



香港城市大學  
City University of Hong Kong



**CISPA**  
HELMHOLTZ CENTER FOR  
INFORMATION SECURITY

# Teacher Model Fingerprinting Attacks Against Transfer Learning

Yufei Chen<sup>1,2</sup>, Chao Shen<sup>1</sup>, Cong Wang<sup>2</sup>, Yang Zhang<sup>3</sup>

<sup>1</sup>Xi'an Jiaotong University

<sup>2</sup>City University of Hong Kong

<sup>3</sup>CISPA Helmholtz Center for Information Security

# Huge Success of Deep Learning



# Reality: A DL Model is Expensive

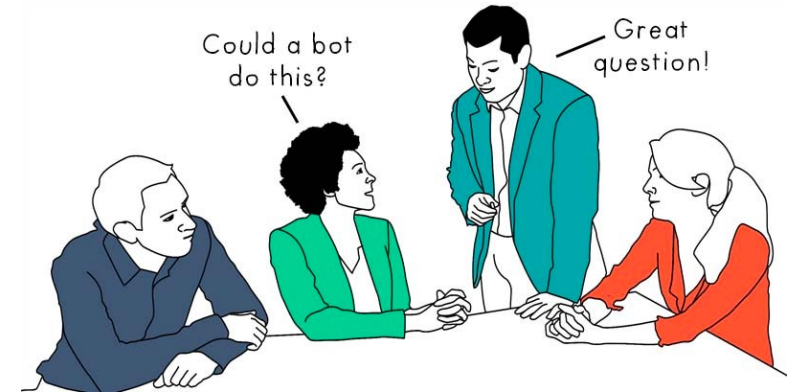


**Data Hungry**  
**(ImageNet ~14M)**



**GPT-3:**  
**# Parameters: 175B**  
**Estimated Cost: \$12M**

**High**  
**Computational Cost**  
**(~355 years on a single**  
**NVIDIA Tesla V100 GPU\*)**



**Experts**



# Reality: A DL Model is Expensive



Data H



Experts

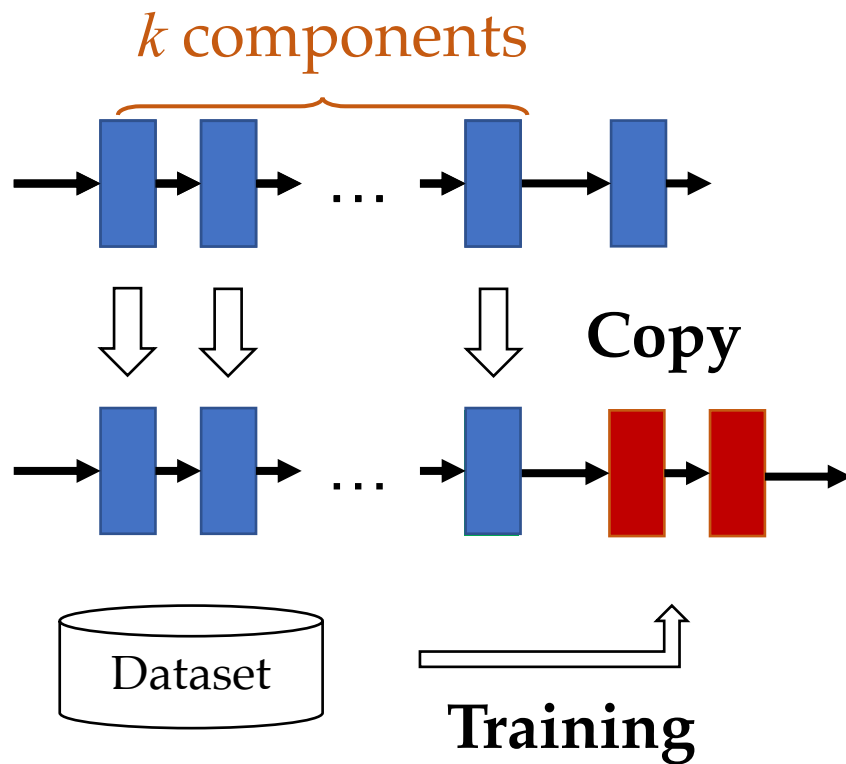
# Transfer Learning -- An Affordable Solution

