

# LEARNING PRIVATE REPRESENTATIONS WITH FOCAL ENTROPY (APPENDIX)

**Anonymous authors**

Paper under double-blind review

In the following sections, we add additional details omitted in the main paper due to space restrictions. In Sec. 1, we present an *ablation* study on different components of the loss function, underlining the importance of each term. In Sec. 2, we analyze the impact of varying the neighborhood size of  $k$ -NN in focal entropy on the adversarial accuracy with experiments conducted on the CelebA dataset. In Sec. 3, the sanitization convergence behavior of different classifiers involved in the adversarial minimax game is analyzed. In Sec. 4, we analyze the effect of the *classifier-strength* on the privacy leakage and dependence on training time. In Sec. 5, we present more visualizations around the concept of *hub formation* and the connection to focal entropy. It contains a visualization of hub forming identities and zoom-in visualization of the adversarial remapping of the identities in CelebA. Next, we generated reconstructions on CelebA dataset samples generated in Sec. 6. Finally, *architectural details* are presented in Sec. 7.

## 1 ABLATION ANALYSIS ON LOSS COMPONENTS

To assess the contribution of each component of our objective function, we evaluated each module’s performance separately, gradually adding components: *reconstruction loss*, *target classification loss*, *adversary loss*. For that, we report the ablation study of the loss components on CelebA in Tab. 1. Furthermore, Fig. 1a shows the dependency between adversarial accuracy and the number of training epochs. As can be seen, beyond 1000 epochs, the adversarial accuracy drops below  $< 0.01$ , nearing chance level.

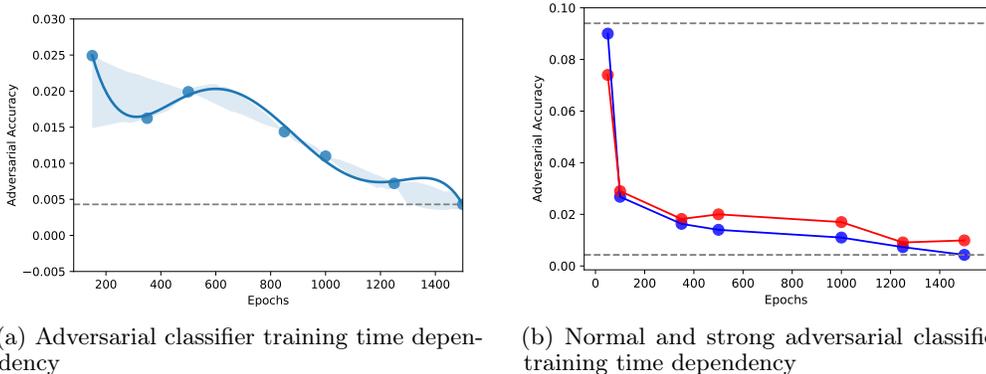


Figure 1: **Left:** Relationship between adversarial accuracy and the number of training epochs on CelebA. The translucent band corresponds to 50% confidence minimum and maximum adversarial accuracy, respectively. **Right:** Relationship between adversarial accuracy for strong (red) and normal classifier (blue) w.r.t. the number of training epochs on CelebA. The translucent band corresponds to 50% confidence interval. Dashed lines correspond to minimum and maximum adversarial accuracy, respectively.

## 2 NEIGHBORHOOD SIZE

Focal entropy entails a notion of similarity tied to the integration of  $k$ -NN. Here, we study the effect of varying  $k$  on focal entropy and the associated the adversary accuracy. See Fig. 2 for a visualization of this relationship on the CelebA dataset. As can be seen, the adversary accuracy has oscillatory behavior with various local minima, reaching optimum around  $k = 16$ . This can be attributed to the superpositioning of different hubs, each exhibiting a different similarity pattern. Analysis of hub formation is explained in Sec. 4.2 in the main paper and Fig. 4 in the main paper.

