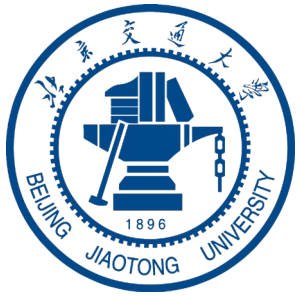
Benign Adversarial Attack: Adversarial Privacy-preserving

Jiaming Zhang

Co-Contributors: Xian Zhao, Shangxi Wu, Zhiyu Lin, Xiaowen Huang, Jitao Sang

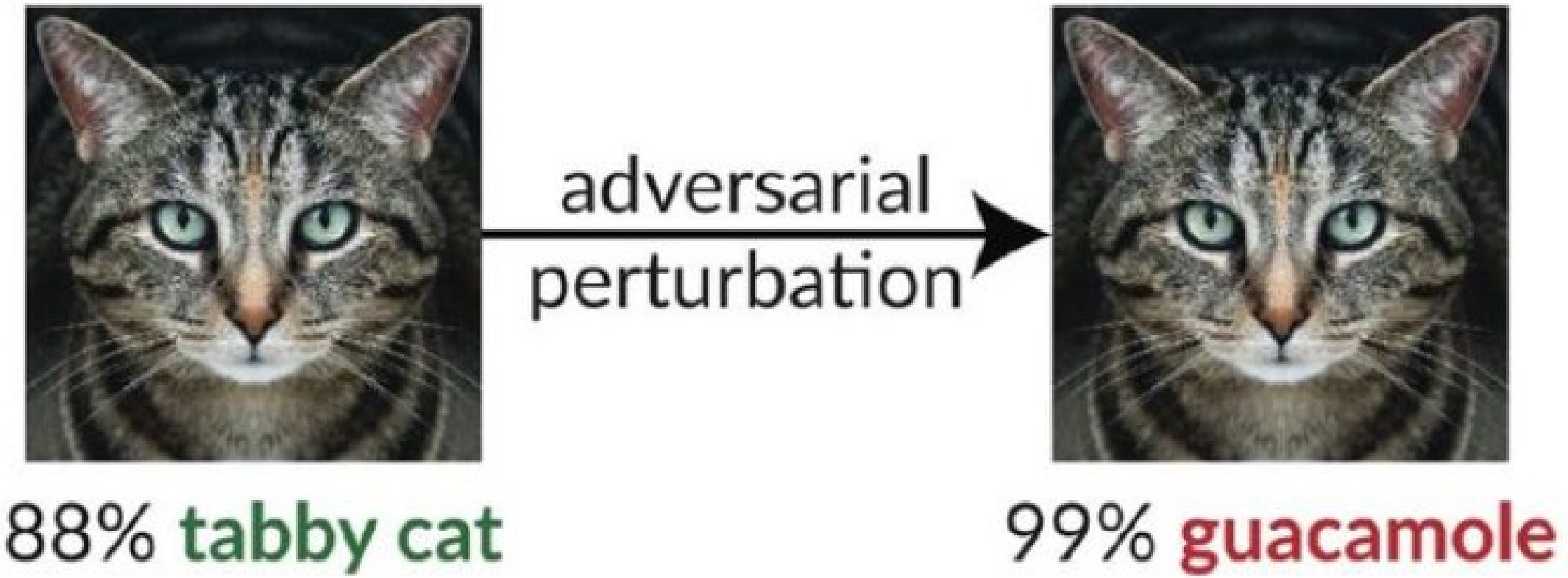
Beijing Jiaotong University

jiamingzhang@bjtu.edu.cn

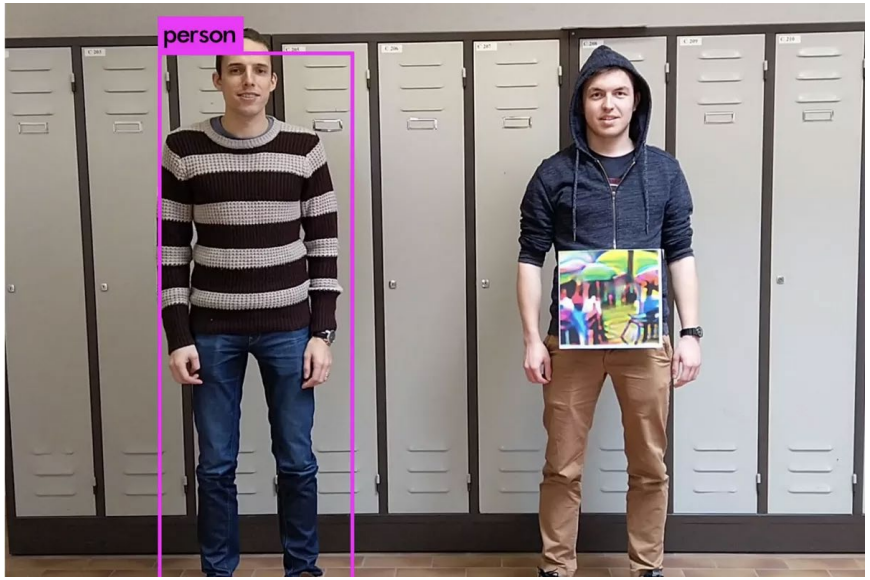


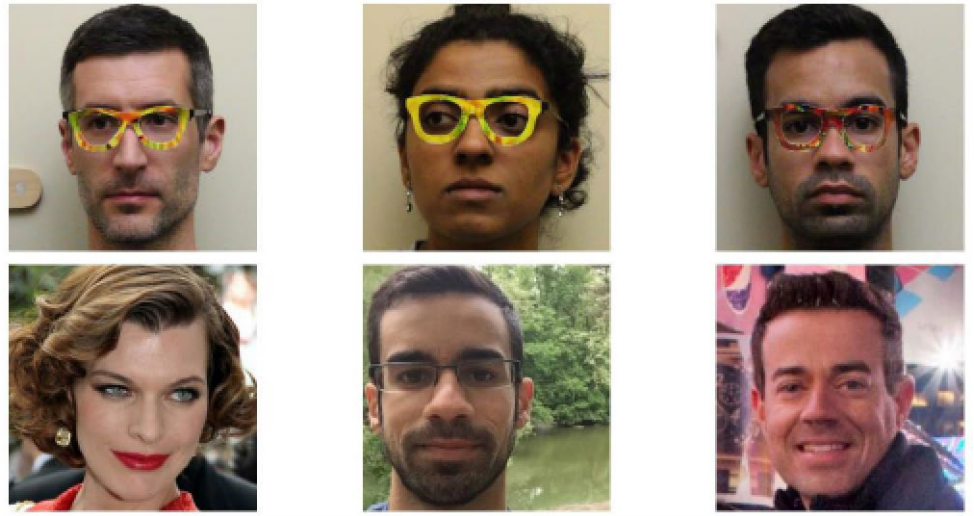
What is Adversarial Example?

Imperceptible to human eyes, but fool the deep leaning model.



Devil: Adversarial Example





Adversarial attack to traffic signs. This is a great threat to autopilot. [1]

Adversarial attack to impersonate a person by a brand.[2]

Adv-glasses: generate adversarial glasses to impersonation or dodging.[3]

[1] Lu et al. "NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles." , 2017.
[2] Thys et al. "Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection." Workshops (CVPRW) , 2019.
[3] Sharif et al. "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition." ACM SIGSAC, 2016.

Adversarial Example Research

2014

Szegedy first proposed adversarial example[1]:

- To algorithm
- Imperceptible

2015

Goodfellow proposed FGSM:
Fast Gradient Sign Method [2]

2015

Adversarial Defense:
Adding adversarial example into training set[2]

2017

Iterative-BIM:
Stronger adversarial example[3]

2017

Physical Adversarial Perturbation:
Adversarial example in 3D world [9]

2018

Stochastic Activation:
Destroy gradient[4]

2018

Madry proposed **Adv-Training** [5]:
Min-max optimization

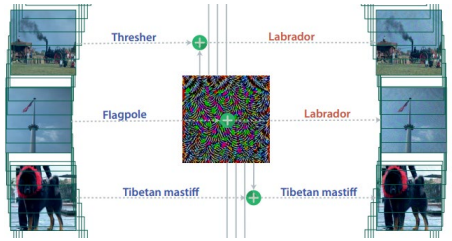
$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|x'-x\|\leq\epsilon} \mathcal{L}(f_\theta(x'), y) \right]$$

2019

Obfuscated Gradients (*Carlini*,[6]):
Breaking 7 defenses, except Adv-Training

2019

Universal Adversarial Perturbation:
Image-independent noise[10]



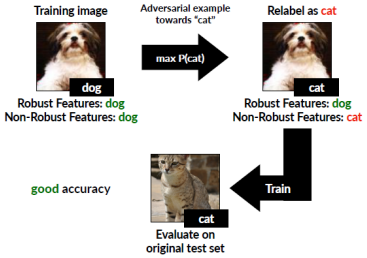
2020

Adv-Training Extend [7]:
Adaptive evaluation on robustness

2020

Unreliable Defense (*Carlini*,[8]):
No universal defense

Adversarial perturbation is feature (*Ilyas*,[11]):
Adversarial features can be generalized

