A Hybrid Model for Identity Obfuscation by Face Replacement
Supplementary Materials

Qianru Sun   Ayush Tewari   Weipeng Xu
Mario Fritz   Christian Theobalt   Bernt Schiele

Max Planck Institute for Informatics, Saarland Informatics Campus
{qsun, atewari, wxu, mfritz, theobalt, schiele}@mpi-inf.mpg.de

A Network architectures

In Fig. 1, we present the U-Net architecture for the Head Generator G, which corresponds to the Inpainter in Fig.1 of the main paper. Note that the output of the deep network is the image (256x256x3) including the body and the head. In the final layer, the output is cropped based on the head mask and pasted onto the obfuscated image (one of the inputs). Therefore, only the head region can provide any feedback during back-propagation. This follows from [1].

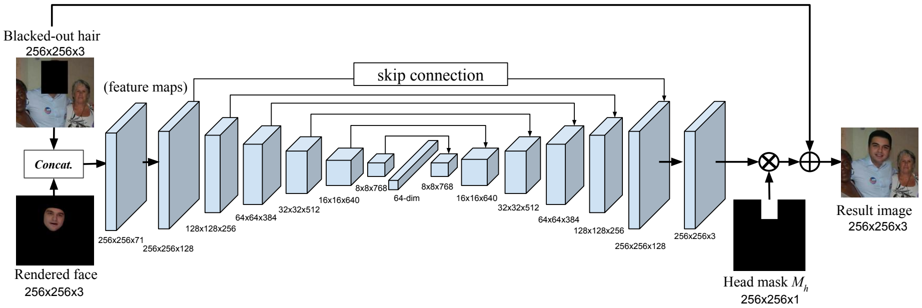

Fig. 1: The network architecture of Head Generator G used in the stage-II.

B Implementation details

For the Stage-I network (Section 3.1 in the main paper), AlexNet is used as the encoder ("Conv Encoder" in the Fig.1 of the main paper). We use AdaDelta [2] (200k iterations) to optimize the weights of the network with a batch size of 5 and a learning rate of 10^{-3}.

In the Inpainter (Section 3.2 in the main paper), the Head Generator is trained using the Adam optimizer [3] with $\lambda_H = 1000$ (in the main paper Eq.

(7)). Initial learning rates (for both generator G and discriminator D) are 2×10^{-5}, which decays to half every $5,000$ iterations. The batch size is set to 6; optimization stops after $10,000$ iterations; each iteration consists of 5 and 1 parameter updates for the generator and the discriminator respectively. It takes around 9 epochs for training the generator and around 2 epochs for training the discriminator.

To prepare a 256×256 body crop (Section 5.1 in the main paper), keeping the ratio of the head (width/height) unchanged, we first resize the original image such that the head height is 1/4 of 256. Then, we crop a 3 head weight × 4 head height region from the input image, making the head lying in the upper middle region of the crop. We zero-pad the image if its dimensions are smaller than the crop size, thereby obtaining the final crop with the desired 256×256 size.

C Obfuscation performance against AlexNet

In the experiments provided in the main paper, we focused on the obfuscation performance using a GoogleNet-based recognizer. However, as we have mentioned, our approach is *target-generic*: it is expected to work against a generic system.

Therefore, in this section, we additionally present the obfuscation performance with respect to an AlexNet-based recognizer. Following the same "feature extraction - SVM prediction" framework as in the main paper, we replace the feature extractor with AlexNet. Table 1 shows the quantitative comparisons between GoogleNet and AlexNet recognizers on different versions of our approach. Note that in this table, **v1**$\sim$**12** are the same representative versions as shown in Table 1 of the main paper. All other versions are also added here. Note that **v13** and **v14** are indexing two hybrid models, different with the GAN models for stage-I ablation study in the main paper.

Some recognition rate differences exit between the two recognizers. First of all, on original (ground truth) images, AlexNet performs worse than GoogleNet ($81.6\% < 85.6\%$). On images generated by our method (**v1**$\sim$**21**), AlexNet performs similarly when using head features, achieving a higher recognition rate for 12 input modalities out of 21, compared to GoogleNet. However, while using head+body features, GoogleNet recognition rates are higher for 18 different input modalities. The possible reason could be that the 1024-dimensional GoogleNet features are more compact than the AlexNet features, which are 4096-dimensional. From the discriminative head images, less compact features can extract more information in the additional feature dimensions. On the other hand, concatenation of features from the noisy body images could reduce the final recognition rates.

D Visualization results

In this section, we show visualization results using different modalities (**v2** to **v21**), corresponding to Table 1.

Respectively in Figure 2, Figure 3 and Figure 4, we show results with rendered faces from Original images, Blurred face images and Blacked-out face images. Note that, the results are consistently cropped to have small zero-padded regions. In most cases, the best visual quality is achieved in the second column which uses Original hair images. The largest visual differences compared to the original faces are visible in the last column when the rendered faces are replaced and the hair regions are entirely obfuscated by blacking-out.