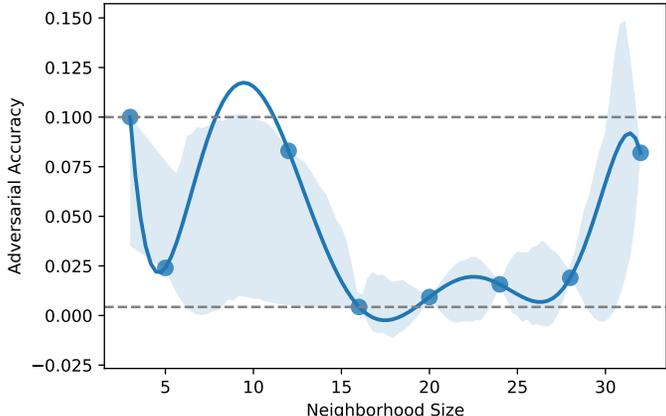


Figure 2: Relationship between adversary accuracy and  $k$ -NN size on CelebA dataset. Translucent band corresponds to 50% confidence interval. Dashed lines correspond to minimum, and maximum adversary accuracy, respectively.

## 3 SANITIZATION CONVERGENCE BEHAVIOUR

This section explores the behavior of standard entropy and the proposed focal entropy for sanitization. Fig. 3 depicts the classification performance during the training of different classifiers involved in the minimax optimization scheme: target classifier accuracy, adversarial sensitive attribute accuracy, and sensitive attribute accuracy. As can be seen, employing standard entropy for sanitization results in re-occurring patterns of oscillations. This can be attributed to degenerate/trivial solutions and “shortcuts”. In contrast to that, focal entropy shows a relatively smooth convergence behavior. More details on the behavior can be found in Sec. 4.2 in the main paper.

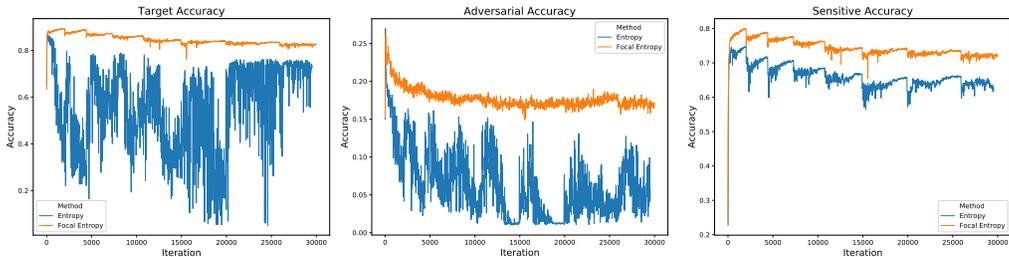


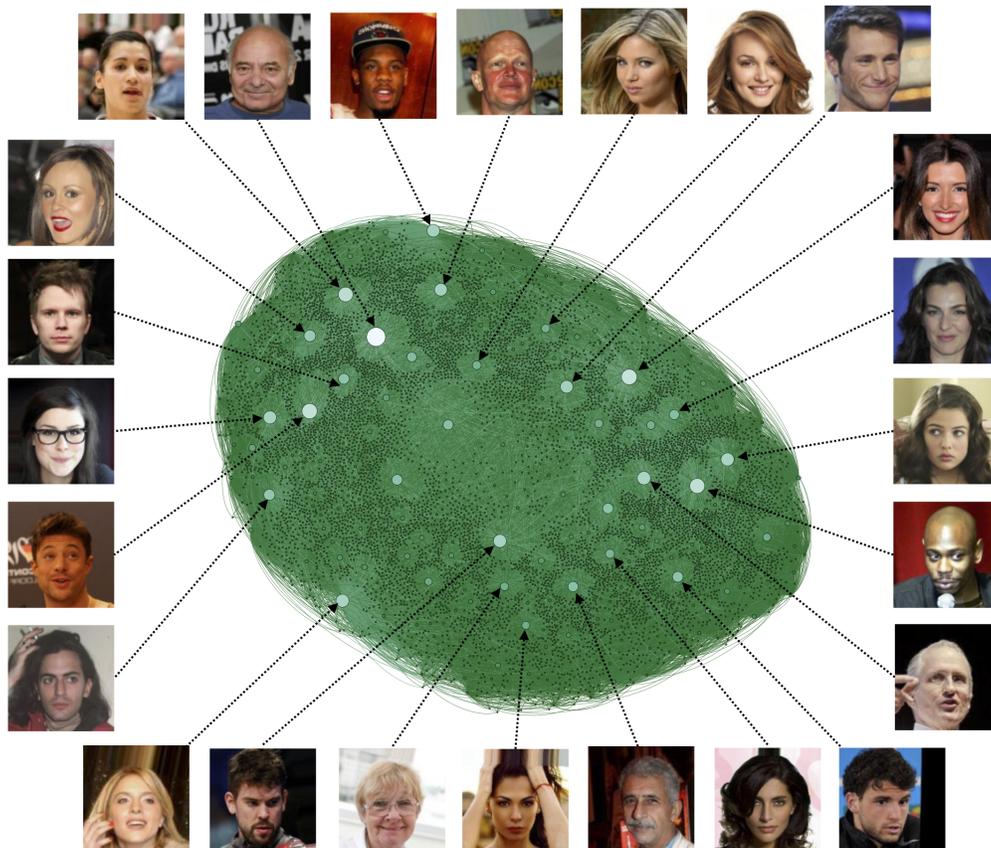
Figure 3: Sanitization convergence behavior of standard entropy and focal entropy on CIFAR-100 for different classifiers: **Left:** Target accuracy, **Center:** Adversarial accuracy, **Right:** Sensitive attribute accuracy