Google

Teacher



Student



Pretrained components

Newly trained components

Fine-tuned components

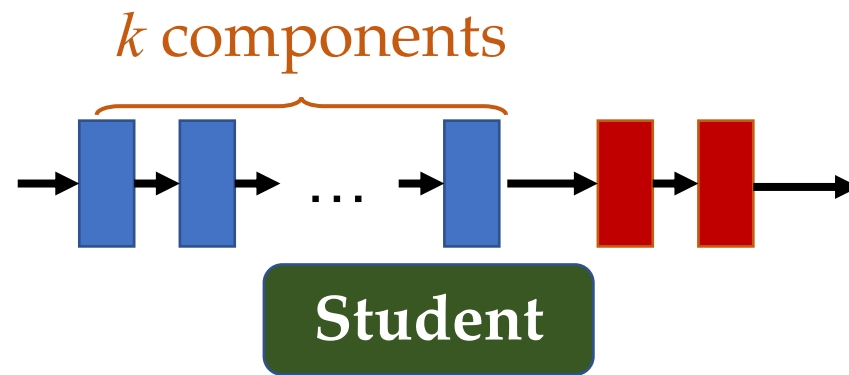
Recommended by

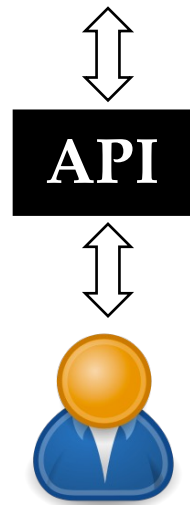
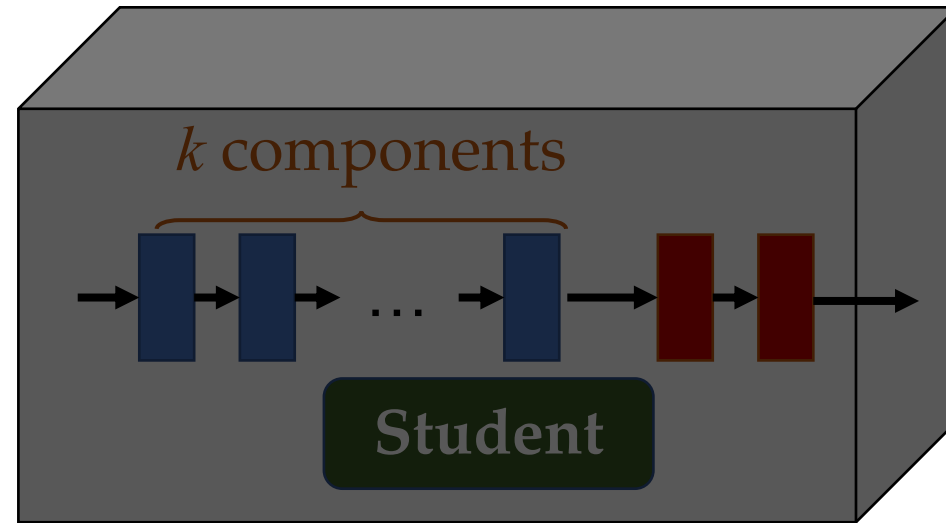
IBM Google Microsoft