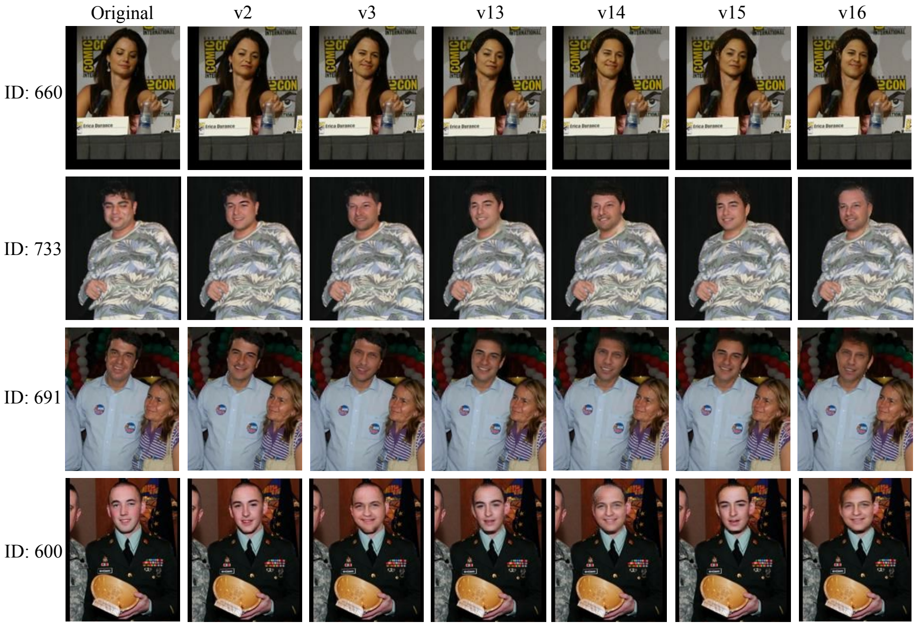

Fig. 2: Result images of methods **v2** to **v15** in the block named “Original in the Stage-I” in Table 1, compared to original images.

References

1. Sun, Q., Ma, L., Oh, S.J., Gool, L.V., Schiele, B., Fritz, M.: Natural and effective obfuscation by head inpainting. In: CVPR. (2018)
2. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. arXiv **1212.5701** (2012)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv **1412.6980** (2014)

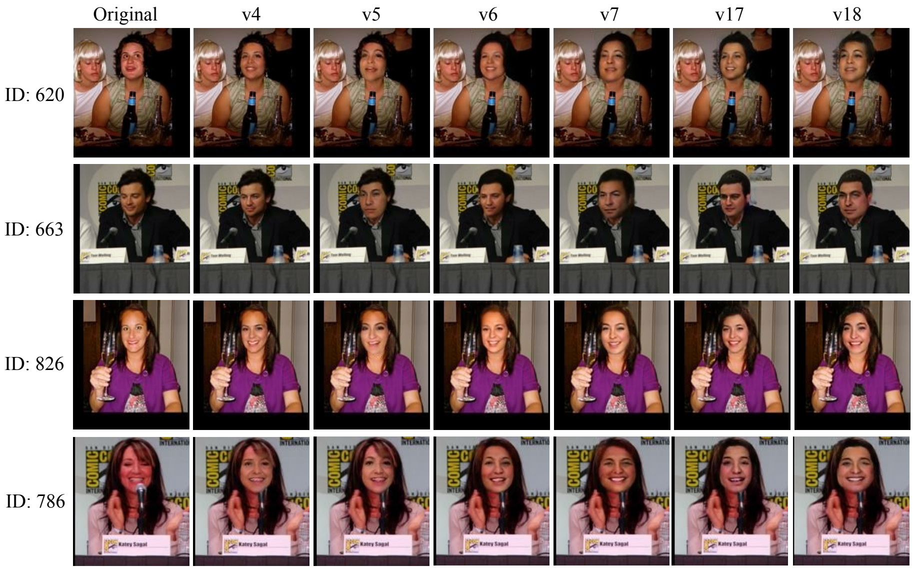

Fig. 3: Result images of methods **v4** to **v18** in the block named "Blurred in the Stage-I" in Table 1, compared to original images.