## 4 PROBING ANALYSIS WITH STRONG CLASSIFIER

This section provides more detail on assessing the classifier strength in terms of privacy leakage and the dependence on training time. We thereby largely follow the protocol of Harsh Jha et al. (2018); Sadeghi et al. (2019). Specifically, we employed a *stronger* post-classifier (Tab. 3 in the main paper) compared to the one used for learning the representation. The stronger post-classifier is endowed with additional layer stack (see Tab. 5c for architectural details), trained for 100 epochs. The results in Tab. 3 of the main paper suggest no significant changes in target and adversarial accuracy. Figure 1b extends these results, depicting the relationship w.r.t. the number of epochs. As can be seen, the difference between the *strong* and *normal* classifier is largely constant, independent of the epoch.

Method	CelebA Guo et al. (2016)	
	Target Acc.	Adversarial Acc.
Upper-bound / Random Chance	1.0	< 0.001
Our Method (only <i>Rec.</i> loss)	0.88	-
Our Method ( <i>Rec.</i> + <i>Tar.</i> loss)	0.91	0.751
Our Method[full] ( <i>Rec.</i> + <i>Tar.</i> + <i>Adv.</i> loss)	0.90	< 0.01

Table 1: Ablation analysis for loss components on CelebA dataset.

Figure 4: Visualization of CelebA identities of adversary classification network. The network (green) corresponds the  $k$ -nearest neighborhood size  $k = 5$ .

## 5 EXTENDED HUB ANALYSIS

This section provides a visually more detailed analysis of how the application of focal entropy promotes the formation of “hubs” (explained in Sec. 4.2 in the main paper). To study that, we analyzed the identity remapping of IDs on the CelebA dataset. Employing focal entropy results in a surjective ID confusion pattern by taking similar IDs into account for privacy sanitization.

### 5.1 VISUALIZATION OF HUB FACES:

To study hubs’ semantics, we visualize the CelebA identities of the network corresponding to focal entropy with  $k$ -nearest neighborhood size  $k = 5$ . See Fig. 4 for the visualization of the hub faces. As can be seen, the hubs exhibit a rich diversity in facial properties.