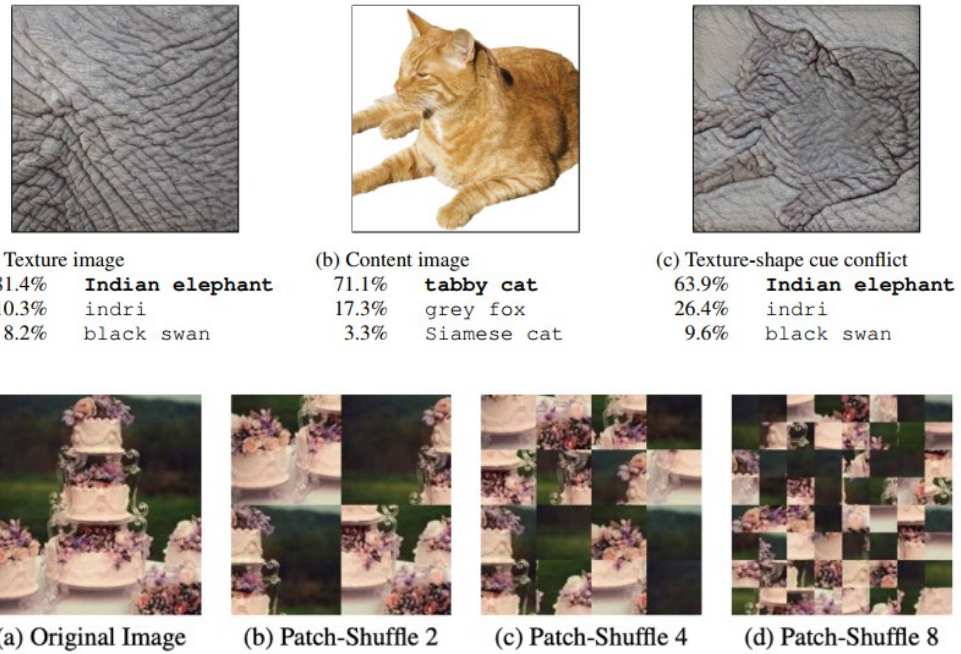


[1] Intriguing properties of neural networks. ICLR 2014
[2] Explaining and harnessing adversarial examples, ICLR 2015
[3] Adversarial examples in the physical world, ICLR 2017
[4] Stochastic activation pruning for robust adversarial defense, ICLR 2018.
[5] Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018.
[6] Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, ICML 2018 Best Paper
[7] Resisting adversarial attacks by k-winners-take-all, ICLR 2020
[8] On Adaptive Attacks to Adversarial Example Defenses, arXiv, 2020
[9] Adversarial examples in the physical world. ICLR 2017.
[10] Universal adversarial perturbations. CVPR 2017.
[11] Adversarial Examples Are Not Bugs, They Are Features, NeurIPS 2019

Human VS Algorithm

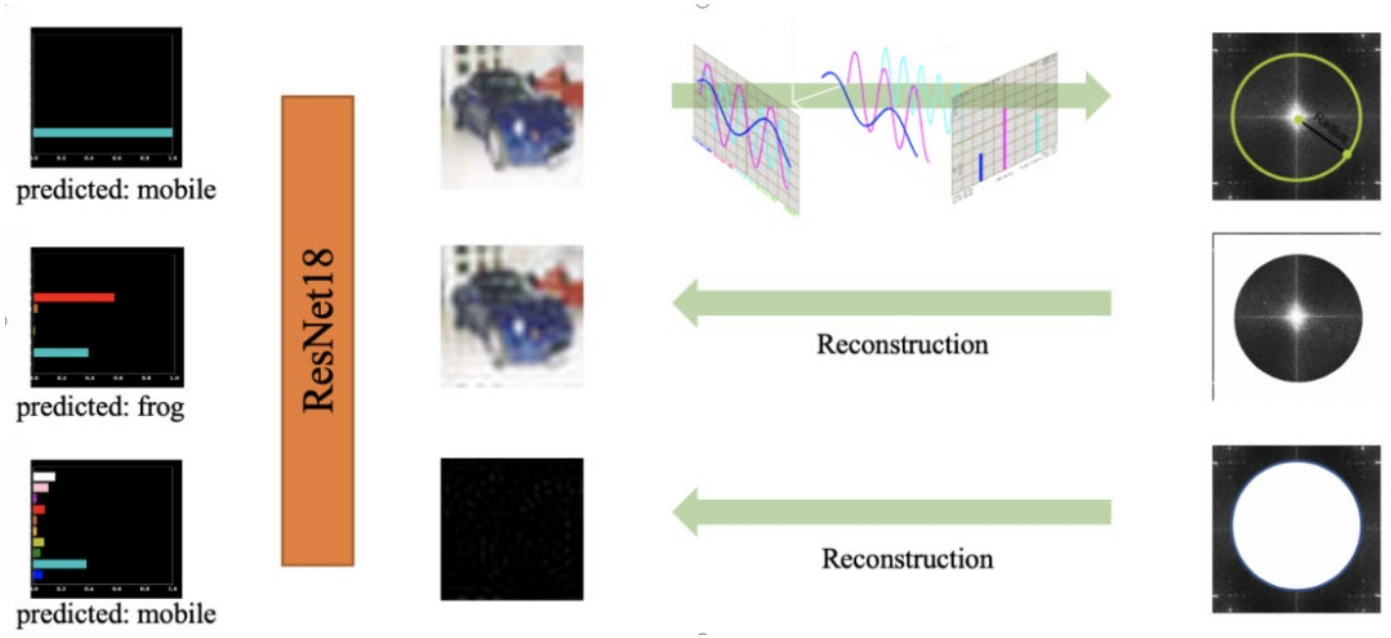
Algorithm is the knowledge distillation to human:
Human label data, algorithm learn from data.

Algorithm: texture vs Human: shape



The algorithm relies more on texture[1]

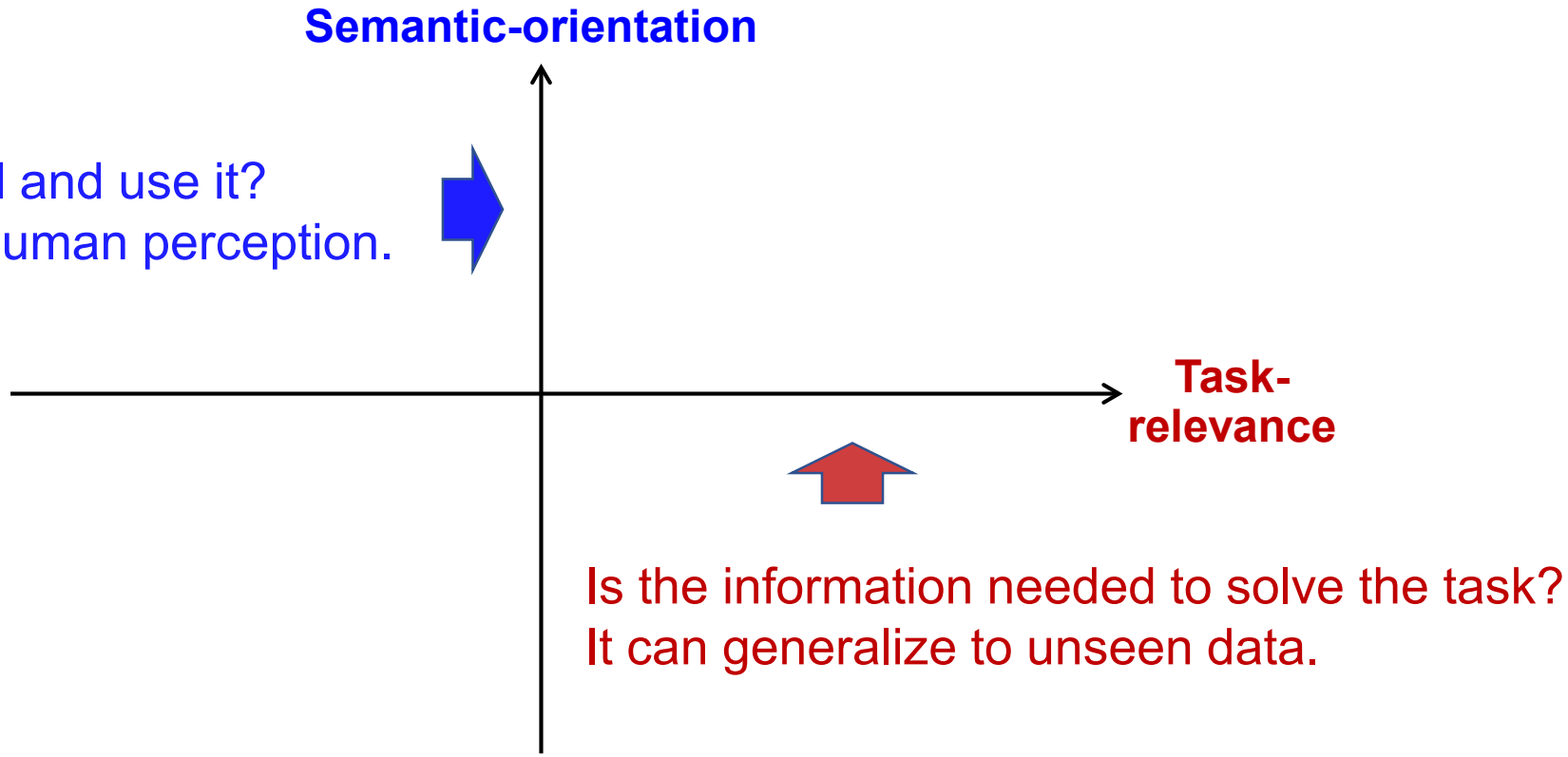
Algorithm: high frequency vs Human: low frequency



The algorithm can use high frequency features[2]

[1] ImageNet-trained CNNs are biased towards texture. ICLR 2019
[2] High Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. CVPR 2020.

