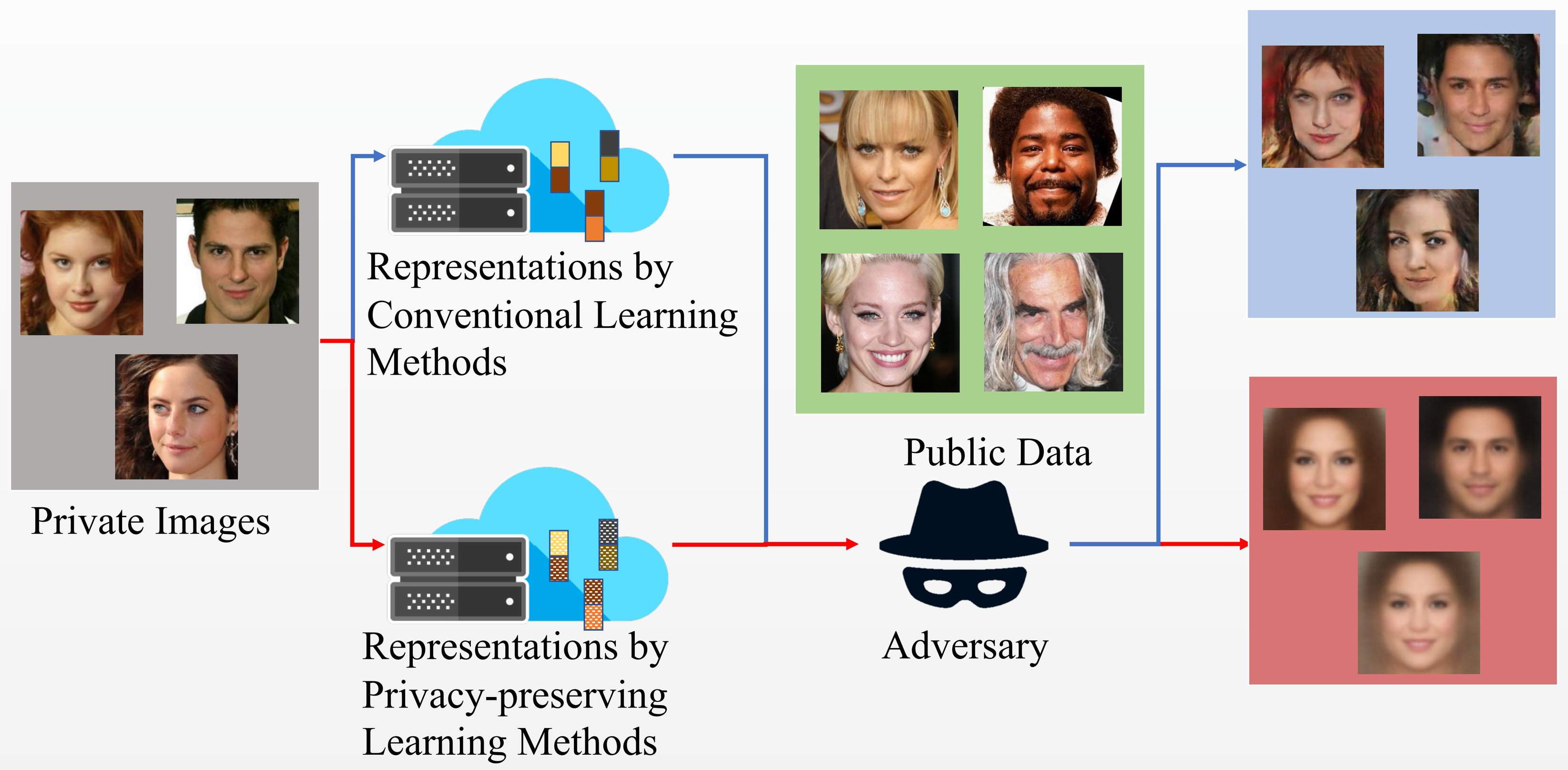
# Adversarial Learning of Privacy-Preserving and Task-Oriented Representations


