On the (Im)Practicality of Adversarial Perturbation for Image Privacy

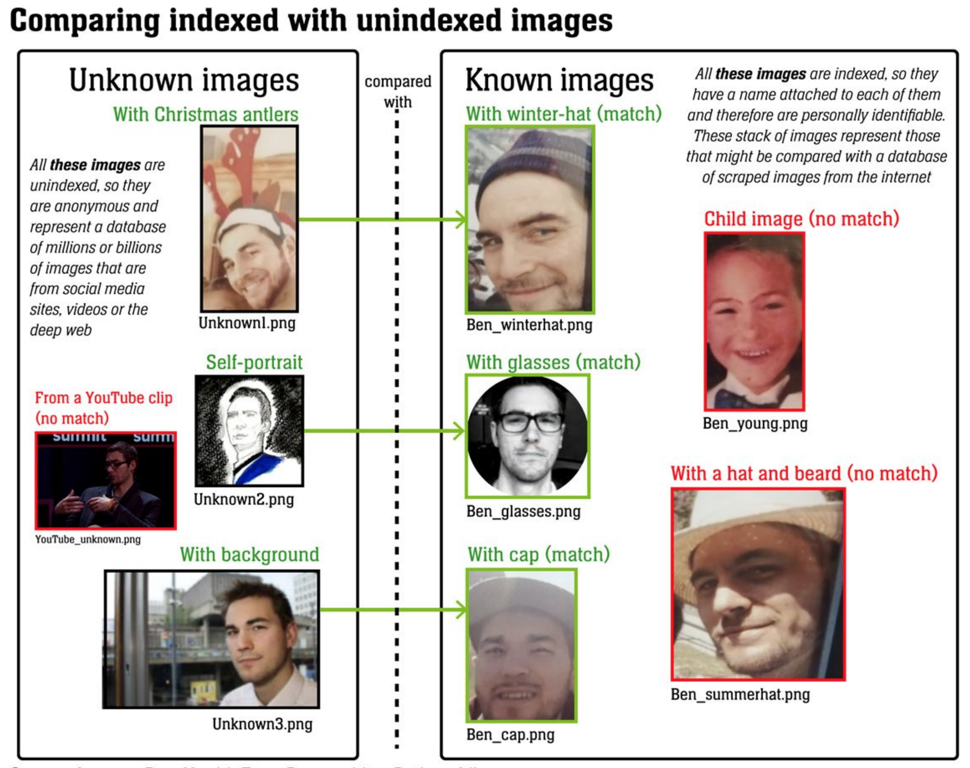
42nd IEEE Symposium on Security and Privacy

Arezoo Rajabi[1], Rakesh Bobba[2], Mike Rosulek[2], Charles V. Wright[3], Wu-chi Feng[3]
University of Washington[1], Oregon State University[2], Portland State University[3]

Introduction

Automated face recognition models can be used for tracking activities and relationships of image sharing platform users [PoPETs2015].

Comparing indexed with unindexed images



Source: Images Ben Heubl; Face Recognition Python Library

Figure 1. Facial-recognition models could endanger our privacy [E&T2020].

Adversarial Perturbation As Image Privacy Defense

- Convolutional Neural Networks (CNNs) are susceptible to adversarial perturbation
- Previously proposed adversarial perturbation-based approaches are not practical for real world applications

Practical Requirements:

Black-box Attack	Users do not know about target CNNs
Low Computational Cost	Users only have a few personal images (family and friends) and limited computational resources
Low Storage Cost	Users do not want to keep a perturbation per image (storage burden)
Recoverability	Users want to recover the original images
Recognizability	Users want to have recognizable images
Compatibility	The proposed approach must be practical on all platforms

Proposed Schemes

Universal Ensemble Perturbation (UEP):

- Uses small CNNs trained only on 10 classes ⇒ Low computational cost
- Trains CNNs locally ⇒ Black-box scheme
- Learns a universal transferable perturbation ⇒ Low storage cost
- Adds perturbation to arc-tangent hyperbolic space of image ⇒ Low loss recovery

$$x_{i,perturbed} = \frac{1}{2}\left(\tanh\left(\operatorname{arctanh}(2 \times (x_i - 0.5)) + \beta \times \delta\right)\right) + 0.5$$