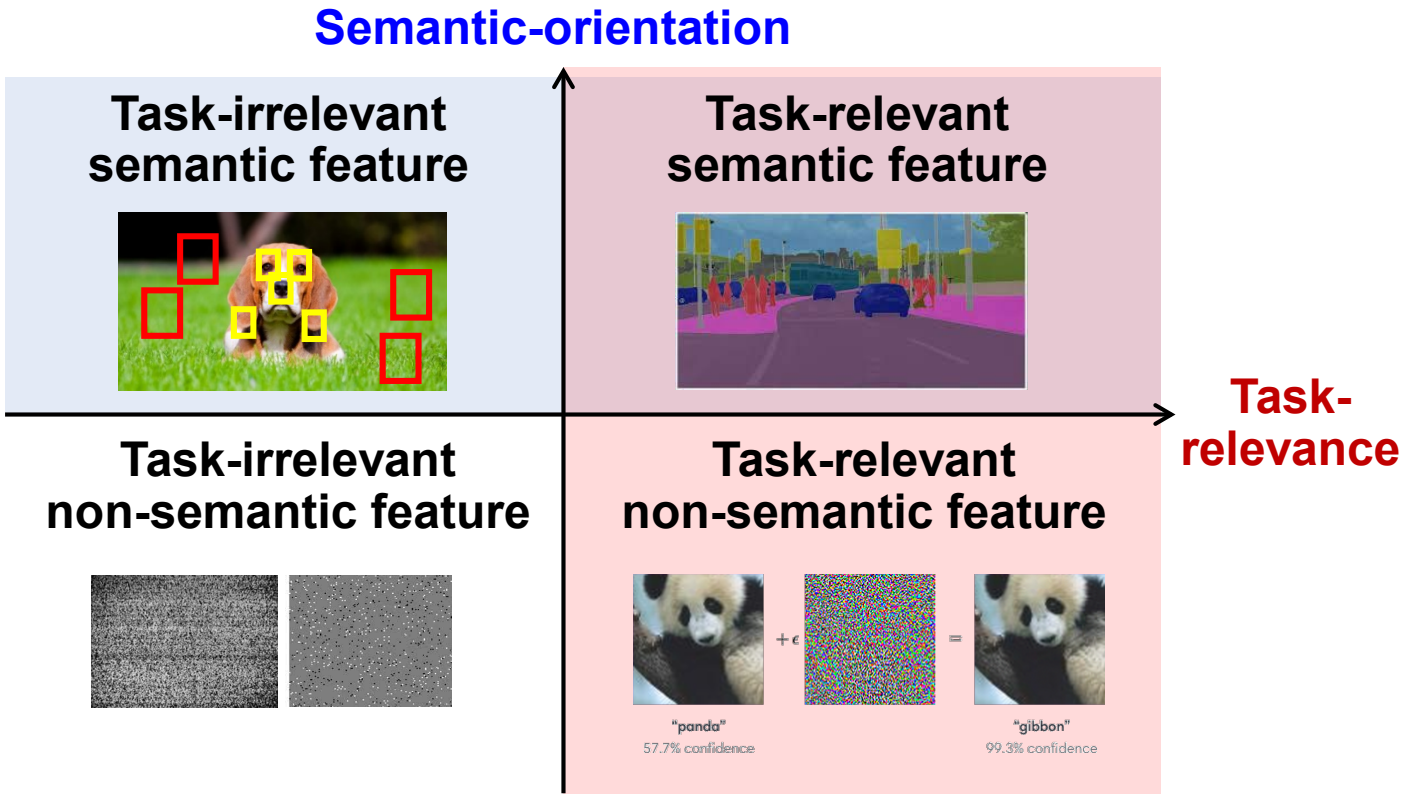
Task-Semantic Coordinate



Can humans understand and use it?
Perturbation can affect human perception.

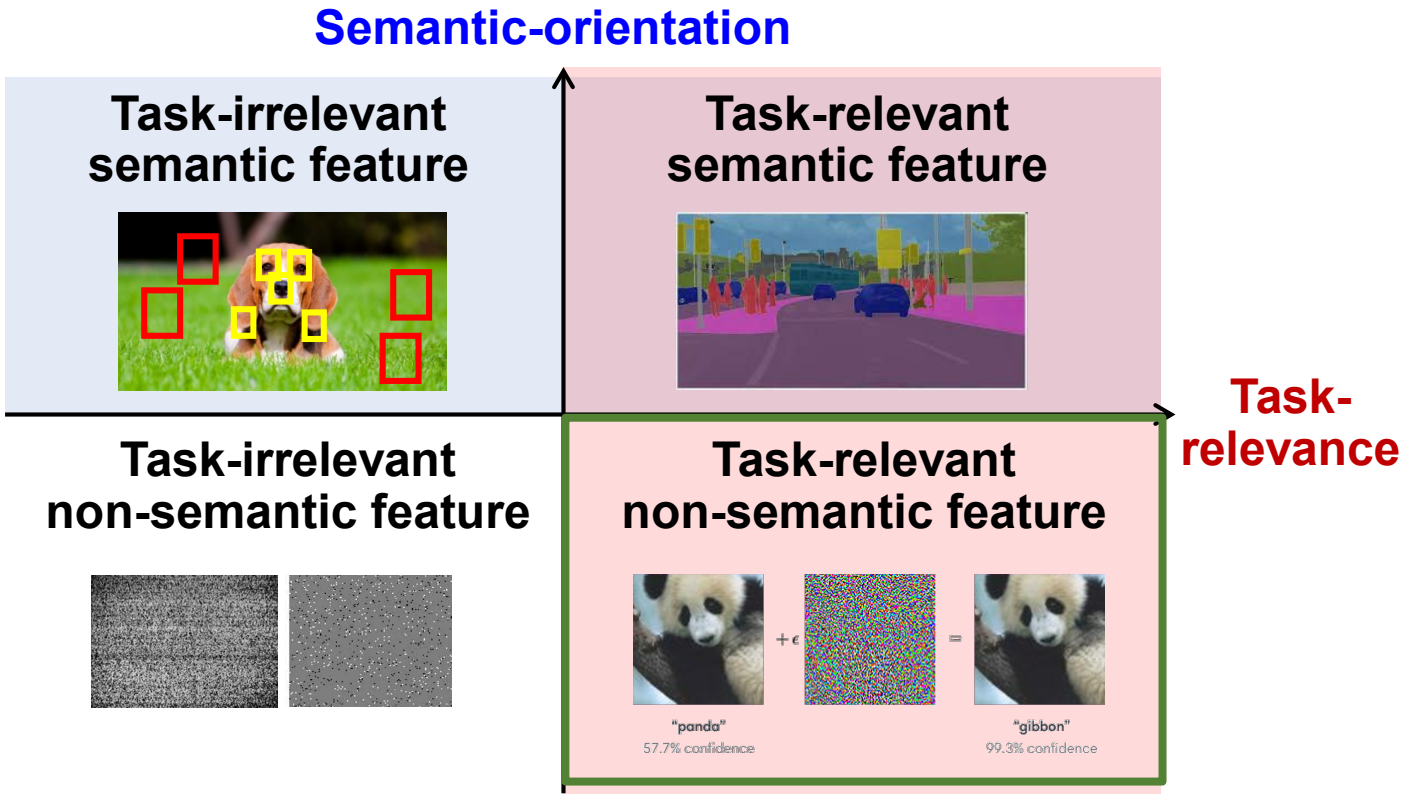
Is the information needed to solve the task?
It can generalize to unseen data.

Task-Semantic Coordinate



Task-Semantic Coordinate

Task-relevant non-semantic feature is exclusive to algorithm:
Algorithm can exploit Task-relevant non-semantic feature and be influenced by it.



Characteristics of Adversarial Example

☐ Utilizable as feature

- ➢ Classical machine learning focuses on semantic feature utilization.
- ➢ From the perspective for solving tasks, non-semantic features play a complementary role.

⇨ **Adversarial data augmentation**

☐ Exclusive to algorithm

- ➢ Machine learning algorithms solve tasks using both semantic and non-semantic information.
- ➢ Humans utilize semantics to solve tasks.

⇨ **Adversarial Turing Test**

☐ Inevitable for vulnerability

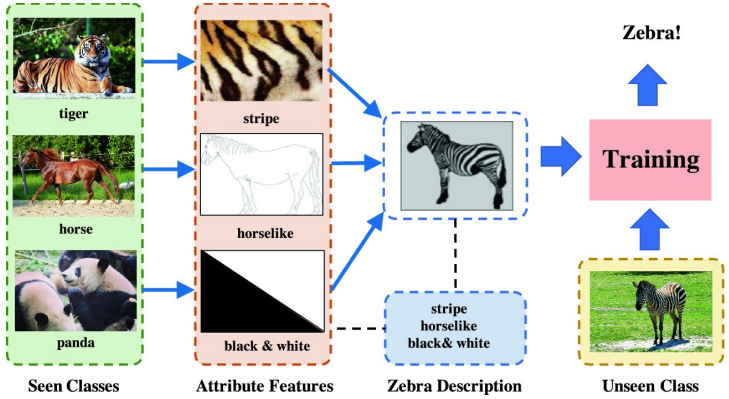
- ➢ Training mechanisms of machine learning cannot prevent algorithms from learning non-semantic features.
- ➢ Non-semantic perturbation of inputs affects the inference of algorithm.

⇨ **Privacy-preserving (Rejecting malicious algorithm)**

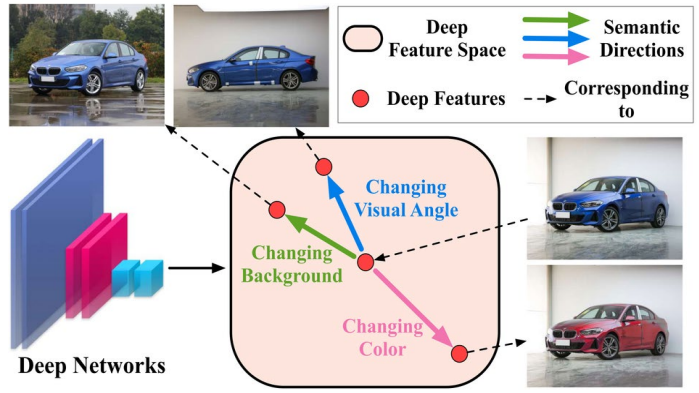
Adversarial data augmentation

Adversarial Data Augmentation

Solving data hunger with semantic information.

