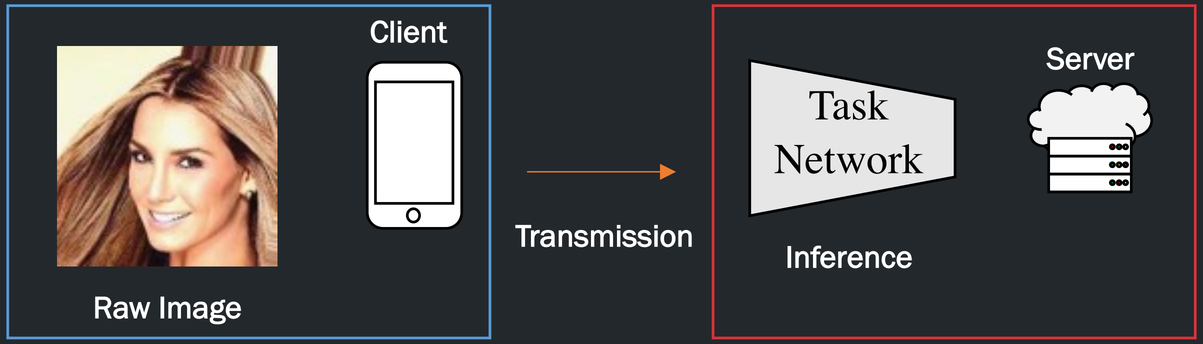
Privacy Safe Representation Learning via Frequency Filtering Encoder

Jonghu Jeong, Minyong Cho, Philipp Benz, Jinwoo Hwang, Jeewook Kim, Seungkwan Lee, Tae-hoon Kim
Deeping Source Inc.

Problem

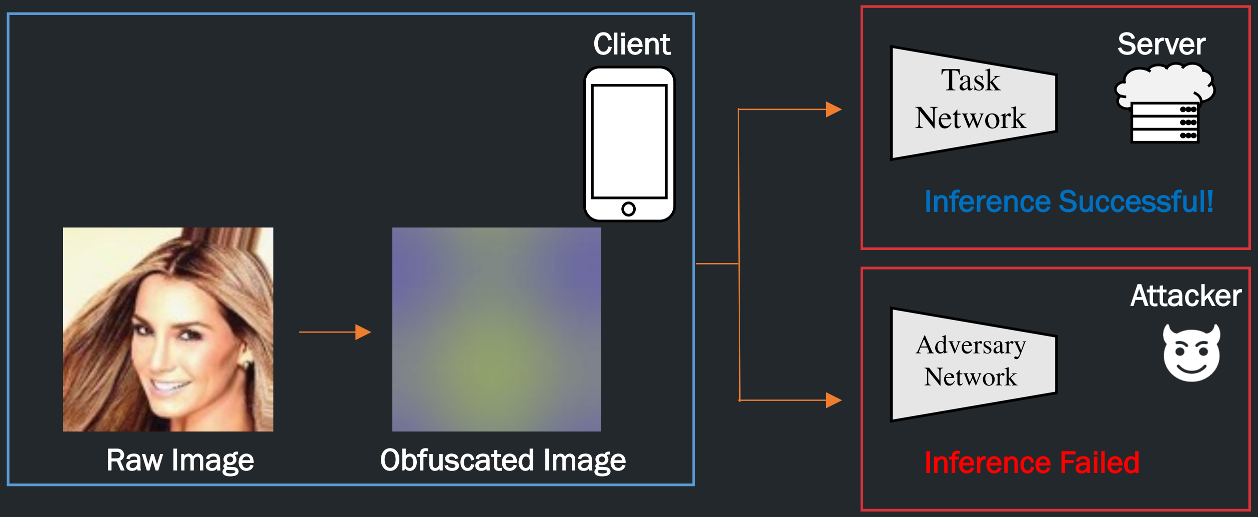


For server-side neural network inference, raw images should leave the client-side.

Due to the direct sharing of raw images, potential privacy leakage is a big concern.

➢ Privacy-safe representation of the raw images needs to be generated on client-side and then be transmitted to the server-side.