Taihong Xiao[1], Yi-Hsuan Tsai[2], Kihyuk Sohn[2], Manmohan Chandraker[2,3], Ming-Hsuan Yang[1]

[1]University of California, Merced [2]NEC Laboratories America [3]University of California, San Diego


## Introduction




### Motivation

- Privacy risk in machine learning cloud services
- Learning deep features that protect the privacy

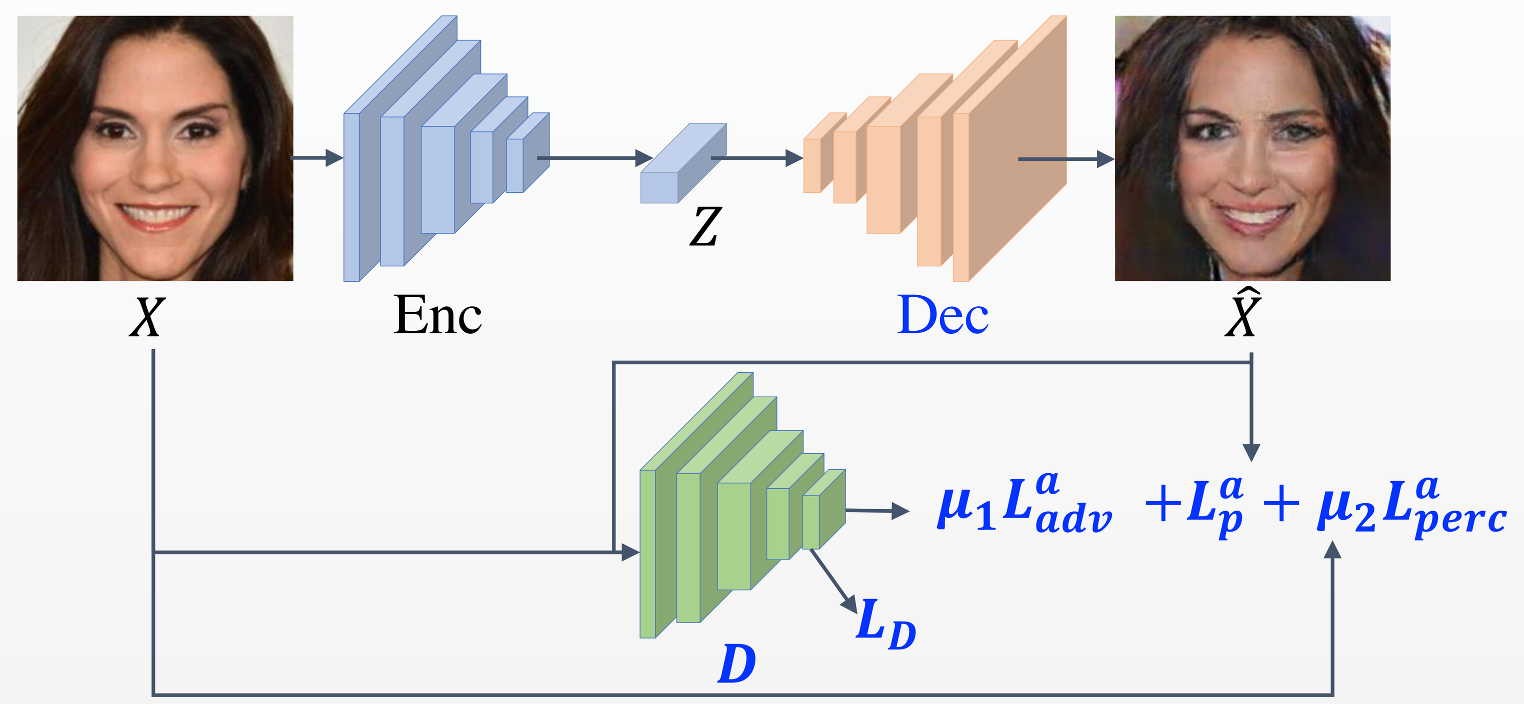
### Problem Context

- Black-box model inversion: the adversary can make unlimited inferences of their own data to recover input from acquired features of private user data
- Defense against black-box model inversion attacks in the context of face attribute analysis via adversarial learning