Meta TensorFlow

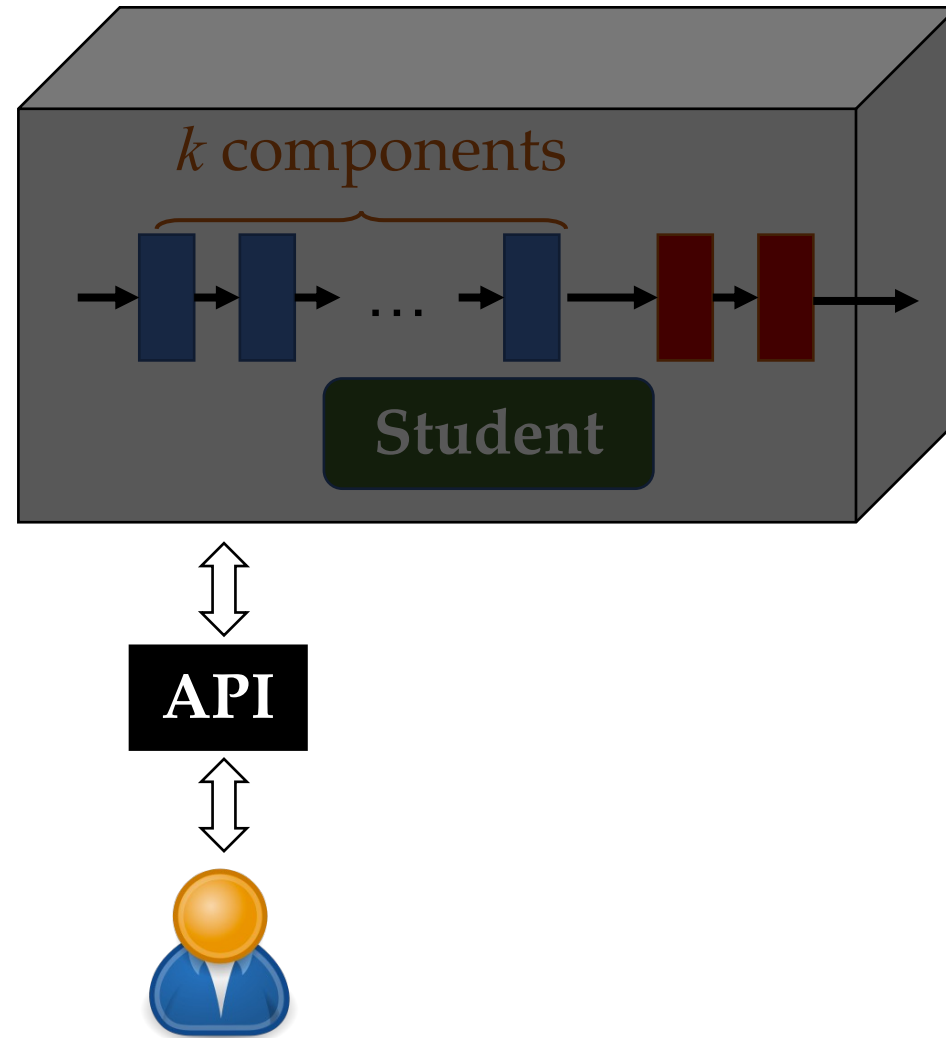
PyTorch Model Zoo

飞桨 PaddleHub ...



