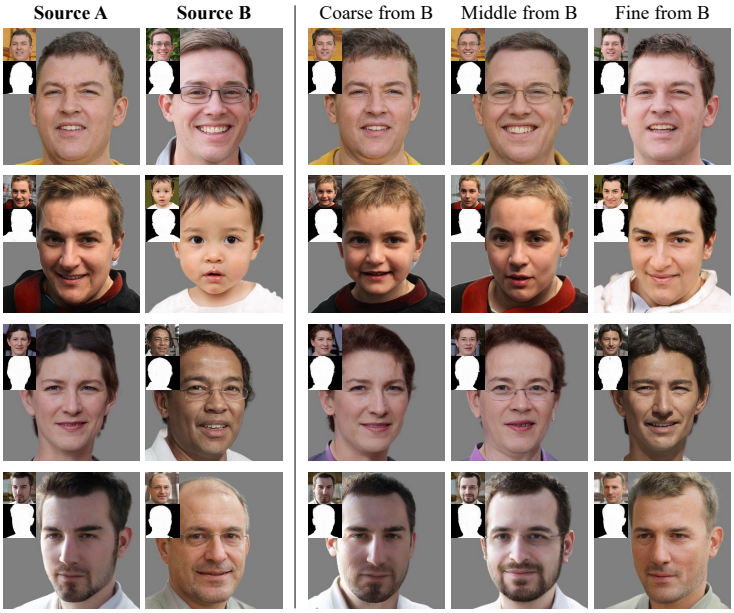
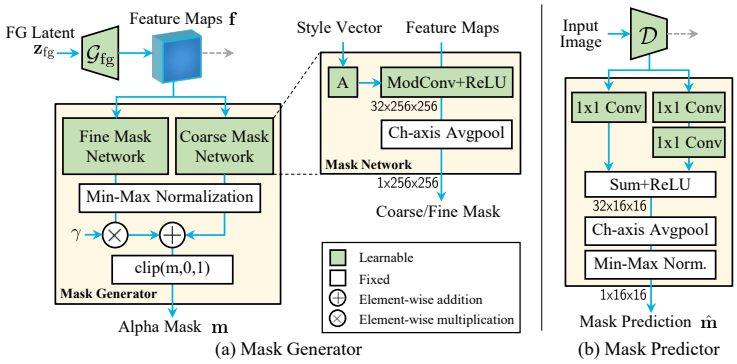


Fig. S3: The two leftmost columns are source images denoted by A and B. The right side of the figure is the result of using the latent code of B instead of the latent code of A in the coarse (4^2-8^2), middle (16^2-32^2), and fine (64^2-256^2) layers, respectively. We demonstrate masked foreground images to show the changes in the foreground mask according to different style mixing. In addition, we provide the composite image and mask in the upper left corner of each image.



(a) Mask Generator (b) Mask Predictor

Fig. S4: **Architecture of the mask generator and the mask predictor.** Coarse and fine mask networks use the same structure shown in the upper right corner of (a). γ is defined in Eq. (3). For brevity, we omit the LeakyReLU activation function between the convolution layers of the right branch in (b).

E Details about masks

In this section, we present the motivation for introducing fine masks and show additional mask visualizations. We assumed that the binarization loss (Eq. (6)) makes it difficult for the model to learn the matting-like details in the mask. These fine details are expected to occupy only a small part around the object boundary. Accordingly, we do not use binarization loss for the fine masks and use a very low threshold value for the inverse area loss (Eq. (9)).

We show some examples of coarse and fine masks in Fig. S5. As mentioned in Eq. (9), we penalize the area where the fine mask actually contributes to the final mask (the rightmost column of Fig. S5). Our generator can produce detailed alpha masks using the fine mask as needed. Finally, Fig. S4 illustrates the architectures of the mask generator and the mask predictor.