### Our Solution

- Propose to consider perspectives from both the adversary and protector to learn privacy-preserved models
- Seek for balancing utility on face attribute classification while protecting the facial privacy
- Provide extensive study to analyze the impact on privacy protection in the proposed framework

## Proposed Algorithm

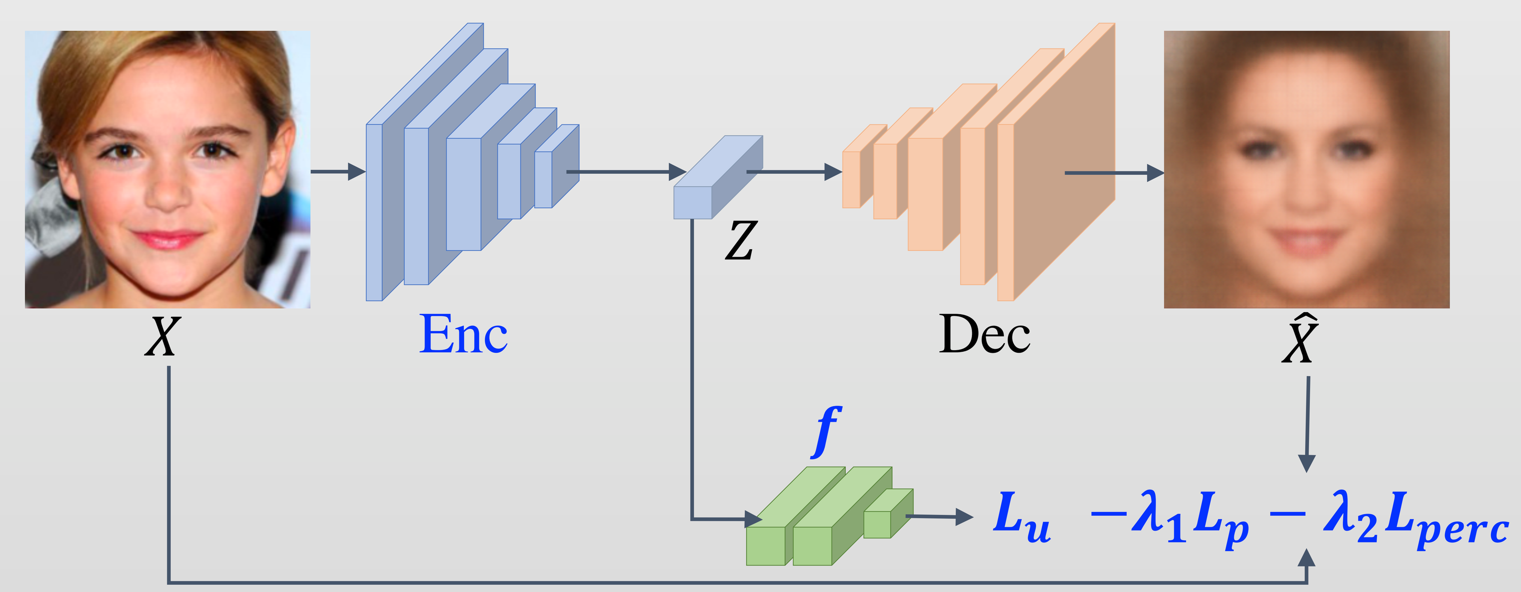



Adversary: updating Dec and D using public data $\mathcal{X}_2$ while fixing Enc and $f$

- Privacy loss: $L^a_p = \mathbb{E}_{\{(X \in \mathcal{X}_2, Z)\}}[\| \hat{X} - X \|^2]$
- GAN loss: $L^a_{\text{adv}} = \mathbb{E}_Z[\log(1 - D(\hat{X}))]$
- Perceptual loss: $L^a_{\text{perc}} = \mathbb{E}_{\{(X \in \mathcal{X}_2, Z)\}}[\| g(\text{Dec}^a(Z)) - g(X) \|^2]$