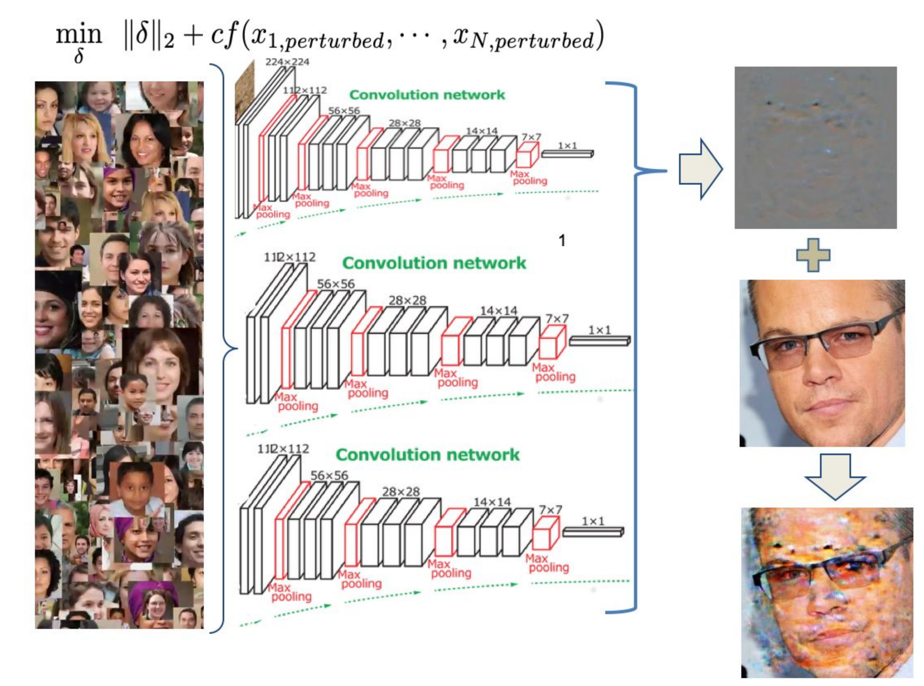


Figure 2. UEP Scheme.

K-Randomized Transparent Image Overlay (k-RTIO):

- Semantic-based adversarial perturbation⇒ Low computational cost
- Uses a secret key and ID of the source image to generate a unique overlay images ⇒ Low storage cost
- Easy to recover ⇒ Reversibility
- No CNNs required for generating perturbations ⇒ Black-box scheme

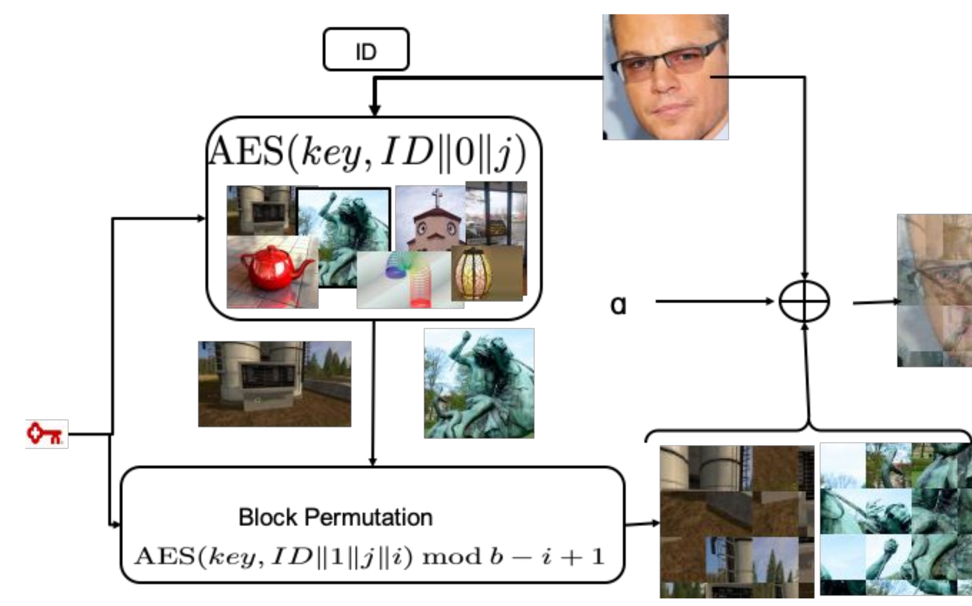


Figure 3. k-RTIO Scheme.