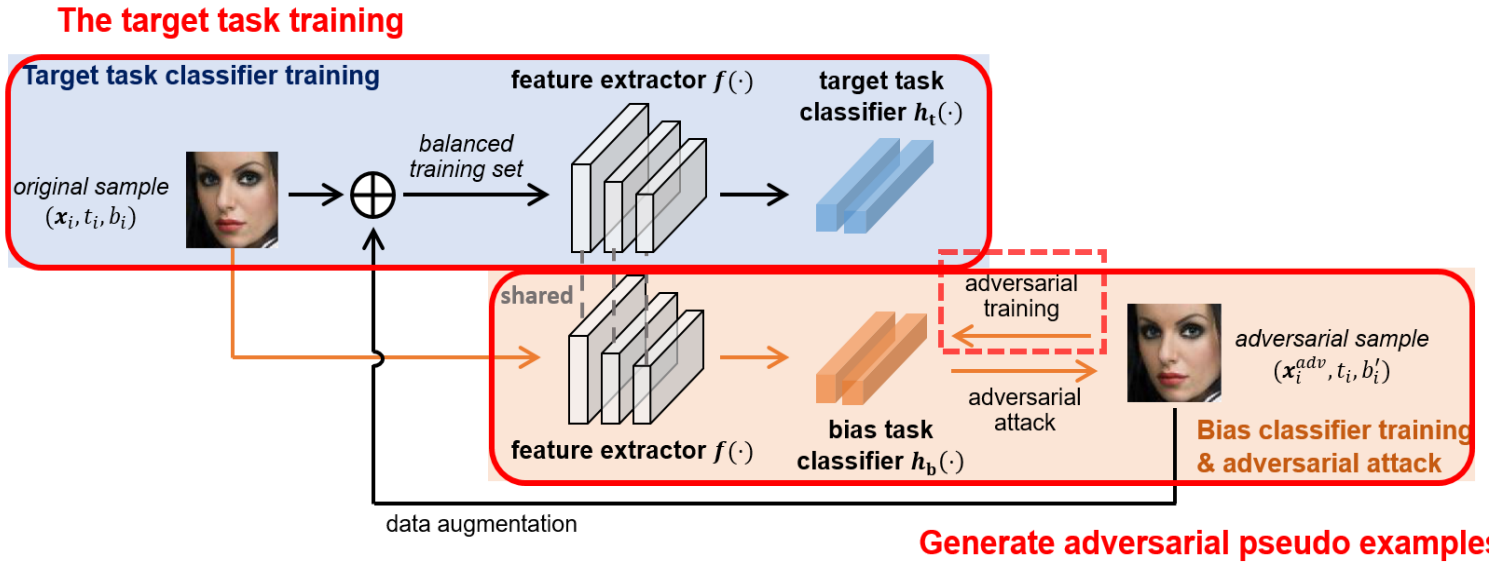


transfer semantic features

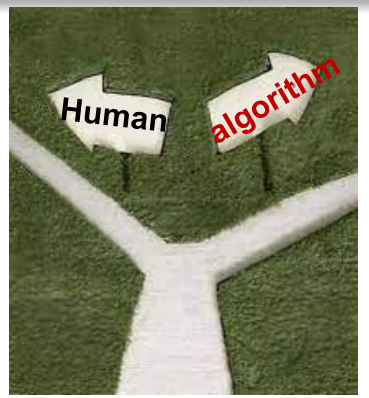


semantic data augmentation

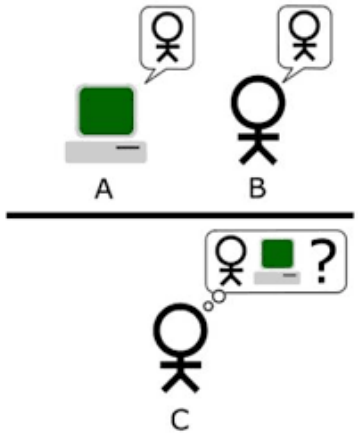
Adversarial example can also provide useful features for data augmentation.



Adversarial Turing test

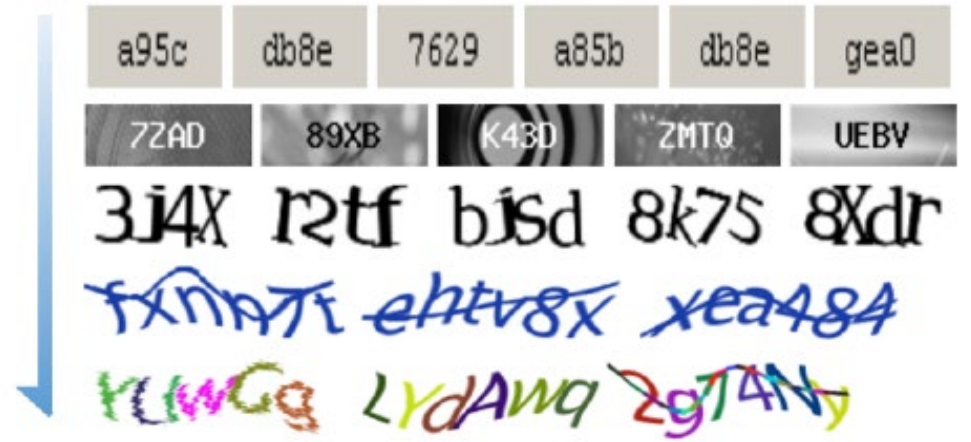


Motivation



Turing Test

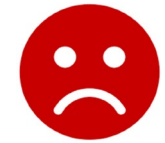
powerful computer vision technology



CAPTCHA

adversarial attack

Malicious algorithm

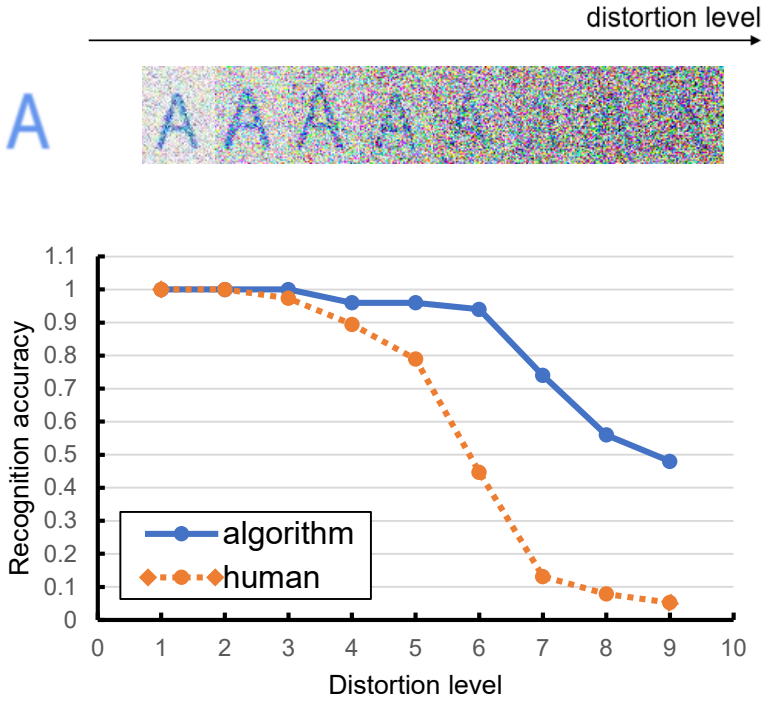


Adversarial Turing test:

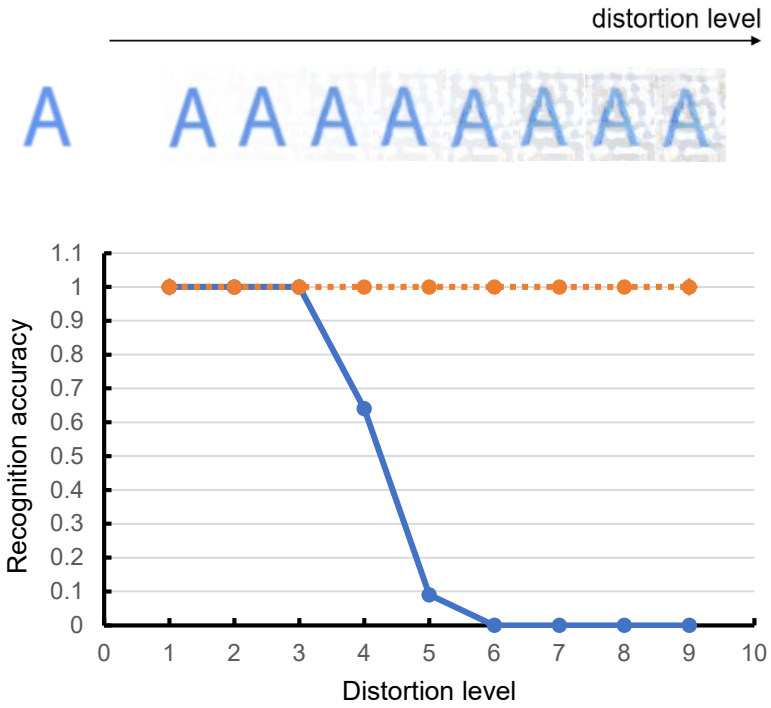
The Turing Test task is adjusted based on adversarial attack, and the original algorithm is invalid.

Data Analysis

Adversarial perturbation can tell the different sensitiveness from human and algorithm.



