Noisy Adversarial Representation Learning for Effective and Efficient Image Obfuscation

Jonghu Jeong[*1] **Minyong Cho**[*1] **Philipp Benz**[1] **Tae-hoon Kim**[1]

[1]Deeping Source Inc., Seoul, Republic of Korea

A COMPLEX TASKS OTHER THAN BINARY CLASSIFICATION

A.1 MULTI-CLASS CLASSIFICATION

While previous experiments are conducted with at least one of the two tasks (privacy and utility) being binary classification, we show that our method can also be applied to more complex task settings, such as multi-class classification, for both the privacy and utility task. The FairFace [Karkkainen and Joo, 2021] dataset provides "Age" and "Race" classification tasks that consists of 9 classes ("0 to 2", "3 to 9", "10 to 19", ... ,"more than 70") and 7 classes ("Black", "East Asian", "Indian", "Latino Hispanic", "Middle Eastern", "Southeast Asian", and "White"), respectively. Table 1 shows the results using "Age" and "Race" as the privacy and the utility task, respectively. Table 1 also shows the results of a controlled study by switching the privacy and the utility tasks. We conducted this experiment to investigate how different combinations of tasks affect the privacy-utility trade-off.

MaxEnt [Roy and Boddeti, 2019] has the highest utility accuracy for both experiments compared to the other methods, excluding the upper bound. The result is in line with the results shown in the main paper with the other datasets. DeepObfs. [Li et al., 2021] suffers from a low privacy-utility trade-off which is consistent with its previous results. DISCO [Singh et al., 2021] also has a hard time defending against the privacy leakage attack, contrary to its performance for the binary classification tasks. We provide results for our method using $\sigma = 960$ for RN18_3, and $\sigma = 15$ for RN18_4, respectively. We note that ours with RN18_4 has the highest Δ among all methods, reaffirming that our proposed noise module effectively facilitates the obfuscator to learn privacy-preserving representations. RN18_3 also shows a comparable Δ even with lower computational cost and memory usage. The result confirms that our method also applies to more complex tasks other than simple tasks such as binary classification while reducing the resource on the client side.

Figure 1 shows reconstruction attack results on the representations generated by obfuscators trained with the privacy task "Race" and utility task "Age". DeepObfs. and DISCO failed to defend against reconstruction attacks revealing identity and race. MaxEnt successfully removes identity and race while retaining age. Our methods, both RN18_3 and RN18_4, successfully defend against the reconstruction attack, making it difficult for the adversary to identify both age and race on the reconstructed images. This result is consistent with the reconstruction attack on CelebA [Liu et al., 2015] in the main paper.

A.2 FACIAL LANDMARK DETECTION

In addition to classification, we applied our method to a regression task, facial landmark detection (FLD). The CelebA [Liu et al., 2015] dataset provides image coordinates of 5 facial landmarks (left eye, right eye, nose, left mouth corner, and right mouth corner). We set facial landmark detection and gender classification as the utility and privacy tasks. We compare our approach, RN18_3 using $\sigma = 1920$, with the practical upper bound model (RN18) and MaxEnt, showing the best

[*]Equal Contribution.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Table 1: Results with multi-class classification. The practical upper bounds (RN18) of each privacy and utility task are reported by training ResNet18 [He et al., 2016] with original images, respectively.

Method	FairFace (Race / Age)			FairFace (Age / Race)		
	Privacy ↓	Utility ↑	Δ ↑	Privacy ↓	Utility ↑	Δ ↑
RN18	63.57	55.49	-	55.49	63.57	-
MaxEnt	23.40	53.82	30.42	30.82	63.27	32.45
DeepObfs.	53.12	50.40	-2.72	63.54	62.08	-1.46
DISCO	53.37	46.63	-6.74	62.54	57.10	-5.44
Ours ($RN18_3$)	20.68	52.64	31.96	31.89	62.45	30.56
Ours ($RN18_4$)	20.70	52.77	**32.07**	29.75	62.98	**33.23**

Table 2: Facial landmark detection with gender classification as a privacy task. The metrics for the privacy and utility task are accuracy (%) and MSE, respectively.

Method	Privacy (Gender) ↓	Utility (FLD) ↓
RN18	98.14	0.0368
MaxEnt	**57.43**	1.2156
Ours ($RN18_3$)	58.53	**0.1766**

classification task performance among the compared methods. Mean squared error (MSE) and accuracy are used as metrics for FLD and classification, respectively.

In Table 2, our method shows an MSE of 0.1766, roughly 7 times better than MaxEnt. Regarding privacy accuracy, ours shows results that are only 1.1%p higher than MaxEnt. The performance of our approach is comparable to MaxEnt since the accuracy is close to 50%, the lower bound for binary classification. $RN18_3$ is more efficient than MaxEnt regarding computational cost (about 23% less GFLOPs), reducing the client-side resource burden.