Results

- Dataset: 1000 images sampled from FaceScrub celebrities' face dataset
- Face detection and recognition models
  1. DeepFace [CVPR2014]
  2. Clarifai.com
  3. Google Vision API

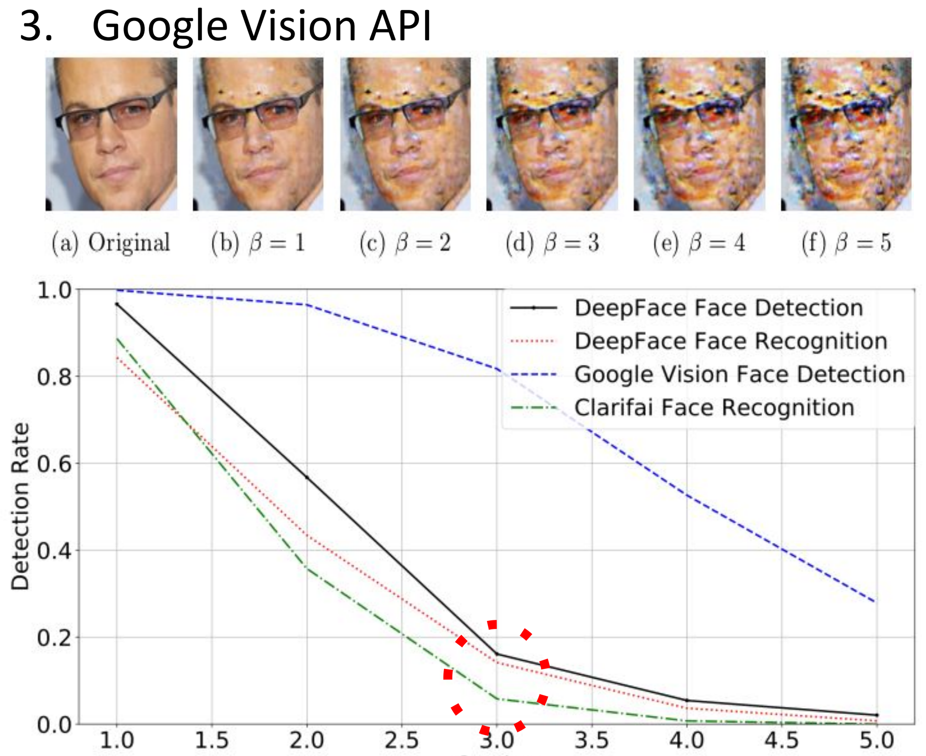


Figure 4. Accuracy of face recognition and detection on perturbed images by UEP

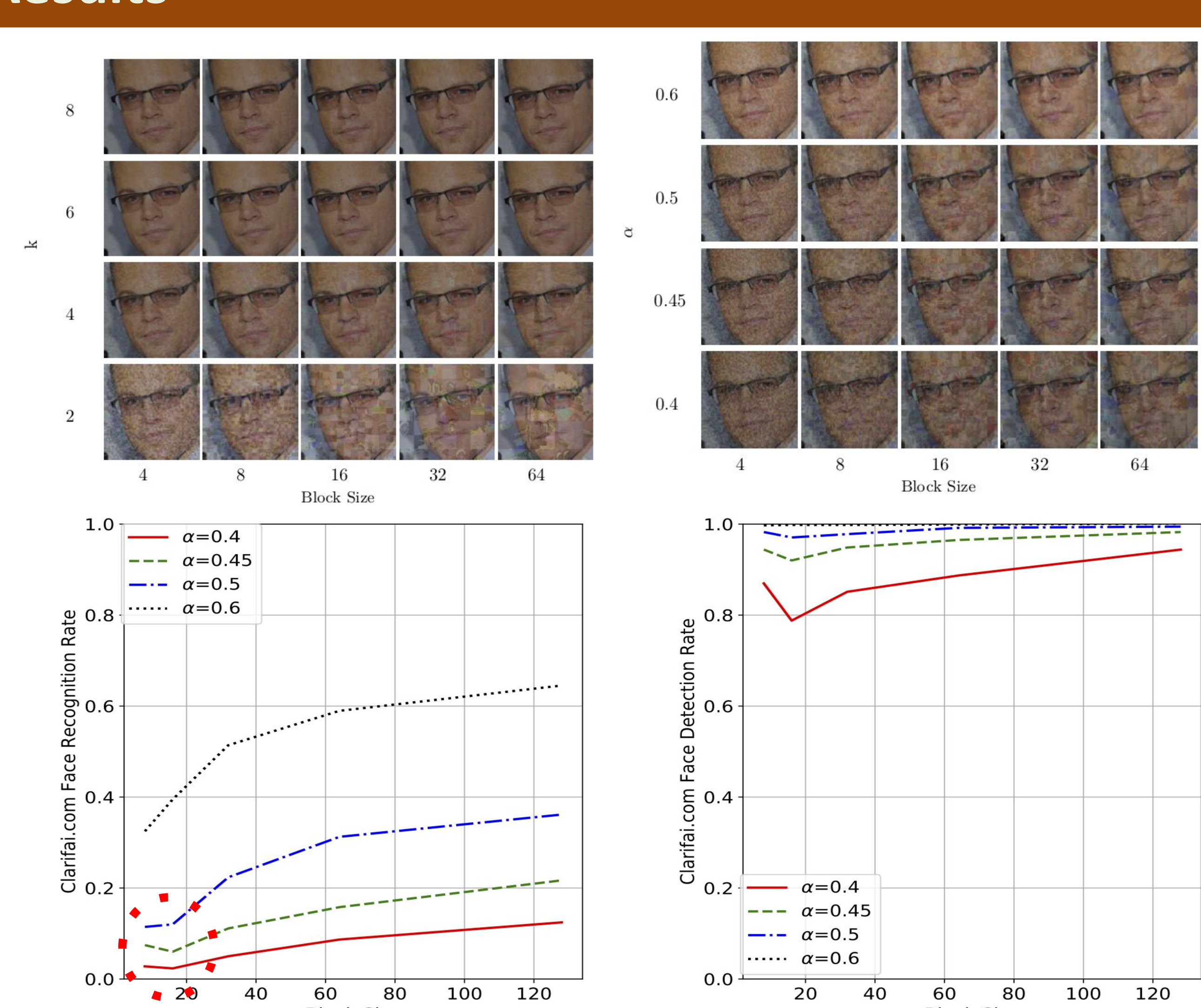


Figure 5. Accuracy of clarifai.com face recognition and detection on perturbed images by k-RTIO

Potential Attacks Against UEP & k-RTIO

- UEP is vulnerable to estimation and removal perturbation method, since it uses a single perturbation for several images.

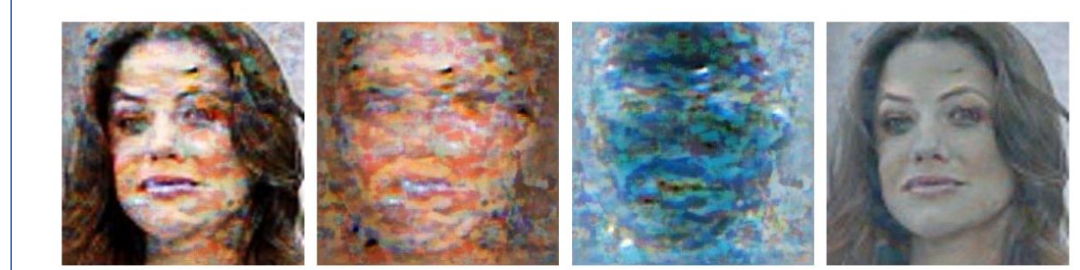


Figure 6. Estimation and removal perturbation method can obtain recognizable images for classifiers

- k-RTIO is robust to filtering methods including estimation and removal perturbation method, since it generates a unique perturbation per image
- The CNNs trained over k-RTIO images can improve their accuracy. But, training robust CNNs is computationally expensive and does not guarantee robustness against all other type of adversarial examples

Conclusions

- Our k-Randomized Transparent Image Overlays can fool well-known face recognition models at least for 85% of the perturbed images
- Our Universal Ensemble Perturbation can fool well-known face recognition models at least for 90% of the perturbed images for $\beta = 4$
- Both UEP and k-RTIO satisfy practical requirements

Future Directions