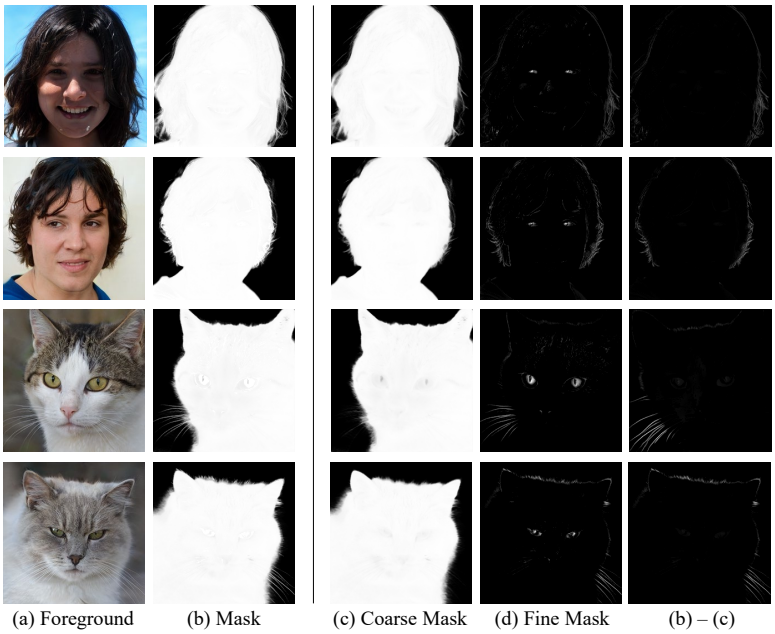
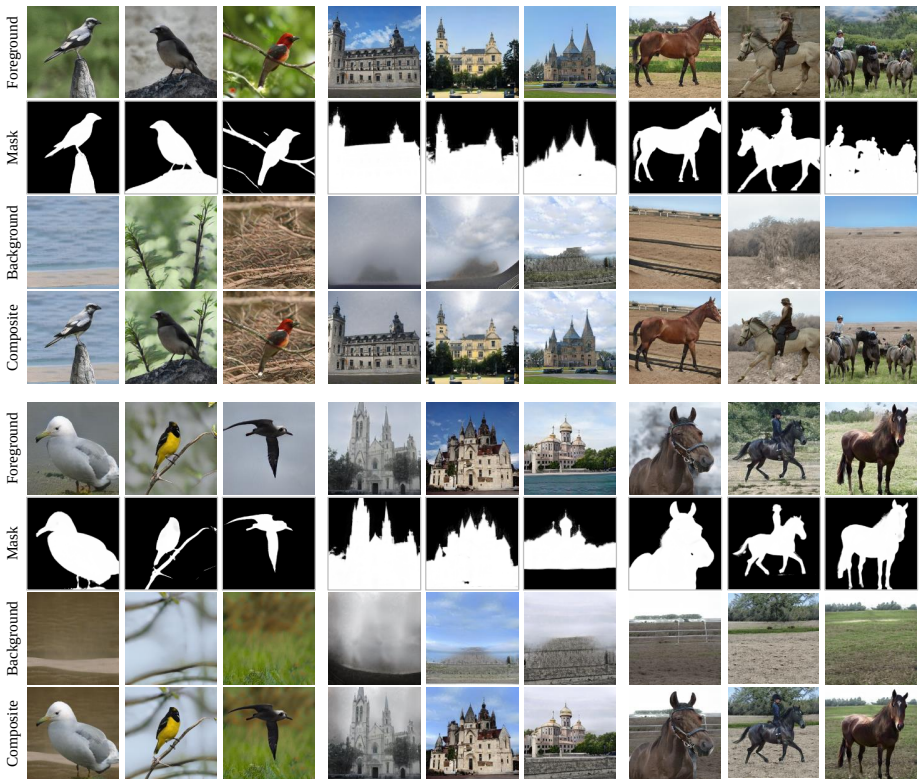


Fig. S5: **Visualization of coarse and fine masks.** We generate a final mask by summing up coarse and fine masks and then clipping it to the range in [0,1]. Due to the clipping operation, the area where the fine mask contributes to the final mask is the difference between the final mask and the coarse mask.

F Results on Unaligned Datasets

We also conducted training on unaligned datasets such as CUB and LSUN-Object. There are some changes in the training setting for this: 1) The coefficient of binarization loss is linearly reduced to 2.0 over the first 5K iterations. (default is 0.5). 2) We apply mask consistency loss after 5K iterations. 3) The average operation of the mask area loss is calculated for the mini-batch (not for each sample). 4) we set $\phi_1 = 0.2$ for LSUN-Object, and $\phi_1 = 0.1$ for CUB (Eq. (7)).

For LSUN-Object datasets, we use the first 100K images. We preprocess all datasets by center cropping and rescaling them to 256×256. Fig. S6 shows the results of selected samples for three unstructured datasets.



(a) CUB (b) LSUN-Church (c) LSUN-Horse

Fig. S6: Curated qualitative results on unaligned datasets.

G User Study on Mask Quality

To further evaluate the mask performance of our model, we asked 50 participants to choose more precise masks between ours and Labels4Free. In Table S3 (a), we report the results for ten random matches of generated image-mask pair used in Fig. S7-S8. In Table S3 (b), we report the results for the quality of masks obtained through the inversion of 20 real images (CelebAMask-HQ). For real image segmentation, both models were trained on FFHQ. Our model outperforms Labels4Free in mask quality of generated images and segmentation results of real images.

Table S3: The reported values mean the preference rate of mask outputs from ours against Labels4Free.

	(a) Generated		(b) Real
	AFHQv2-Cat	FFHQ	CelebA-HQ
Labels4Free	15.8%	11.2%	11.8%
Ours	84.2%	88.8%	88.2%

(a) Labels4Free (b) Ours

Fig. S7: Uncurated qualitative comparison of image composition results on FFHQ, with truncation setting $\psi = 0.7$.

(a) Labels4Free (b) Ours

Fig. S8: Uncurated qualitative comparison of image composition results on AFHQ, with truncation setting $\psi = 0.7$.