### 5.2 ADVERSARIAL IDENTITY MAPPING:

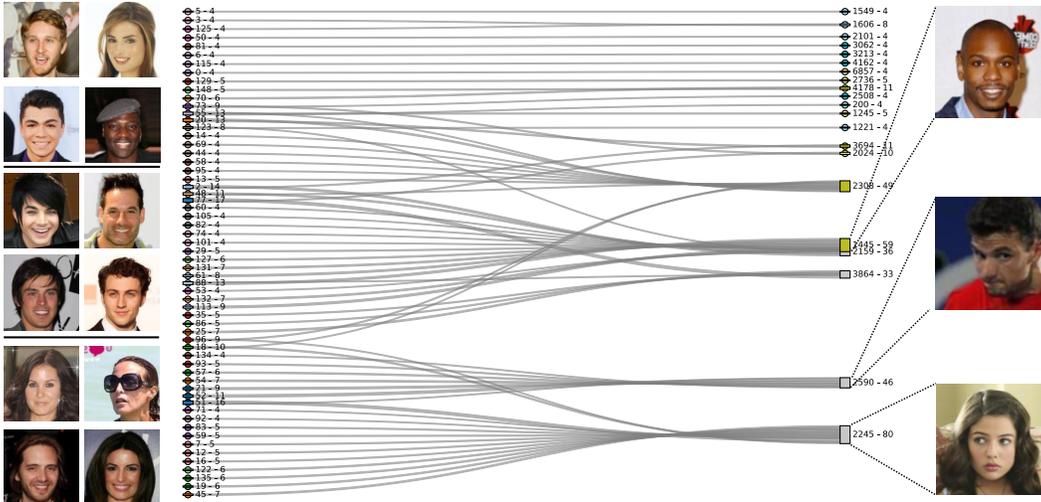


Figure 5: Visualization of the remapping of IDs in CelebA due to adversarial representation learning. Source IDs (left) are remapped to new target IDs (right). Pictures on the left are samples that get mapped to a hub; separation with bar indicates different target hub. Pictures on the right are the visualization of hub identities. Visualization contains a subset of 150 IDs, with targets getting at least four associations. The first number at each node indicates the ID, the second the number of images per ID. Node splicing indicates the remapping of a single ID to multiple adversarial targets.

Figure 5 is a zoom-in version of a graph as shown in Fig. 4 in the main paper, with  $k$ -nearest neighborhood size  $k = 5$ . This visualization provides a more in-depth view of how the adversarial process leads to a remapping of identities. In order to avoid visual clutter, a subset of identities and targets was chosen. As can be seen, instead of being a collapse of a facial stereotype, each hub is associated with a diverse looking set of identities, giving rise to the deep sanitization of the representation.

## 6 QUALITATIVE RESULTS

Figure 6 shows different reconstructions of additional CelebA identities (equal male and female) at different privacy levels. Each column is two different samples from CelebA (one male and one female), and from top to bottom, the privacy disclosure is decreasing for each. It can be noticed that visualizations from residual latent part and target latent parts confirm the sanitization visually.

## 7 ARCHITECTURAL DETAILS

We describe the architectures of each part of our model. Table 4 shows the architectures of the VAE, i.e., the encoder and the decoder. It should be noted that the last two layers of the encoder in Tab. 2 arise from layer splitting to accommodate for partitioning target and residual representations. This is highlighted with dashed lines. Furthermore, we provide the architectures of classifiers in Tab. 5. Architectures for target and adversarial classifiers are identical.