- Evaluating users/humans' ability of recognizing k-RTIO perturbed faces for specific α, k and block-size values.
- Generating synthetic overlay images instead of using a fixed set of overlay images
- Extending differential-privacy based perturbation approaches which provide strong guarantees for image privacy

Contact Information

Arezoo Rajabi
University of Washington
Email: rajabia@uw.edu

Rakesh Bobba
Oregon State University
Email: rakesh.bobba@oregonstate.edu

References

[S7P2017] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." *IEEE symposium on security and privacy (S&P),* pp. 39-57. IEEE, 2017.
[CVPR2014] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.
[PoPETS2021] Rajabi, Arezoo, Rakesh B. Bobba, Mike Rosulek, Charles Wright, and Wu-chi Feng. "On the (im) practicality of adversarial perturbation for image privacy." *Proceedings on Privacy Enhancing Technologies* (2021).
[PoPETs2015] Shoshitaishvili, Yan, Christopher Kruegel, and Giovanni Vigna. "*Portrait of a privacy invasion: Detecting relationships through large-scale photo analysis*." *Proceedings on Privacy Enhancing Technologies* 2015.1 (2015): 41-60.
[E&T2020] Heubl, Ben." How online facial-recognition systems could endanger our privacy." *E&T Engineering and Technology*. August 2010.
[CVPR2017] Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Acknowledgements

This work has been published in Privacy Enhancing Technologies Symposium (PoPETS) 2021 when Arezoo Rajabi was PhD student at Oregon State University