Table 3: Results on highly correlated privacy and utility tasks. Our method shows the biggest privacy-utility gap (Δ) among the other methods while having less computational cost.

Method	CelebA (Mouth open / Smiling)		
	Privacy ↓	Utility ↑	Δ ↑
RN18	94.20	93.48	–
MaxEnt	80.82	93.29	12.47
DISCO	78.30	90.70	12.40
DeepObfs.	94.06	91.99	-2.07
Ours ($RN18_3$)	76.99	91.56	**14.57**
Ours ($RN18_4$)	57.44	91.61	**34.17**

B CORRELATION BETWEEN PRIVACY AND UTILITY TASKS

We present an experiment with privacy and utility tasks that are highly correlated on the CelebA [Liu et al., 2015] dataset. The experiment in the main paper is conducted with “Gender" as the privacy task and “Smiling" as the utility task. The two classes show a Cramér’s V [Cramér, 2016] correlation coefficient of 0.1367. Here, we test with “Mouth open" as a privacy task and “Smiling" as a utility task, which is a highly correlated task, as made evident by the correlation coefficient of 0.5316.

A result comparison is shown in Table 3. Although DeepObfs. shows relatively high accuracy for the utility task, it fails to defend against the privacy leakage attack. MaxEnt has the best utility accuracy but performs poorly in defense compared with other approaches. DISCO has the lowest privacy while showing the lowest utility among all baselines, which leads to a

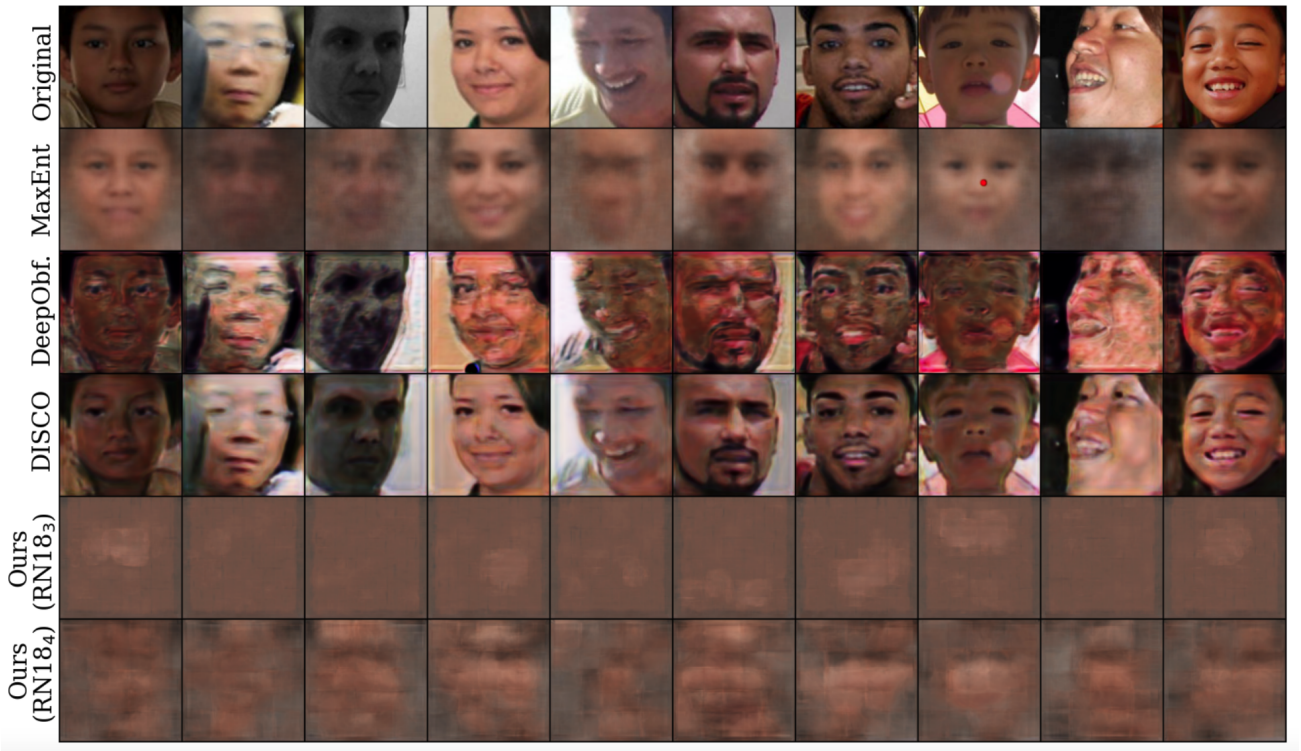

Figure 1: Reconstruction attack on the complex task obfuscators. The reconstructor architecture is from DeepObfs. [Li et al., 2021] and is trained with MSE loss between original image and reconstructed image. Our method is the only one that successfully defend the reconstruction attack by concealing identity, race, age, and face shape.

small privacy-utility gap (Δ). Our methods, RN18$_3$ and RN18$_4$ with $\sigma = 15360$, outperform the other methods regarding the privacy-utility gap (Δ). The Δ of RN18$_4$ is roughly 20%p higher than the others presenting a considerable gain. Even in the case of RN18$_3$, utilizing a more computationally efficient model, the Δ is roughly 2%p higher than the other approaches. The results confirm that our method preserves both privacy and utility performance while being efficient regarding the resource burden, even in the case of highly correlated tasks.