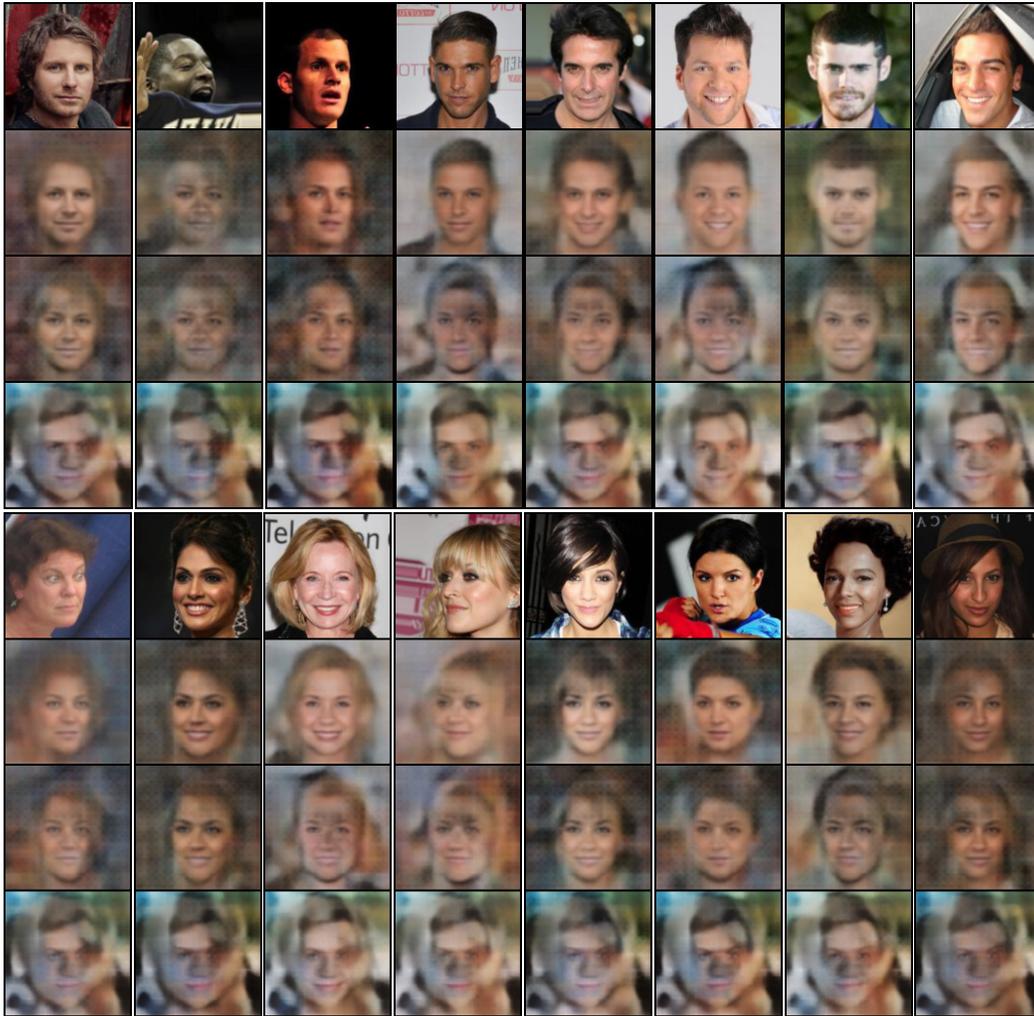


Figure 6: Visualization of CelebA data and reconstructions at different privacy levels. (From top to bottom, privacy revelation is decreasing).

Layer	Output	Parameters
Input: $128 \times 128 \times 3$		
Conv-2d	$64 \times 64$	$64 \times [3 \times 3]$ , st. 2
BatchNorm		
LeakyReLU	negative slope: 0.01	
Conv-2d	$32 \times 32$	$128 \times [3 \times 3]$ , st. 2
BatchNorm		
LeakyReLU	negative slope: 0.01	
Conv-2d	$16 \times 16$	$256 \times [3 \times 3]$ , st. 2
BatchNorm		
LeakyReLU	negative slope: 0.01	
Conv-2d	$8 \times 8$	$512 \times [3 \times 3]$ , st. 2
BatchNorm		
LeakyReLU	negative slope: 0.01	
Linear	$1 \times 4096$	
Linear	$1 \times 512$	
Linear	$1 \times 4096$	
Linear	$1 \times 512$	

Table 2: Encoder

Layer	Output	Parameters
Linear	32768	
BatchNorm		
LeakyReLU	negative slope: 0.01	
DeConv-2d	$16 \times 16$	$256 \times [3 \times 3]$ , st. 2
BatchNorm		
LeakyReLU	negative slope: 0.01	
Conv-2d	$32 \times 32$	$128 \times [3 \times 3]$ , st. 2
BatchNorm		
LeakyReLU	negative slope: 0.01	
Conv-2d	$64 \times 64$	$64 \times [3 \times 3]$ , st. 2
BatchNorm		
LeakyReLU	negative slope: 0.01	
Conv-2d	$128 \times 128$	$3 \times [3 \times 3]$ , st. 2
Tanh		

Table 3: Decoder

Table 4: Architectural details of VAE components. Parameters for convolutions correspond to: number kernels  $\times$  [ kernel size ], and stride

Layer	Output size / Params
Linear	256
BatchNorm	
PReLU	
Dropout	drop-rate: 0.2
Linear	128
BatchNorm	
PReLU	
Linear	#classes

(a) Classifier on CIFAR-100

Layer	Output size / Params
Linear	256
BatchNorm	
PReLU	
Dropout	drop-rate: 0.5
Linear	128
BatchNorm	
PReLU	
Linear	#IDs or #attributes $\times$ [1]

(b) Normal classifier on CelebA

Layer	Output size / Params
Linear	256
BatchNorm	
PReLU	
Dropout	drop-rate: 0.5
Linear	128
BatchNorm	
PReLU	
Dropout	drop-rate: 0.5
Linear	256
BatchNorm	
PReLU	
Linear	#IDs

(c) Strong ID classifier on CelebA

Table 5: Architectural details of the used classifiers

## REFERENCES

- Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pp. 87–102. Springer, 2016.
- Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- Bashir Sadeghi, Runyi Yu, and Vishnu Boddeti. On the global optima of kernelized adversarial representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7971–7979, 2019.