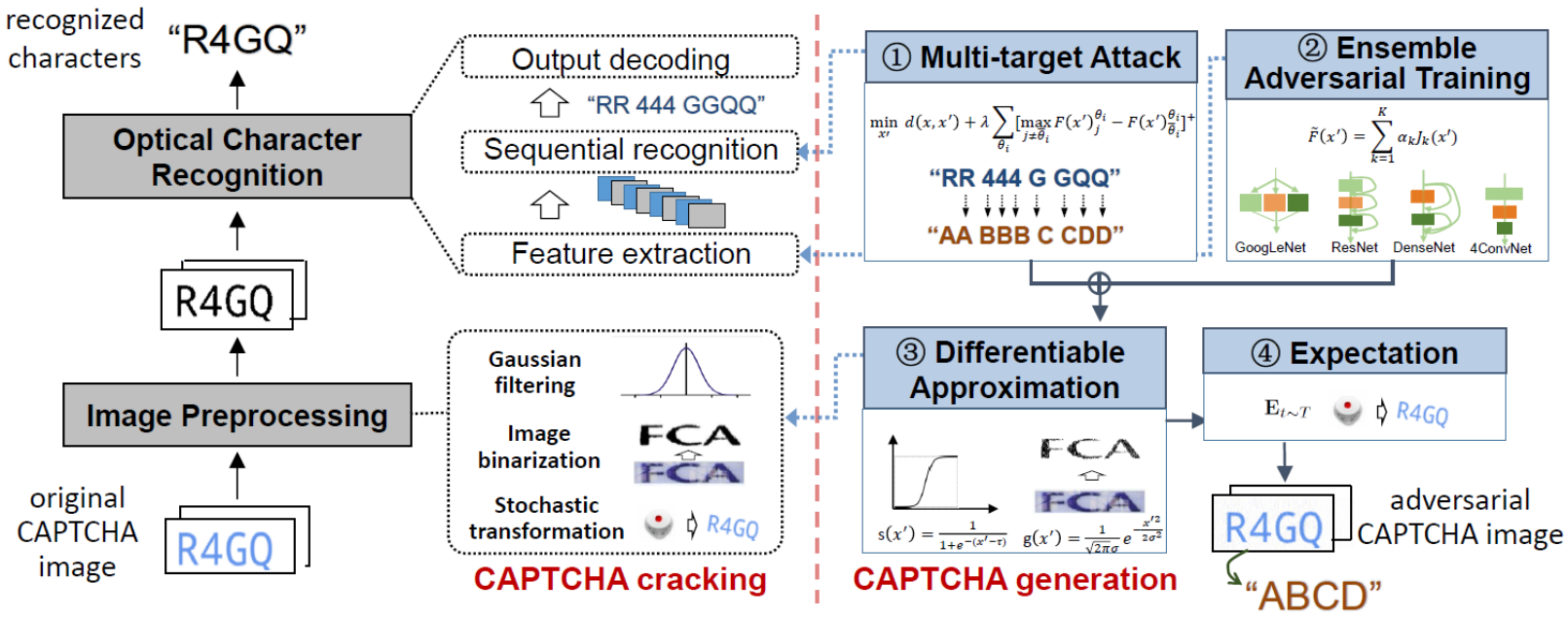
White Gaussian Noise



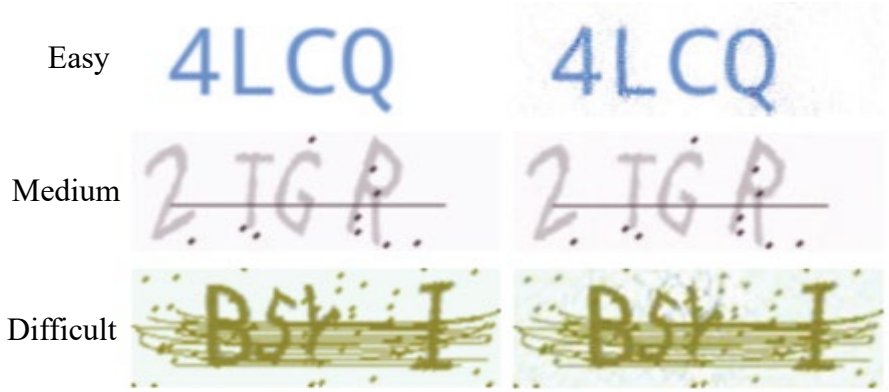
Adversarial Perturbation

[1] Zhang et al. "Robust CAPTCHAs towards Malicious OCR." IEEE Transactions on Multimedia (2020).

Prototype Application



The proposed robust CAPTCHA designing framework



Model vs. Human

		Original images	Adversarial images
Easy	*algorithm*	100%	0%
	human	99%	94%
Medium	*algorithm*	91%	0%
	human	73%	65%
Difficult	*algorithm*	81%	4%
	human	56%	49%

Zhang et al. "Robust CAPTCHAs towards Malicious OCR." IEEE Transactions on Multimedia (2020).

Privacy-preserving

Motivation

User sharing is the main source of face images

Pinterest MediaFire imgur imgfave SmugMug DEVIANT ART 500PX tinypic

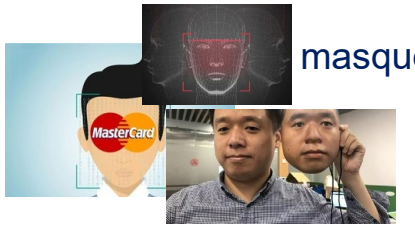
Malicious algorithms using human face images

- DeepFake raises ethical questions.
- Kneron tested that widely used face payment systems like AliPay and WeChat can be fooled by masks and face images.
- Clearview AI crawls face images from Facebook, YouTube and other websites, constructs a database containing more than three billion images, and provides it to 600 law enforcement agencies in the United States.
- The European Parliament passed a bill banning face recognition in public

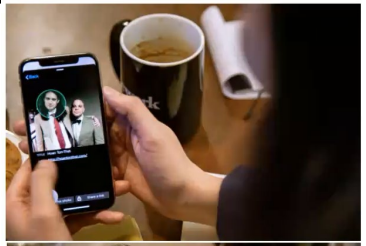
Face images are exposed

face images leakage (face recognition)

swap face



masquerade



collect data

Adversarial Privacy-preserving Filter

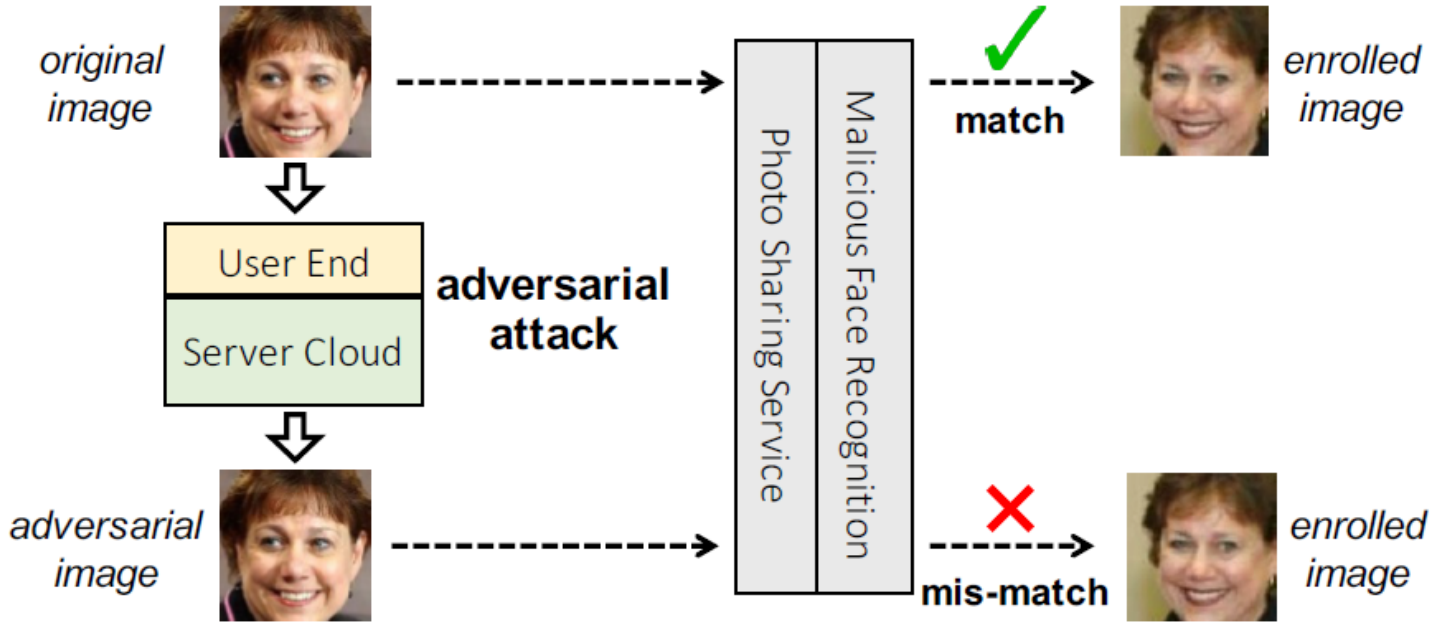
Requirements

- **Privacy**: Unable to identify the user from the shared face image.
- **Utility**: Maintain the original quality of images without affecting sharing.

Adversarial example characteristics

- **Fool algorithm**: Disable malicious face recognition algorithms.
- **Imperceptible:** Adversarial perturbation is almost invisible to human.

- **Non-accessibility**: Only the user device has access to the original image
  - → Absolute data security
  - ✓ Adversarial attack requires a lot of computation and is usually carried out on the server cloud.
  - ✓ Exposure of original images to the cloud increases the risk of leaks.
  - ✓ **End-cloud collaborated:** The user end obtains the image gradient on the small model, and the cloud enhances the gradient through the large model.



Adversarial Privacy-preserving Filter

- Obtain Image-specific gradient on the probe model: $g = \epsilon \cdot sign(\nabla x I(x, x_e))$
- Compatible with different adversarial attack methods
- To transfer the image gradient from the probe model to the server model by transfer network T : $\hat{g} = T(g)$
- Transferred gradient $\hat{g}$ is effective against potentially malicious face recognition models