Problem: Privacy-Safe Representation



Privacy-safe representations (aka obfuscated images)

- can not be abused by malicious attackers
  - Attribute inference attack (ex; gender, age)
  - Reconstruction attack (from privacy-safe representation to original image)
- can be utilized with ML models for pre-designated tasks
  - Ex) Facial expression classification

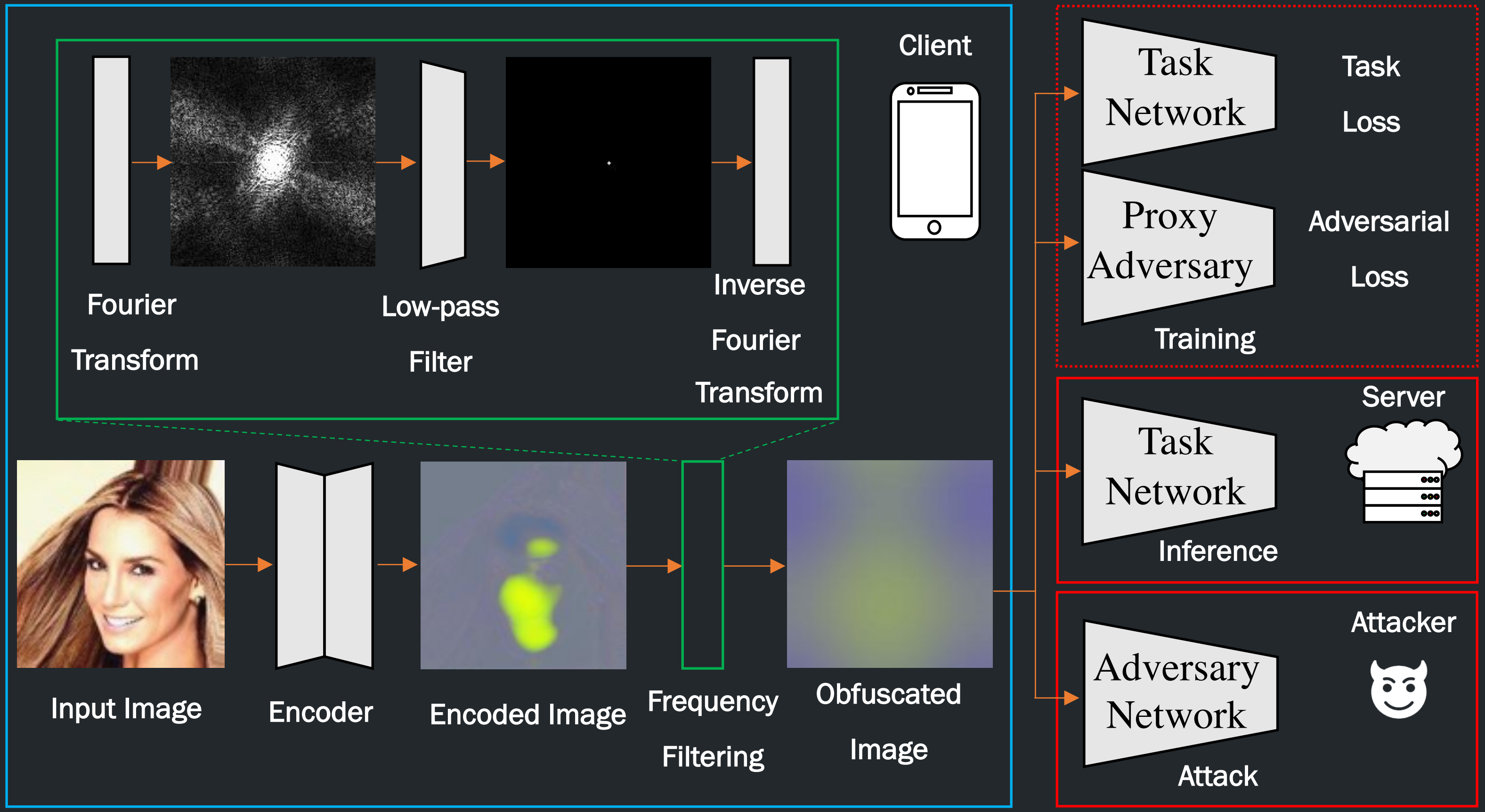
* [Bertran et al., 2019], [Singh et al., 2021]

Existing Solution: Adversarial Representation Learning

1. Predefine *utility task* and *adversary task*.
   - *Utility task* is what the server will eventually inference from the encoded images.
   - We can consider *adversary task* as a proxy for possible adversary attacks.
2. Train the encoder, utility task model, and adversary task model simultaneously.
   - The utility task model tries to minimize utility task loss.
   - The adversary task model tries to minimize adversary task loss.
   - The encoder tries to minimize utility task loss and maximize adversary task loss.
3. After the training is done,
   - Deploy the encoder to the client.
   - Deploy the utility task model to the server.