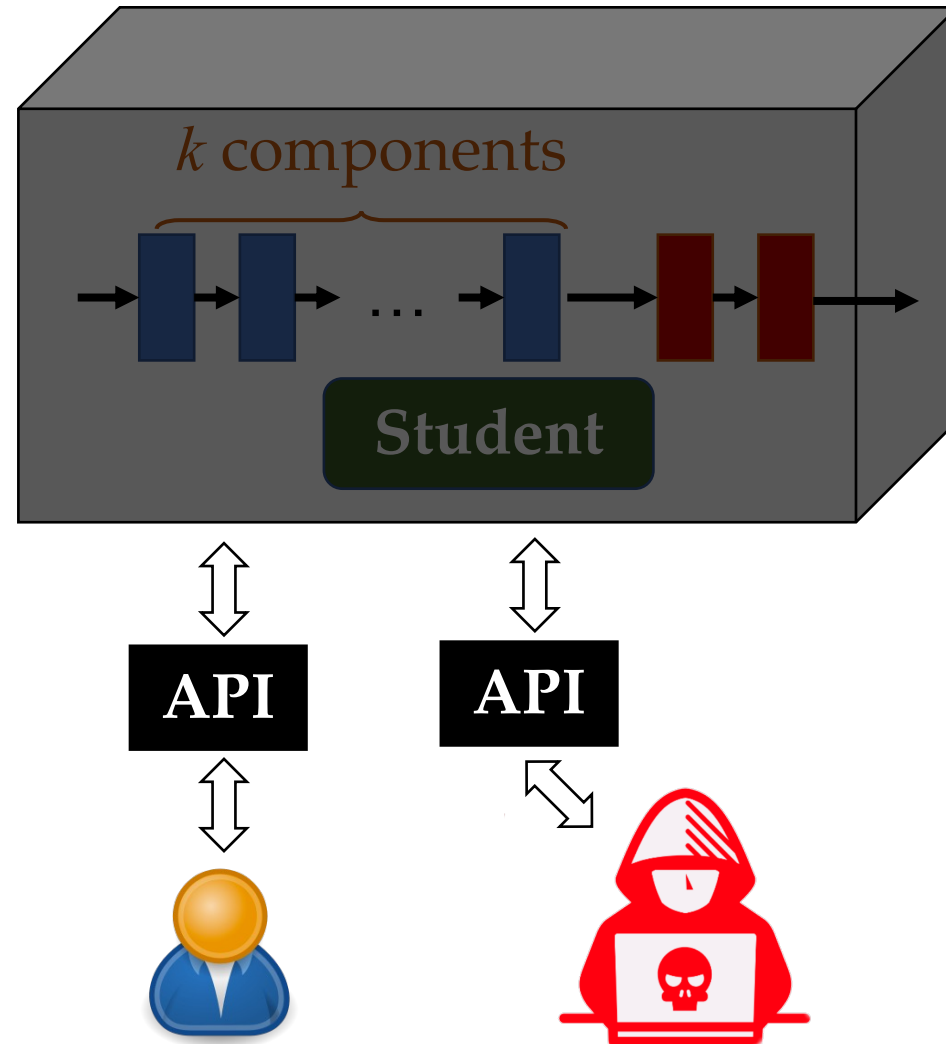


# Transfer Learning -- A **SAFE** Solution?

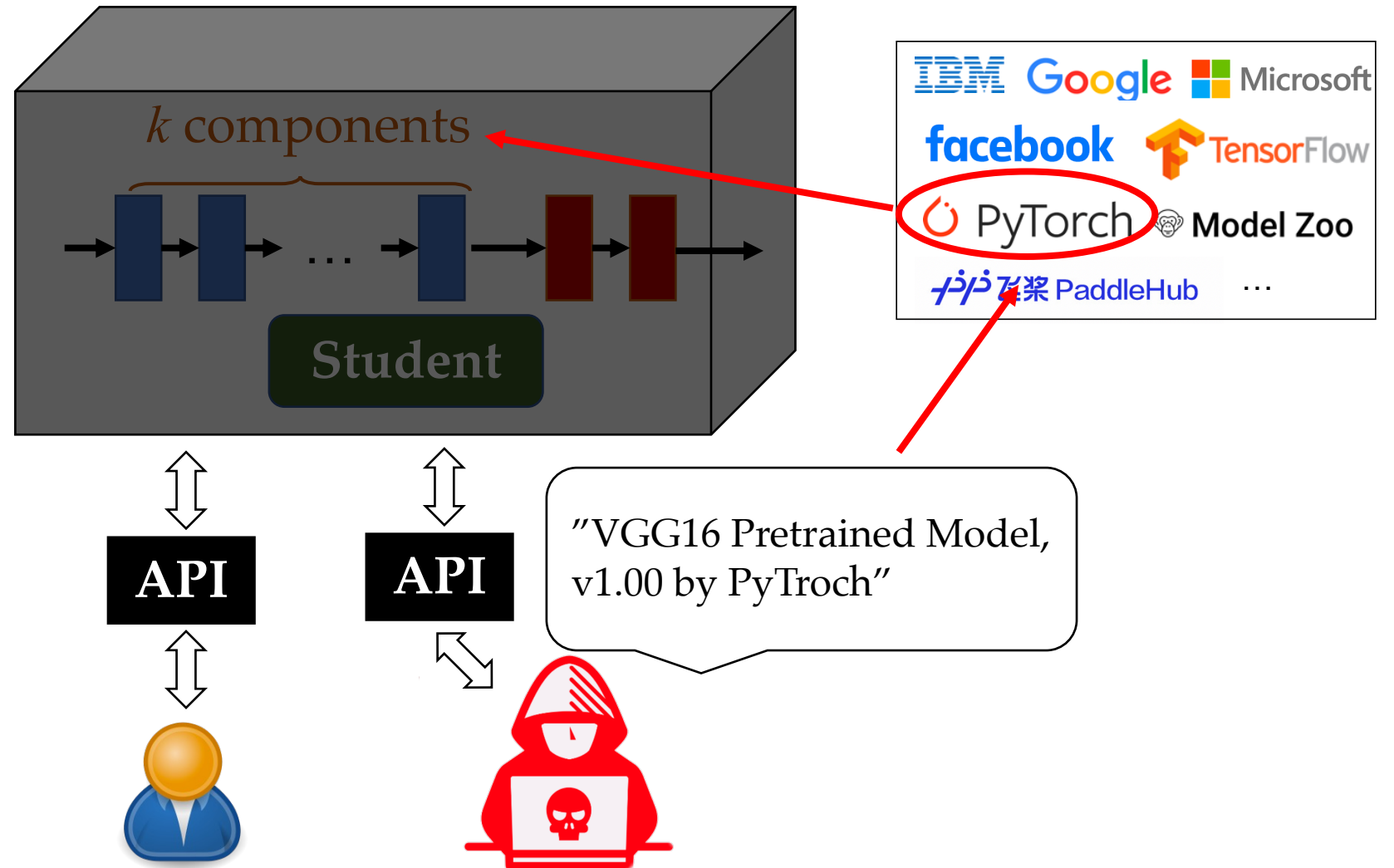