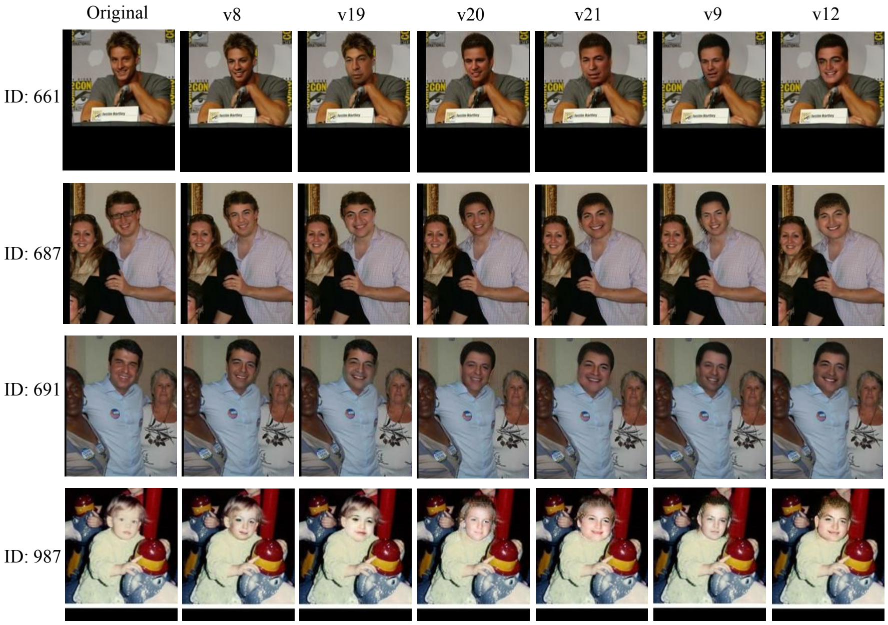

Fig. 4: Result images of methods **v8** to **v12** in the block named "Blacked in the Stage-I" in Table 1, compared to original images.

Table 1: Quantitative results comparing with the state-of-the-art methods [1]. Image quality: Mask-SSIM, SSIM and HPS scores (both the higher, the better). Obfuscation effectiveness: recognition rates of machine recognizers (lower is better). **v*** simply represents the method in that row, noting that supplementary methods are numbered after **v12** according to the Table 1 of our main paper. To save space, we use some abbreviations of input data as Rendered.=Rendered Face, Orig.=Original, Blu.=Blurred, Bla.=Blacked and Overlay-No-Inp.=Overlay-No-Inpainting, while full names were used in the Table 1 of our main paper.

Obfuscation method			Evaluation					
	Stage-II		Image quality		Google Net		Alex Net	
Stage-I	Hair	Rendered.	Mask-SSIM	SSIM	`head`	`body+head`	`head`	`body+head`
Orig.	-	-	1.00	1.00	85.6%	88.3%	81.6%	85.3%
[1]	Blu.+Detect		0.68	0.96	43.7%	51.7%	49.0%	48.9%
[1]	Blu.+PDMDec.		0.59	0.95	37.9%	49.1%	45.1%	45.6%
[1]	Bla.+Detect		0.41	0.90	10.1%	21.4%	11.4%	20.5%
[1]	Bla.+PDMDec.		0.20	0.86	5.6%	17.4%	7.4%	16.6%
[1], our crop	Blu.+Detect		0.64	0.98	40.5%	47.8%	43.6%	43.2%
[1], our crop	Blu.+PDMDec.		0.47	0.97	30.6%	38.6%	35.4%	37.0%
[1], our crop	Bla.+Detect		0.43	0.97	12.7%	24.0%	15.1%	23.4%
[1], our crop	Bla.+PDMDec.		0.23	0.96	9.7%	19.7%	10.5%	19.2%
v1, Orig.	Overlay-No-Inp.		0.75	0.96	66.9%	68.9%	64.0%	54.9%
v2, Orig.	Orig.	`Own`	0.87	0.99	70.8%	71.5%	66.6%	58.3%
v3, Orig.	Orig.	`Replacer15`	-	-	47.6%	57.4%	45.1%	47.9%
v13, Orig.	Blu.	`Own`	0.58	0.98	36.6%	48.2%	42.4%	43.2%
v14, Orig.	Blu.	`Replacer15`	-	-	18.0%	30.8%	22.9%	30.2%
v15, Orig.	Bla.	`Own`	0.50	0.97	22.5%	35.5%	30.3%	33.9%
v16, Orig.	Bla.	`Replacer15`	-	-	7.1%	21.3%	13.2%	21.5%
v4, Blu.	Orig.	`Own`	0.86	0.99	59.9%	65.2%	57.8%	52.0%
v5, Blu.	Orig.	`Replacer15`	-	-	26.3%	41.7%	24.1%	31.8%
v6, Blu.	Blu.	`Own`	0.55	0.98	25.8%	38.0%	28.2%	33.5%
v7, Blu.	Blu.	`Replacer15`	-	-	12.7%	29.3%	14.8%	23.3%
v17, Blu.	Bla.	`Own`	0.44	0.97	15.7%	28.2%	19.7%	25.7%
v18, Blu.	Bla.	`Replacer15`	-	-	7.2%	20.7%	9.9%	18.9%
v8, Bla.	Orig.	`Own`	0.85	0.99	59.3%	64.4%	54.8%	49.1%
v19, Bla.	Orig.	`Replacer15`	-	-	27.0%	41.4%	25.0%	31.3%
v20, Bla.	Blu.	`Own`	0.53	0.98	28.1%	38.6%	31.0%	34.7%
v21, Bla.	Blu.	`Replacer15`	-	-	11.2%	25.9%	14.7%	22.1%
v9, Bla.	Bla.	`Own`	0.47	0.97	14.2%	25.7%	19.1%	24.4%
v12, Bla.	Bla.	`Replacer15`	-	-	**7.1%**	**18.1%**	**9.7%**	**16.5%**