* The existing methods vary from specific loss formulation, model architecture, and training scheme.

Proposed Method : ARL + Frequency filtering



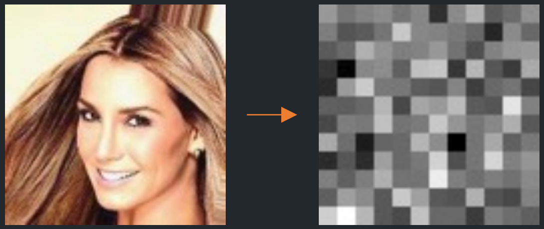
Intuition: Information encoded in low or high-frequency range is enough for CNN to learn. [Yin et al., 2019], [Wang et al.,2020]
We used U-Net [Ronneberger et al., 2015] for the encoder.
Different from previous ARL methods that utilized special 1) loss function 2) neural net model architecture 3) training scheme, we simply added frequency filtering module to the basic ARL scheme.
The frequency module helps with removing privacy-related information effectively.

Evaluation Protocol

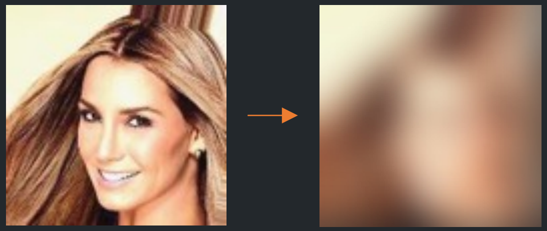
1. Choose an image dataset with various classification attributes.
   - Ex) CelebA [Liu et al., 2015] dataset has 40 various facial attributes.
2. Select attributes for utility task and adversary task, respectively.
   - From CelebA, use ‘smile’ as a utility task and ‘gender’ as an adversary task.
3. Train an encoder, utility task model, and adversary task model with the ARL training scheme.
4. Report utility task accuracy and adversary task accuracy, and their difference.
   - Utility task accuracy is higher the better.
   - Adversary task accuracy is lower the better.
   - Thus, the difference between two is higher the better.
5. Train a reconstruction attacker and report quantitative/qualitative results.
   - Report visual dissimilarity scores between the original and reconstructed images.

Evaluation Protocol: Compared methods

1. Noise Addition (*Noise*)*: Add Gaussian noise to the image, so that the result does not show any information to human eye.
2. Low-pass Filtering (*LP*)*: The resulting image does not show any information to human eye.
3. Basic ARL (*U-Net*): Train U-Net [Ronneberger, 2015] with the ARL training scheme.
4. *DISCO* [Singh et al, 2021]: Channel attention based ARL method.
5. Ours : Combination of *LP* and *U-Net.*