# Transfer Learning -- A **SAFE** Solution?



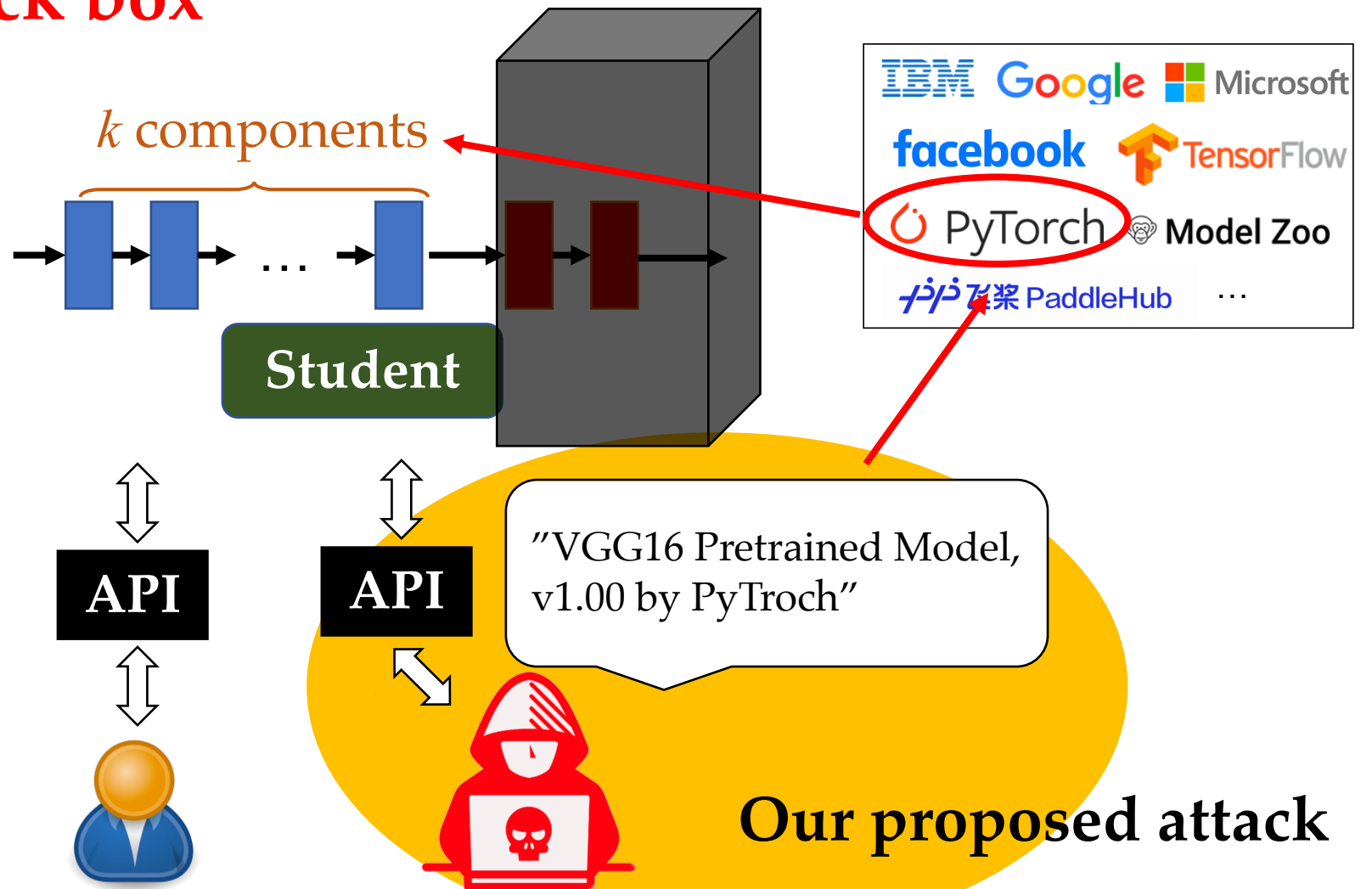
# Transfer Learning -- A **SAFE** Solution?