The overall objective of an adversary is

$$\min_{\text{Dec}^a} L^a_p + \mu_1 L^a_{\text{adv}} + \mu_2 L^a_{\text{perc}}$$




Protector: updating Enc and $f$ using private data $\mathcal{X}_1$ while fixing Dec

- Utility loss: $L_u = \mathbb{E}_{\{(X \in \mathcal{X}_1, Y)\}}[\mathcal{L}(f(Z), Y)]$
- GAN loss: $L_p = \mathbb{E}_{\{(X \in \mathcal{X}_1, Z)\}}[\| \text{Dec}(Z) - X \|^2]$
- Perceptual loss: $L_{\text{perc}} = \mathbb{E}_{\{(X \in \mathcal{X}_1, Z)\}}[\| g(\text{Dec}(Z)) - g(X) \|^2]$

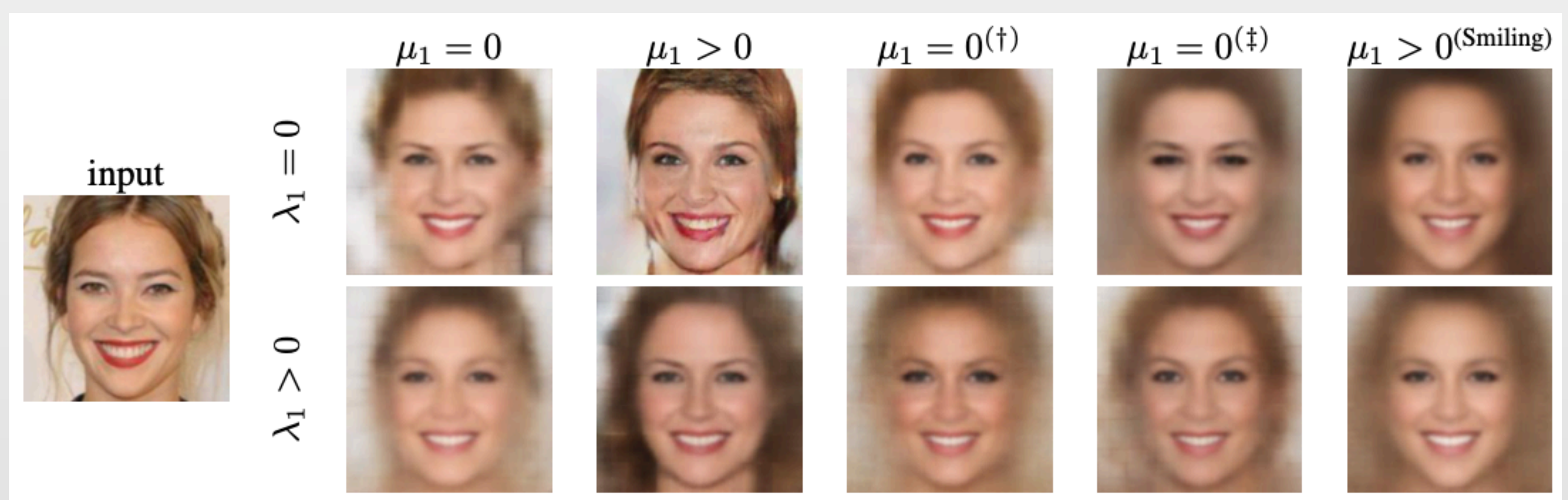
The overall objective of an adversary is