< Noise >



< LP>

** Neural network is unnecessary for these methods.*

Results: Attribute Inference Attack

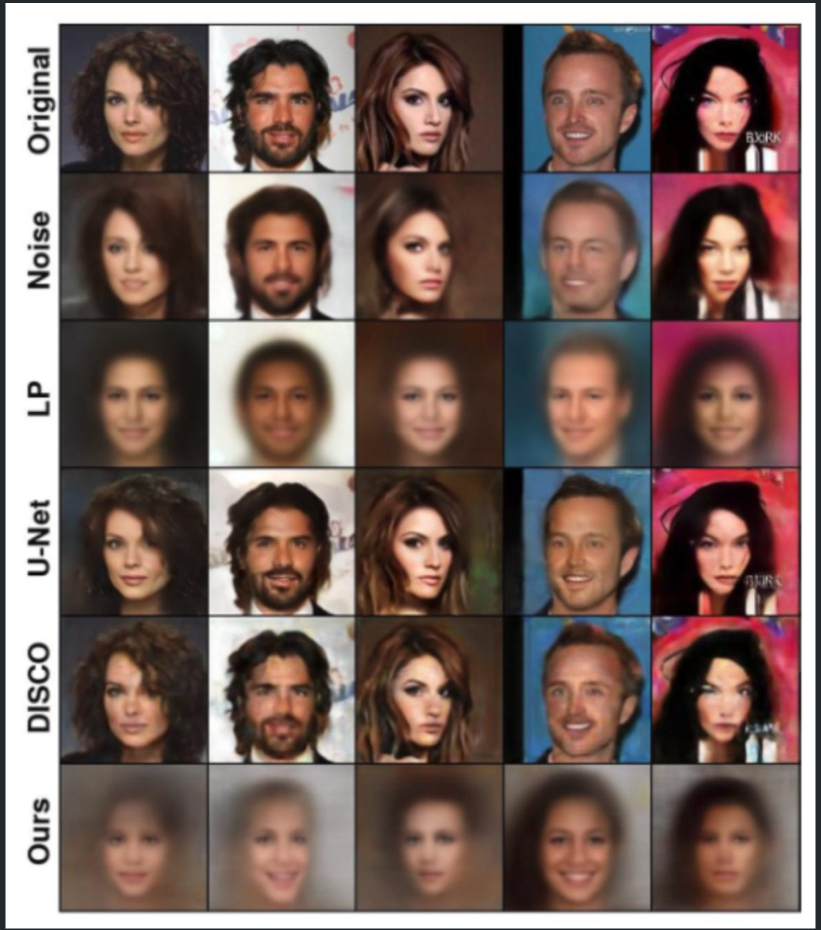
Method	Fairface Privacy ↓	Fairface Utility ↑	Fairface Δ ↑	CelebA Privacy ↓	CelebA Utility ↑	CelebA Δ ↑	CIFAR10 Privacy ↓	CIFAR10 Utility ↑	CIFAR10 Δ ↑
Perf. Bounds	19.03	90.16	71.13	57.43	93.32	35.89	10.00	98.79	78.79
Noise	42.61	74.33	31.72	91.71	85.38	-6.33	54.37	87.77	33.40
LP	31.93	64.77	32.84	76.52	63.69	-12.83	47.05	85.76	38.71
U-Net	51.52	86.40	34.88	87.21	93.12	5.91	85.05	95.45	10.40
DISCO	19.00	81.50	62.50	61.20	91.00	29.80	22.30	91.98	69.68
Ours	23.63	89.67	**66.04**	61.60	93.27	**31.67**	22.58	92.95	**70.37**

Privacy-Utility trade-offs among various methods and datasets.
Ours shows the biggest gap which means it effectively removes private information while retaining utility.

We followed the same task setting from DISCO [Singh et al, 2021].

- FairFace [Karkkainen and Joo, 2021]: Utility=Gender classfication, Privacy=Race classification
- CelebA [Liu et al., 2015]: Utility='Smile' classification, Privacy=Gender classification
- CIFAR10 [Krizhevsky , 2009]: Utility=Living vs. non-living classification, Privacy=10 classes classification