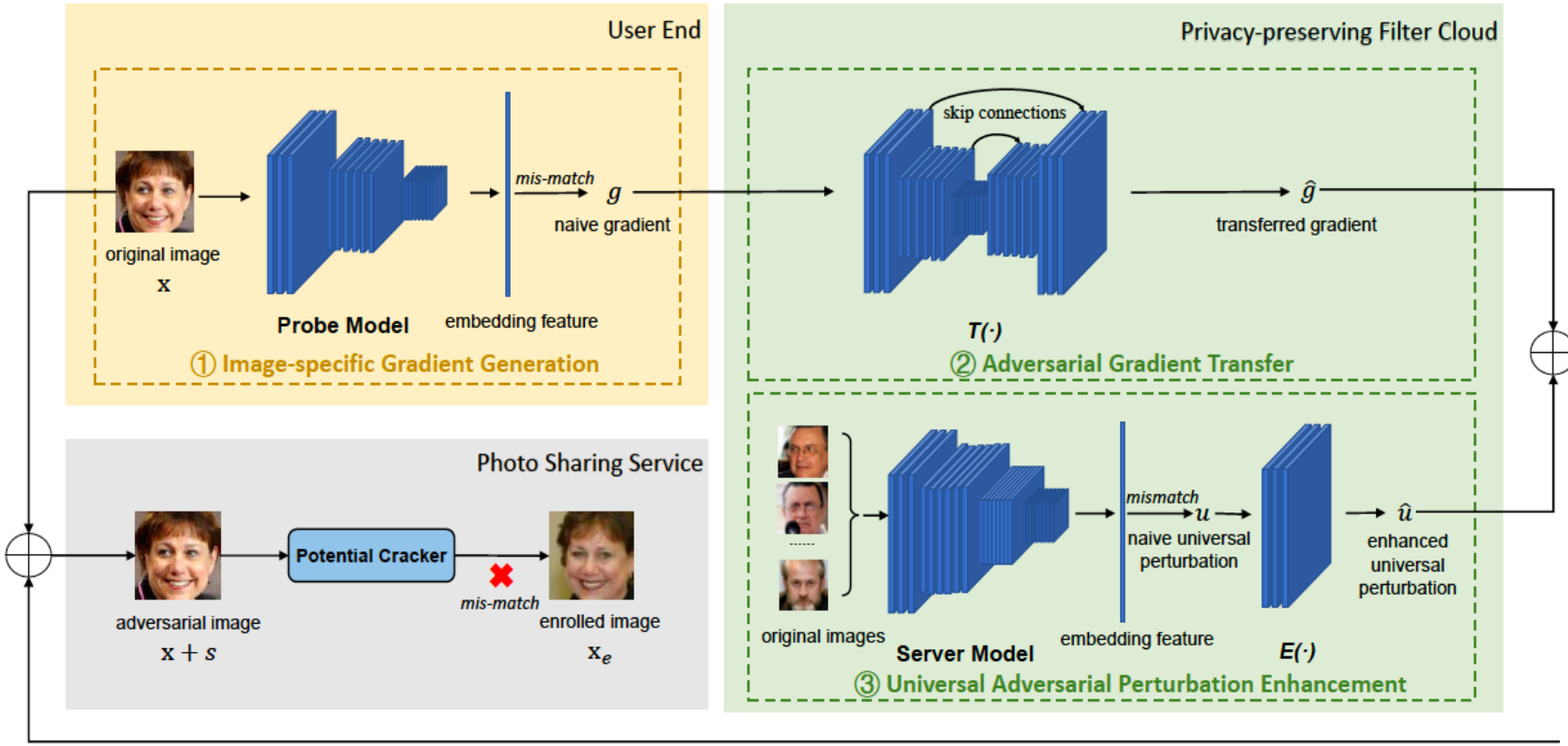


Image-independent universal adversarial perturbation $\hat{u}$:

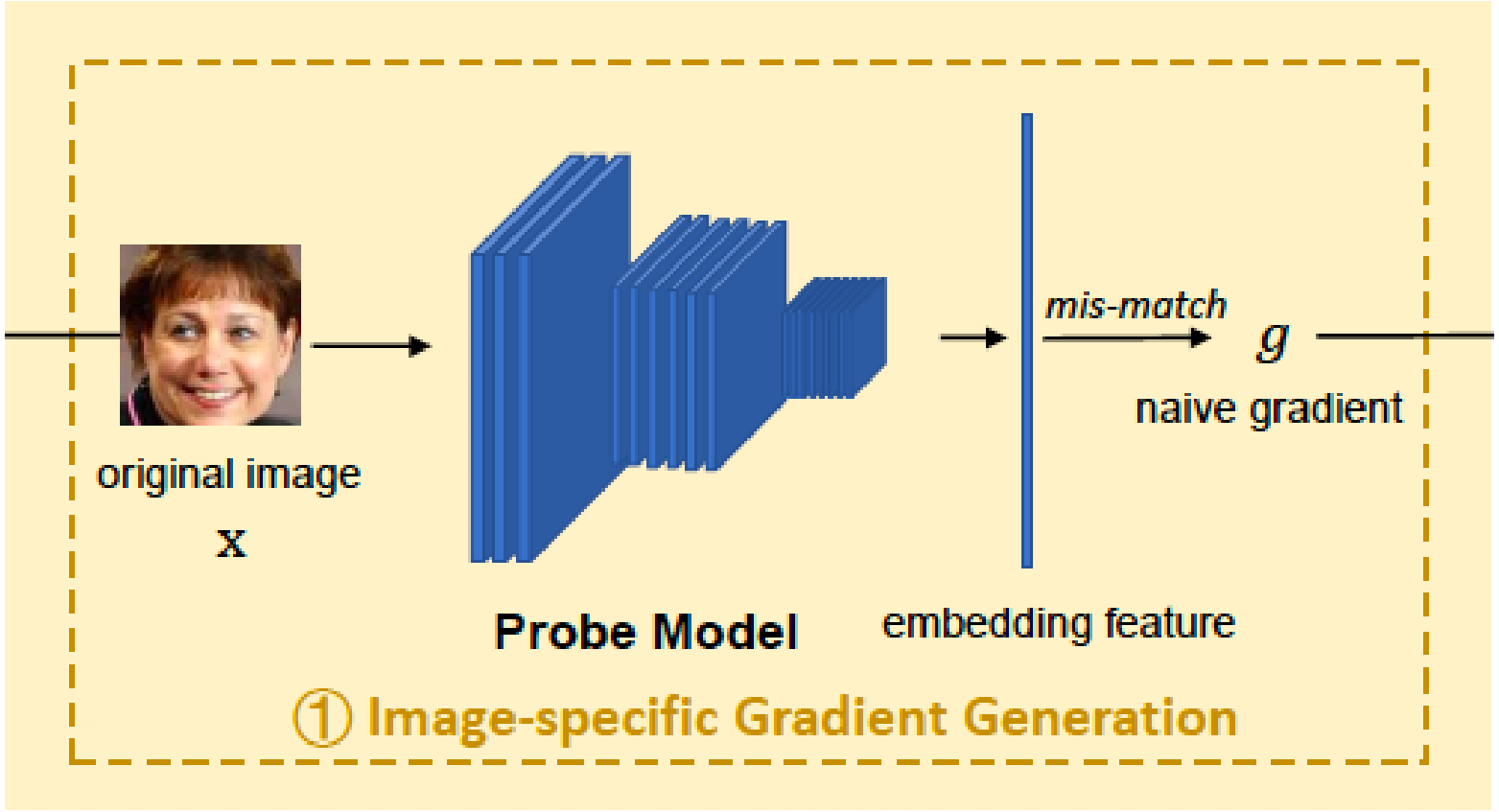
- Can be combined with image-specific gradient
- Support the training of transfer network T (accelerate convergence)

Zhang et al. "Adversarial privacy-preserving filter." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Method-Image-specific Gradient

- To extract image-specific adversarial gradient g
- Different adversarial attack algorithms are allowed

$$g = \epsilon \cdot sign(\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{x_e}; \theta))$$

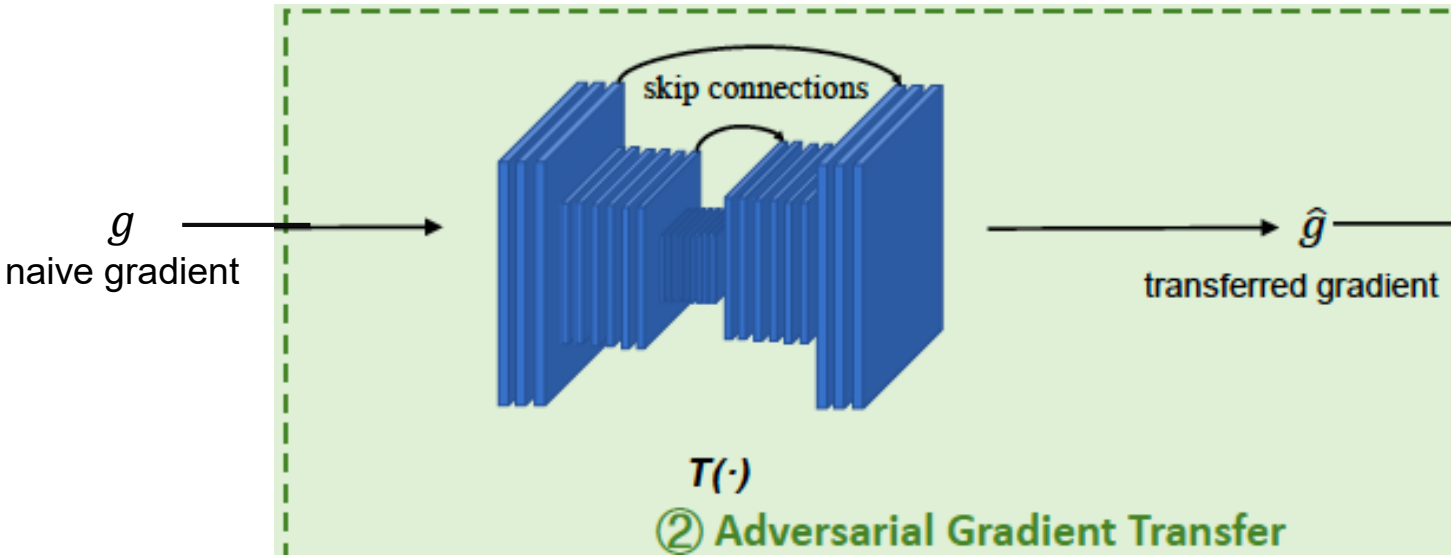


Zhang et al. "Adversarial privacy-preserving filter." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Method-Adversarial Gradient Transfer

- To transfer the image gradient from the probe model to the server model.