C DETAILED SETTINGS FOR USER STUDY

We report detailed settings for user study to show our best efforts to provide impartial results. We first randomly sample 30 images from CelebA [Liu et al., 2015] that ResNet18 classifies correctly. By doing so, we address the possibility that the original images yield ambiguous results by default, which could affect our results. Then, we randomly selected the test subject group of 30 people in their 20s and 30s who live in Seoul, South Korea. Finally, the images are ordered randomly regardless of the obfuscation method to eliminate any bias that may arise from people noticing a pattern.

For the results in the main paper, note that a 50% ratio of "Correct" and "Wrong" can be considered a random guess since the tasks are binary classification and the data we presented is balanced. Additionally, "Cannot judge" can also be considered a random guess since the users would have guessed answers randomly if there was no "Cannot judge" option. Thus, an equal ratio for "Correct" and "Wrong" answers, meaning 50% per each, or all of the "Cannot judge" options are the best we can achieve regarding privacy protection for binary classification.

D PRIVACY-UTILITY TRADE-OFF UNDER DIFFERENT STANDARD DEVIATIONS

Intuitively, the noise intensity is highly relevant to removing information in the obfuscated feature. More privacy would be achieved for the model with increased intensity since the noise would confuse the adversary model and hinder training the adversary task. However, as the loss of information gets severe, utility accuracy would be negatively affected at a tipping point, eventually leading to a lower privacy-utility trade-off. Thus, we must carefully choose the appropriate noise intensity

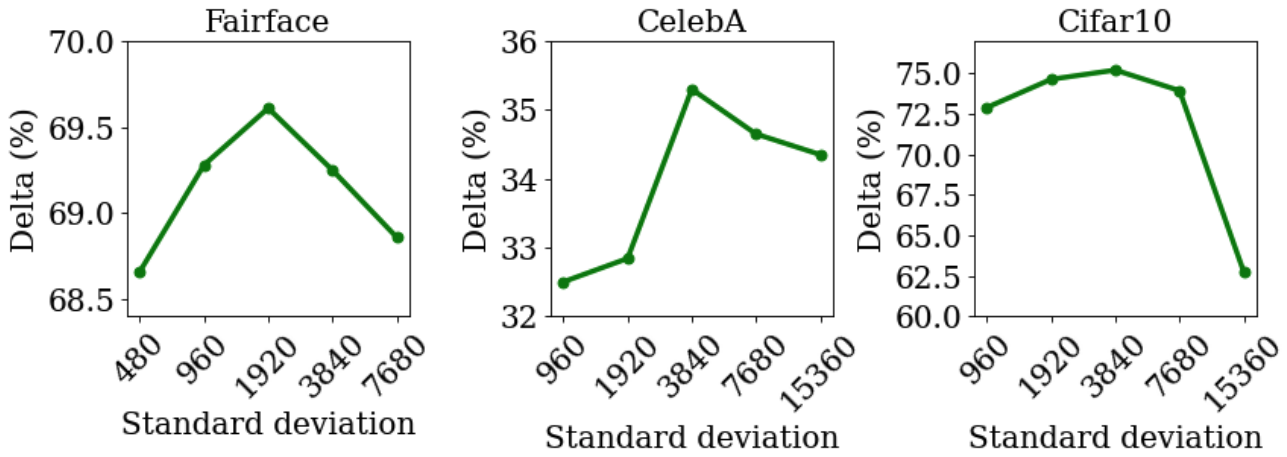

Figure 2: Privacy-utility trade-off under each standard deviation of noise. Delta represents the performance gap between utility and privacy. We report the privacy and utility accuracy in the supplementary materials.

for the best trade-off. We report the privacy-utility trade-off for RN18_3 with various σ in Figure 2. An appropriate σ of noise exists to achieve the best privacy-utility trade-off. Note that the best σ differs for different datasets and models.

References

Harald Cramér. *Mathematical Methods of Statistics*. Princeton university press, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on computer vision and pattern recognition (CVPR)*, 2016.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Ang Li, Jiayi Guo, Huanrui Yang, Flora D Salim, and Yiran Chen. Deepobfuscator: Obfuscating intermediate representations with privacy-preserving adversarial learning on smartphones. In *International Conference on Internet-of-Things Design and Implementation*, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015.

Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Abhishek Singh, Ayush Chopra, Ethan Garza, Emily Zhang, Praneeth Vepakomma, Vivek Sharma, and Ramesh Raskar. Disco: Dynamic and invariant sensitive channel obfuscation for deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.