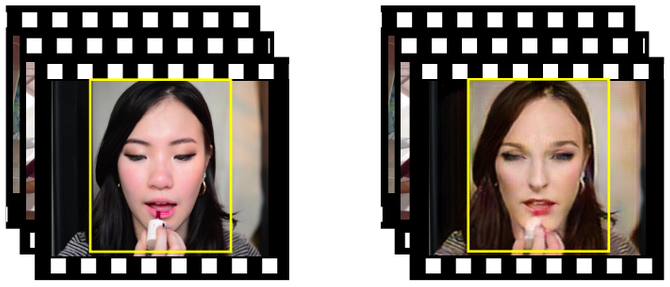
Learning to Anonymize Faces for Privacy Preserving Action Detection

Zhongzheng Ren†*, Yong Jae Lee†*, and Michael S. Ryoo†

†EgoVid Inc. *UC Davis

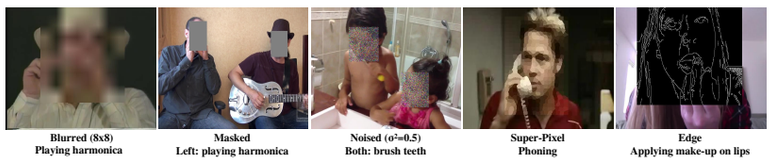
Task: *Learning to **anonymize** videos in a way that does not negatively affect recognition of human activities.*



Identity: Jessica
Action: Applying Make-up on Lips

Identity: ???
Action: Applying Make-up Lips

Observation: *Existing work largely use hand-crafted image processing methods to modify images/videos for privacy protection.*



Blurred (8x8) Playing harmonica | Masked Left: playing harmonica | Noised (σ^2=0.5) Both: brush teeth | Super-Pixel Phoning | Edge Applying make-up on lips

- **Blur** *(down-sample to extreme-low resolution): Ryoo et al. AAAI'17, AAAI'18*
- **Masking** *and strong* **Gaussian Noise**: *Standard practice.*
- **Super-Pixel / Edge map**: *Butler et al. HRI'15*

Multi-task Adversarial Learning:

- *Adversarial loss:*
  (1) modifier M removes privacy-sensitive face information while also being optimized for action detection;
  (2) discriminator D (face classifier) tries to extract privacy-sensitive information from modified faces:

$$L_{adv}(M, D, F) = -\mathbb{E}_{(f\sim F, i_f\sim I)}[L_D(M(f), i_f)] - \mathbb{E}_{(f\sim F, i_f\sim I)}[L_D(f, i_f)]$$

- *Detection loss:*

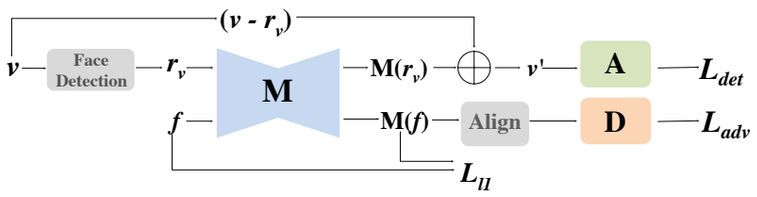
$$L_{det}(M, A, V) = \mathbb{E}_{v\sim V}[L_A(v', \{b_i(v)\}, \{t_i(v)\})]$$

- *L1 loss preserves basic image structure:*

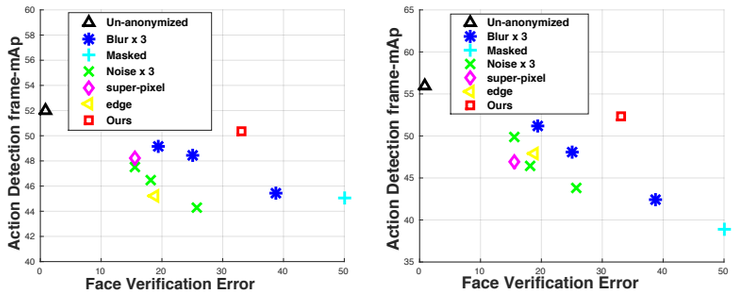
$$L_{l1}(M, F) = \mathbb{E}_{f\sim F}[\ \lambda\ ||M(f) - f||_1]$$

$$\arg\min_{M,A}\max_{D} L_{det}(M, A, V) + L_{adv}(M, D, F) + L_{l1}(M, F)$$

Architecture



Experiments



Left: *JHMDB* **Right:** *DALY*

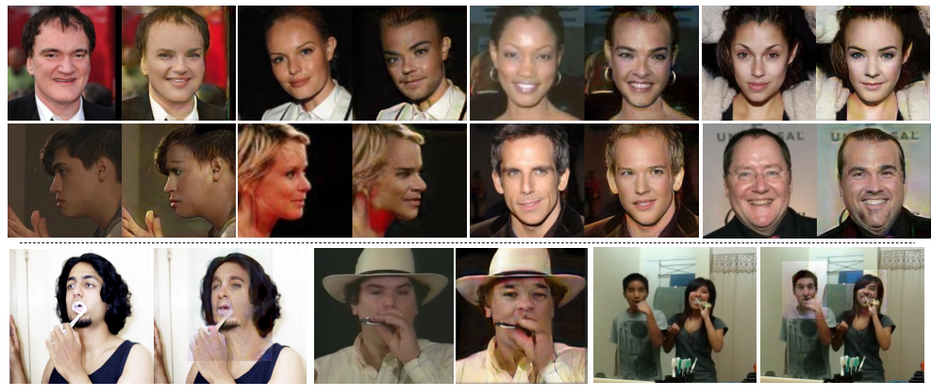
- ***A good model should locate on the top right corner.***
- ***Our method outperforms various baselines on both action & face tasks.***

User Study: *We applied trained modifier on **new** identities and collected 400 answers from 10 users.*

- ***Q1**: We sample a pair of modified images and ask user to do verification.*
  ***Res**: The overall accuracy is **53.3%**, which is close to random guess 50%.*
- ***Q2**: We use our model to modify famous celebrities and ask user to identify.*
  ***Res**: Users could only name **19.75%** correctly based on the modified images.*

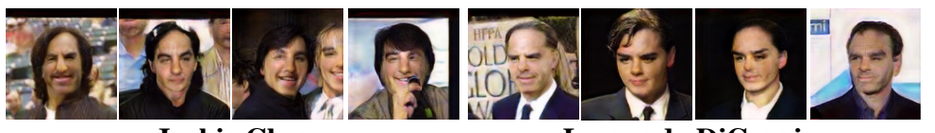
Demo tomorrow! #2 Session 2-A, 10am-12am, Sep.11
Code, demo, and more results available: jason718.github.io/project/privacy/main.html

Qualitative Results #1



Same person before and after modification: *the identity is greatly changed.*

Qualitative Results #2



Jackie Chan **Leonardo DiCaprio**

Serana Williams **Vladimir Putin**

Video 1 **Video 2**

Same person in different frames: *can you still identify the above celebrities?*

Analysis: *After training, discriminator D (face classifier) still get **94.75%** accuracy on LFW; i.e., it can still accurately recognize original faces despite being "fooled" by modified faces.*

Acknowledgment

This research was conducted as a part of EgoVid Inc.'s research activity on privacy preserving computer vision. This work was supported by the Technology development Program (S2557960) funded by the Ministry of SMEs and Startups (MSS, Korea).