$$\min_{\theta_T} \|\hat{g} - \tilde{g}\|_2$$



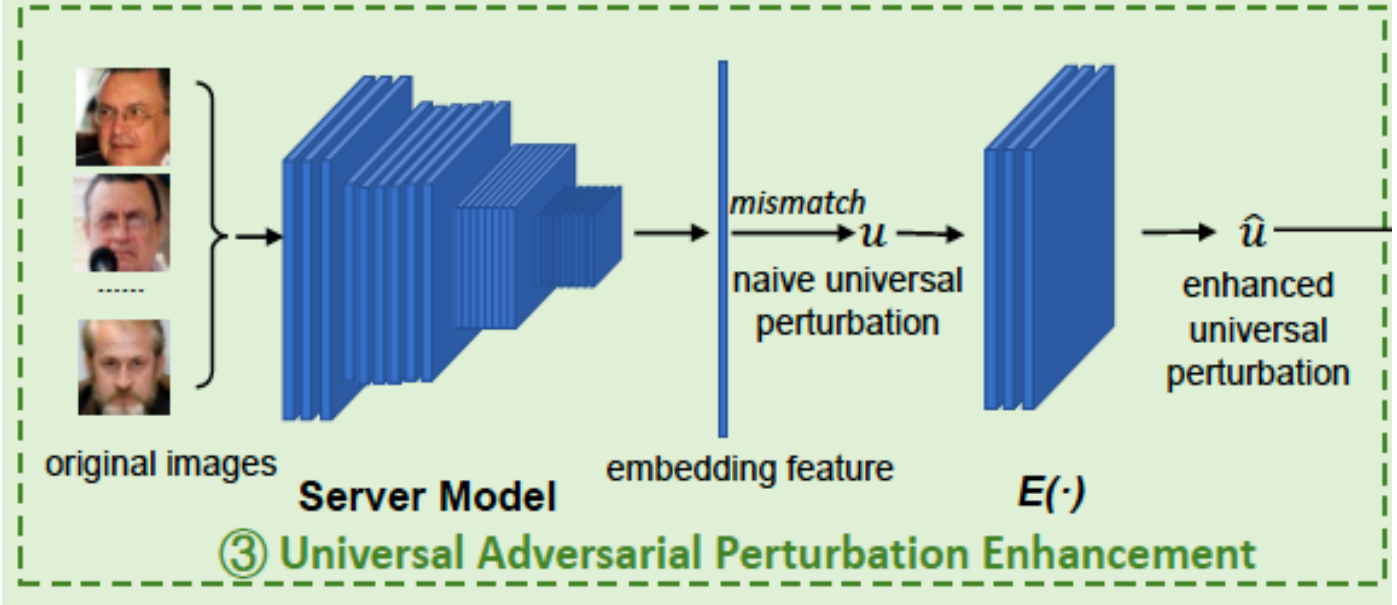
Zhang et al. "Adversarial privacy-preserving filter." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Method-Universal Adversarial Perturbation

- To enhance the performance of adversarial perturbation, we integrate the image-specific information and image-independent information.

$$u = \max_{u} \sum_{i=1}^{n} \sum_{j=1}^{n} d(f_\theta(\mathbf{x}_i + u), f_\theta(\mathbf{x}_j))$$

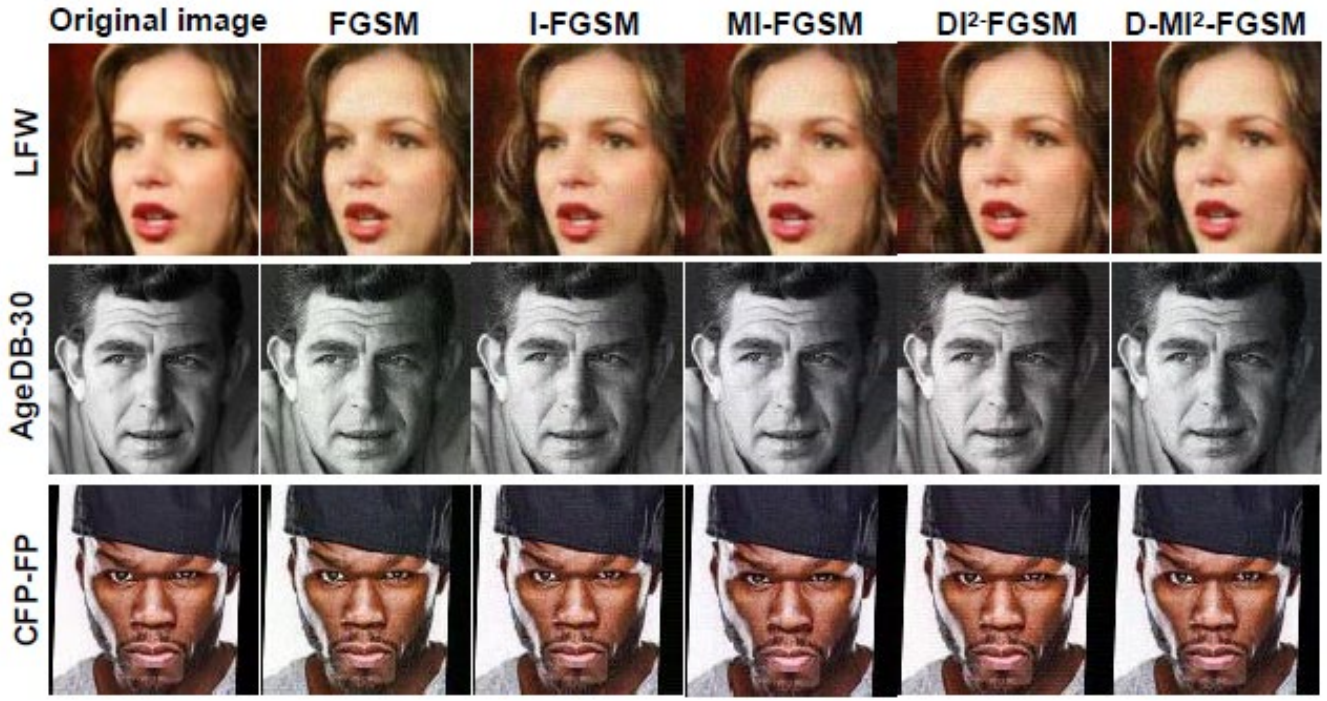
$$\hat{u} = E(u) = \mathrm{conv}(\beta \cdot u + \gamma)$$



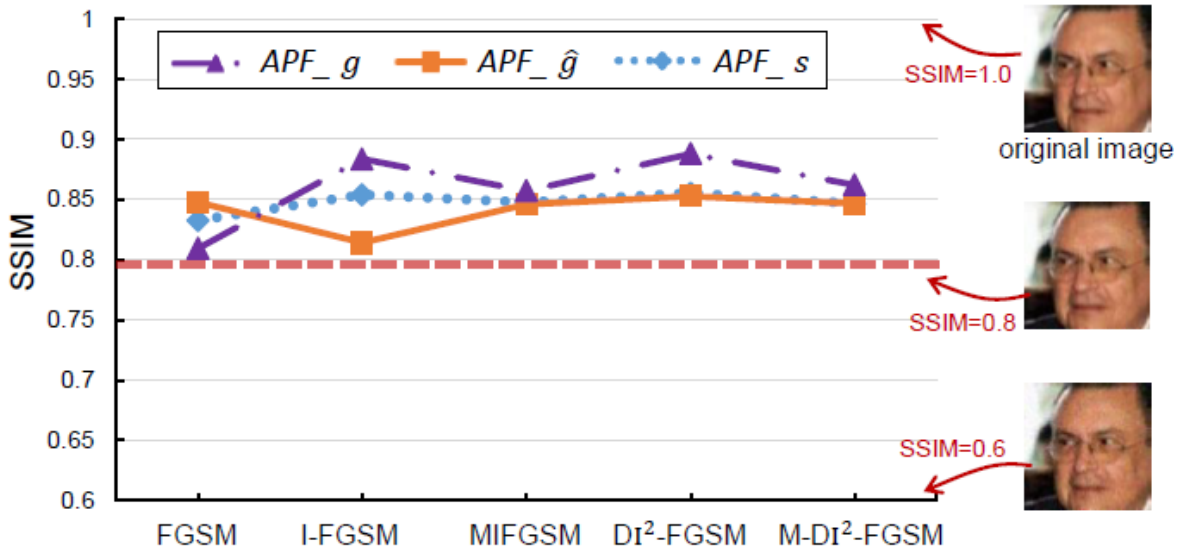
Zhang et al. "Adversarial privacy-preserving filter." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Experimental Results

Face filters can preserve the quality of the image.



The example adversarial images based on APF_s



The SSIMs between the adversarial images and original images

Zhang et al. "Adversarial privacy-preserving filter." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Experimental Results

The ASRs on LFW, AgeDB-30 and CFP-FP datasets

Datasets	Adversarial Images	FGSM	I-FGSM	MI-FGSM	DI^2-FGSM	M-DI^2-FGSM
LFW	*APF_g*	74.5%	86.8%	88.2%	92.4%	89.1%
	APF_ĝ	91.9%	95.4%	89.8%	96.9%	96.5%
	APF_s	94.8%	97.4%	95.7%	**98.8%**	**98.8%**
	Images_u	-	-	-	-	-
	original image accessible	98.5%	99.4%	99.4%	99.4%	99.27%
AgeDB-30	*APF_g*	81.7%	86.3%	88.1%	90.5%	88.6%
	APF_ĝ	82.3%	90.8%	90.8%	94.9%	92.8%
	APF_s	88.3%	93.4%	93.8%	**95.5%**	94.7%
	Images_u	-	-	-	-	-
	original image accessible	95.8%	96.0%	96.0%	96.0%	96.0%
CFP-FP	*APF_g*	48.6%	57.2%	63.8%	68.3%	65.0%
	APF_ĝ	51.8%	72.8%	74.9%	84.7%	78.1%
	APF_s	67.4%	79.6%	82.8%	**88.3%**	85.3%
	Images_u	-	-	-	-	-
	original image accessible	92.5%	93.7%	90.8%	93.2%	93.4%

- ***APF_s* achieves best performance:** proves the validity of each.
- **Compatible with ensemble learning:** improves the black box attack performance.