Results: Reconstruction Attack



* CelebA Dataset
Utility=Smiling, Privacy=Gender

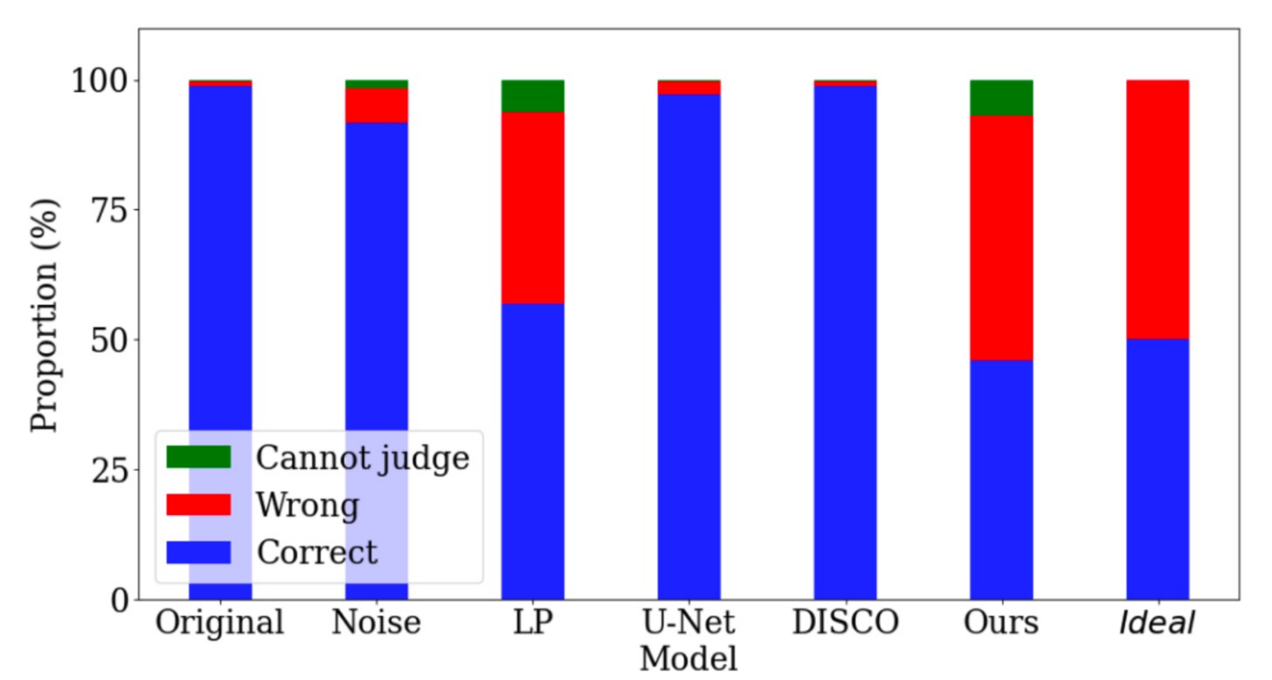
The reconstruction attack is successful on the existing ARL methods (*U-Net, DISCO*).
With naïve methods (*Noise, LP*), the identity is hidden but the 'gender' is still revealed.
Our method: The identity and gender are properly hidden.

Results: Reconstruction Attack

Method	MSE ↑	L_1 ↑	SSIM ↓	MS-SSIM ↓	PSNR ↓	LPIPS ↑
Noise	584.88	16.97	0.6017	0.7776	20.46	0.3714
LP	1889.15	32.10	0.4632	0.5390	15.37	0.5537
U-Net	390.34	13.81	0.7505	0.8839	22.22	0.1809
DISCO	567.17	15.94	0.5765	0.7611	20.60	0.4351
Ours	**3689.50**	**48.08**	**0.4240**	**0.4728**	**12.47**	**0.6145**

We used commonly used visual metrics to assess visual difference.
Ours shows the best dissimilarity score between the original and reconstructed images.
This result reconfirms the qualitative result.

Results: Reconstruction Attack



We asked 30 people to judge 'gender' of the reconstructed images.
While the participants have judged with high accuracy among the compared methods,
ours showed accuracy of 56.9%, which is almost same to random guessing the task is binary classification.

Conclusion

- A novel approach that combines frequency filtering and a neural net for privacy-safe machine learning.
- High privacy-utility trade-off.
- Robustness to the reconstruction attack quantitatively/qualitatively.
- Empirically confirmed the performance with various experiments and user study.

References

1. M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Ro- drigues, G. Reeves, G. Sapiro, Adversarially learned representations for information obfuscation and inference, in: International Conference on Machine Learning (ICML), 2019.
2. A. Singh, A. Chopra, E. Garza, E. Zhang, P. Vepakomma, V. Sharma, R. Raskar, Disco: Dynamic and invariant sensitive channel obfuscation for deep neural networks, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
3. Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: International Conference on Computer Vision (ICCV), 2015.
4. K. Karkkainen, J. 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.
5. A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report, 2009.
6. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, 2015.
7. D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, J. Gilmer, A fourier perspective on model robustness in computer vision, in: Advances in neural information processing systems (NeurIPS), 2019.
8. H. Wang, X. Wu, Z. Huang, E. P. Xing, High-frequency component helps explain the generalization of convolutional neural networks, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2020.