$$\min_{\text{Enc}, f} L_u - \lambda_1 L_p - \lambda_2 L_{\text{perc}}$$

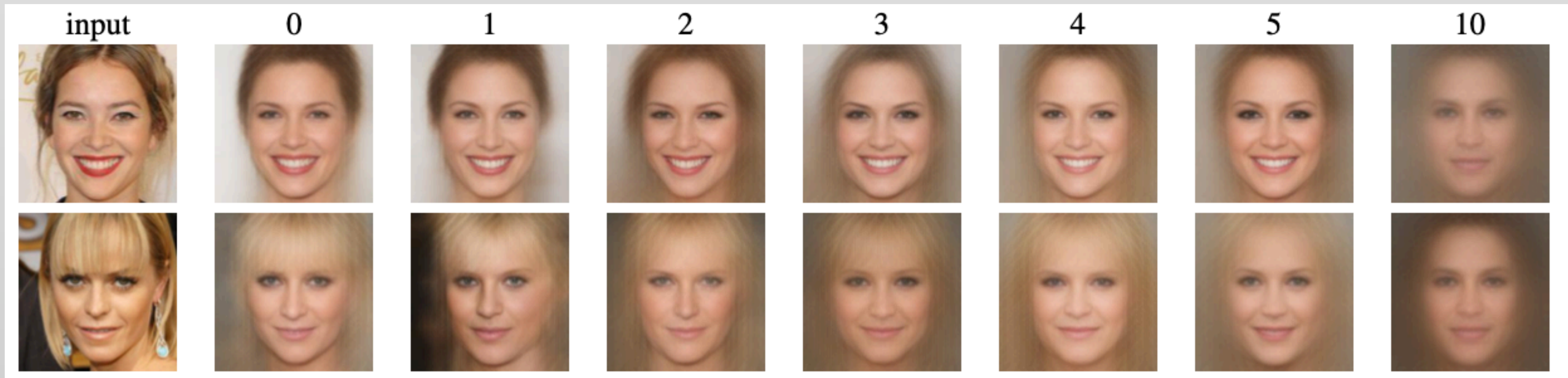
## Experimental Results

| ID | Enc | Dec$^a$ | Mean MCC ↑ | Face Sim. ↓ | Feature Sim. ↓ | SSIM | PSNR |
|---|---|---|---|---|---|---|---|
| 1 | $\lambda_1 = 0$ | $\mu_1 = 0$ | 0.641 | 0.551 | 0.835 | 0.231 | 13.738 |
| 2 | $\lambda_1 > 0$ | $\mu_1 = 0$ | 0.612 | 0.515 | 0.574 | 0.221 | 13.423 |
| 3 | $\lambda_1 = 0$ | $\mu_1 > 0$ | 0.641 | 0.585 | 0.835 | 0.240 | 14.065 |
| 4 | $\lambda_1 > 0$ | $\mu_1 > 0$ | 0.612 | 0.513 | 0.574 | 0.277 | 13.803 |
| With more data for training Dec$^a$ (ID #5 and #6) and both Enc and Dec$^a$ (ID #7 and #8) | | | | | | | |
| 5 | $\lambda_1 = 0^\dagger$ | $\mu_1 = 0$ | 0.641 | 0.594 | 0.864 | 0.250 | 14.132 |
| 6 | $\lambda_1 > 0^\dagger$ | $\mu_1 = 0$ | 0.612 | 0.541 | 0.633 | 0.222 | 13.703 |
| 7 | $\lambda_1 = 0^\ddagger$ | $\mu_1 = 0$ | 0.651 | 0.579 | 0.834 | 0.263 | 14.432 |
| 8 | $\lambda_1 > 0^\ddagger$ | $\mu_1 = 0$ | 0.630 | 0.550 | 0.591 | 0.231 | 13.334 |
| Single (Smiling) attribute prediction. MCC for Smiling attribute is reported in the parenthesis. | | | | | | | |
| 9 | $\lambda_1 = 0$ | $\mu_1 > 0$ | 0.001 (0.851) | 0.460 | 0.494 | 0.204 | 13.214 |
| 10 | $\lambda_1 > 0$ | $\mu_1 > 0$ | 0.044 (0.862) | 0.424 | 0.489 | 0.189 | 12.958 |

Results on facial attribute prediction. We report the MCC over 40 attributes as a utility metric, while face and feature similarities are privacy metrics.




Visualization of reconstruction by Dec$^a$. Examples in the first and second row are results with/without employing the negative reconstruction loss.




Results with different $\lambda_2$ in the training stage. As we increase $\lambda_2$, the model becomes more privacy-preserved.