The black-box ASRs

Training Models	Testing Models			
	FaceNet	SphereFace	ArcFace	Average
raw	83.0%	50.2%	92.5%	75.1%
FaceNet	95.2%	75.6%	98.3%	89.6%
SphereFace	91.5%	85.6%	96.7%	91.1%
ArcFace	93.8%	73.3%	98.8%	88.5%
Ensemble(F+S)	95.1%	81.8%	**98.6%**	**91.7%**

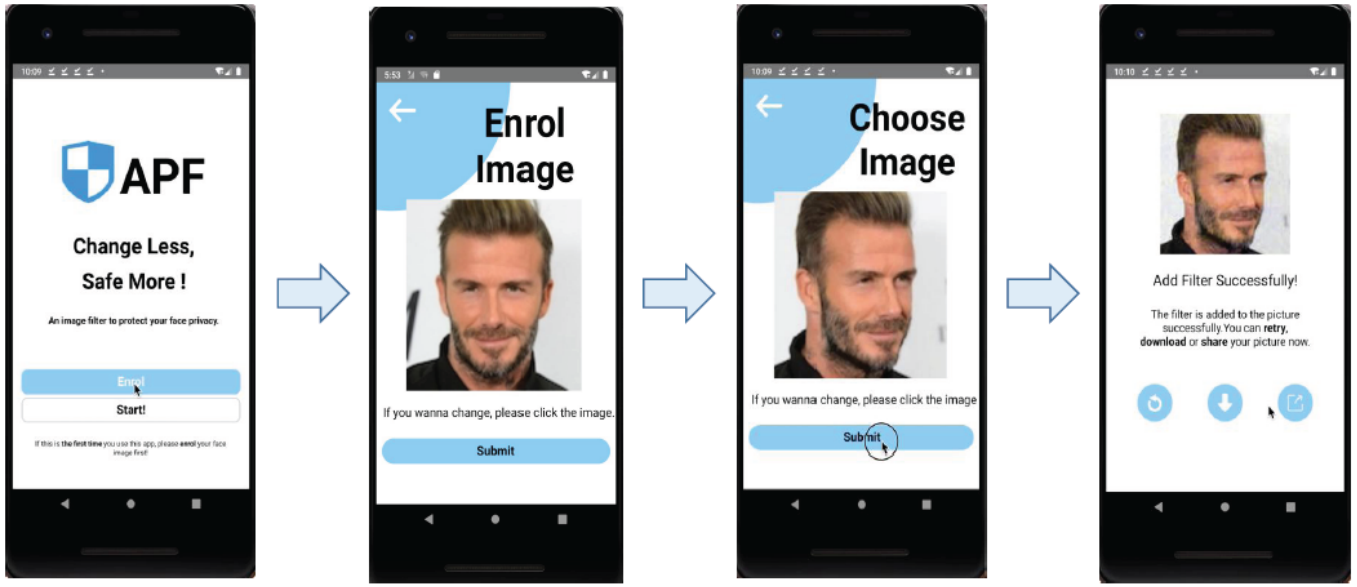
Zhang et al. "Adversarial privacy-preserving filter." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Demonstration

APF - Adversarial Privacy-preserving Filter

Video Program and Demo Session
Time: Oct.22 16:00—18:00

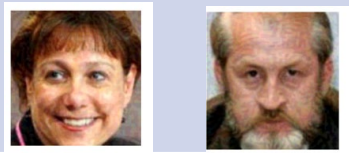
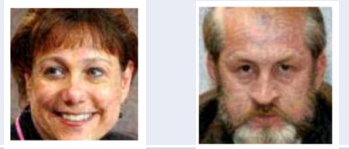
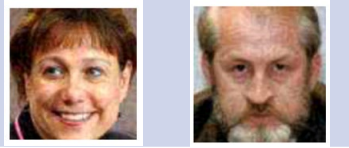
https://github.com/adversarial-for-goodness/APF



Against Compression

Image Compression Operation

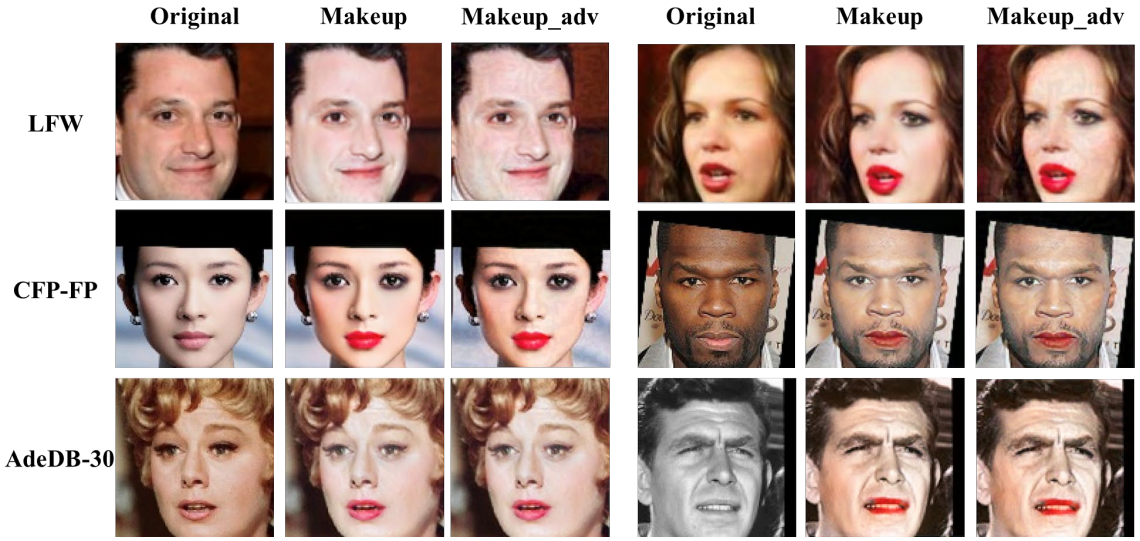
Pinterest MediaFire imgur imgfave SmugMug DEVIANT ART 500px tinypic

	Compression Rate	ASR	Example
Adversarial example	1x	98.4%	
Adversarial example with Compression quality of 0.75	6x	94.35%	
Adversarial example with Compression quality of 0.45	6x	89.95%	
Adversarial example with Compression quality of 0.25	10x	85.05%	

Makeup

Problem

- Pursuit of beauty is increasing.
- Filters applied on social media platforms usually make people look better, while our filters do not have this feature.



Solution

- Makeup can also slightly protect privacy.
- Adversarial perturbation can be compatible with makeup.

Dataset	LFW	CFP_FP	AgeDB_30
Adv	98.5%	92.5%	95.8%
Makeup	5.138%	4.83%	9.50%
Makeup_Adv	98.73%	96.95%	99.44%

Discussion

Discussion

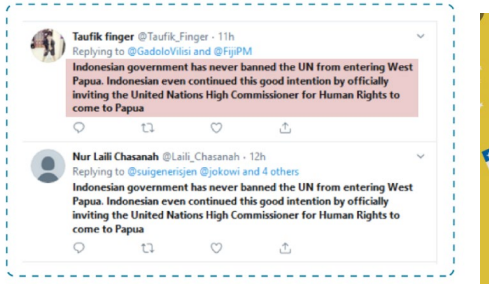
Physical adversarial privacy-preserving filter

Big data discriminatory pricing (BDDP):
The shops deploy face recognition system to model users for discriminating consumers.

Discussion

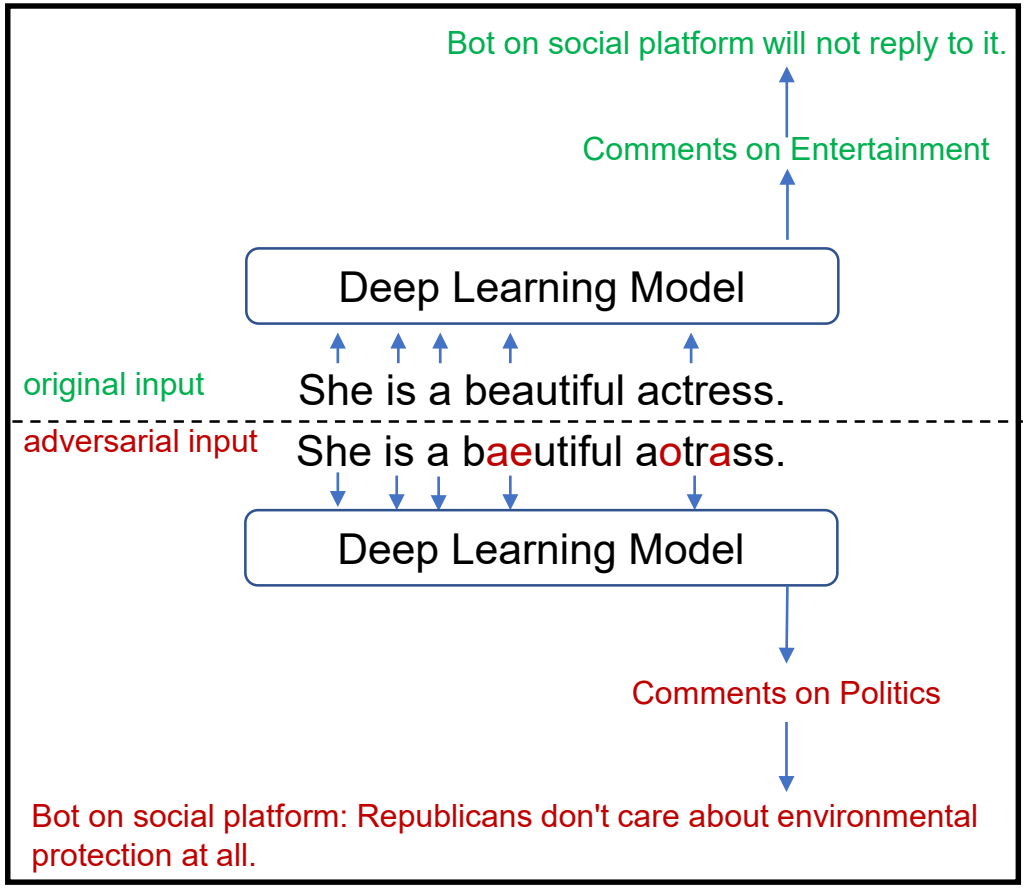
Bots on social platforms:

Bots that incite racial hatred, gender antagonism and political campaigns on social media platforms.



They only post content about Papua and numerous accounts post exactly the same thing.

Source: BBC investigation & Australian Strategic Policy Institute BBC



Sting operation on social platform

Thanks!

Contact me: jiamingzhang@bjtu.edu.cn