# Transfer Learning -- A ~~SAFE~~ Solution?

Most part of the black box is exposed! 😱

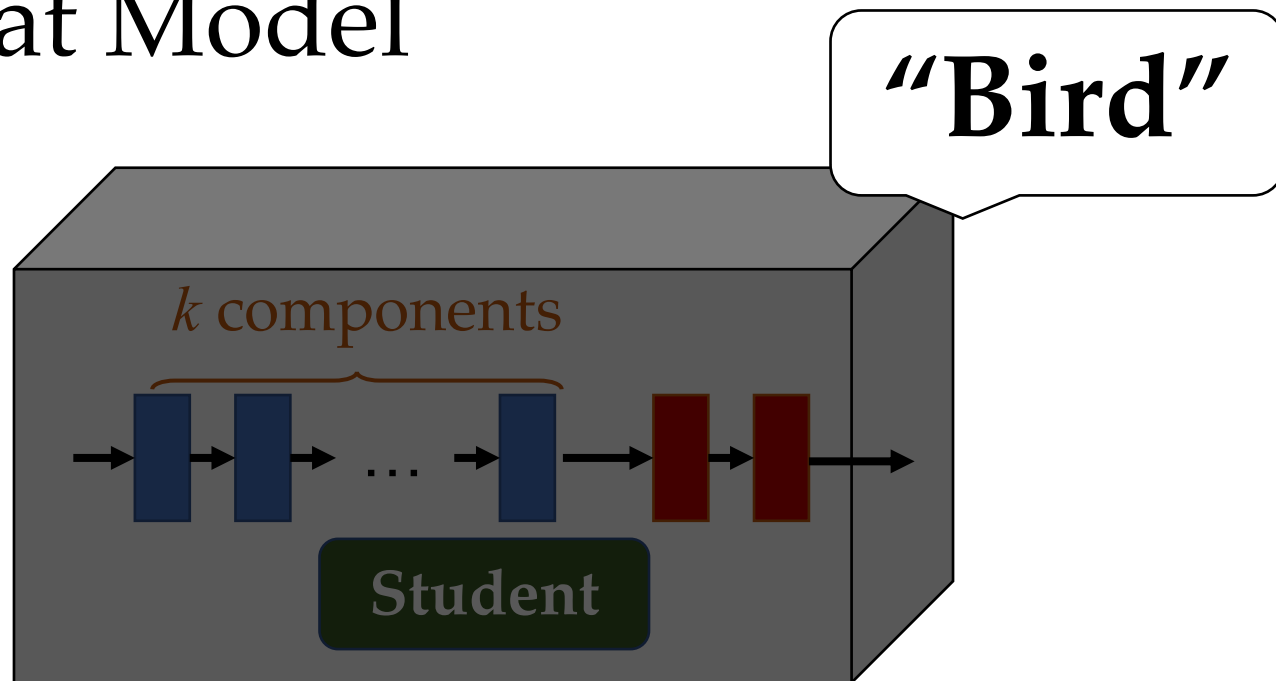
- Vulnerabilities exposure (from the teacher)
- Downstream attacks



# Threat Model

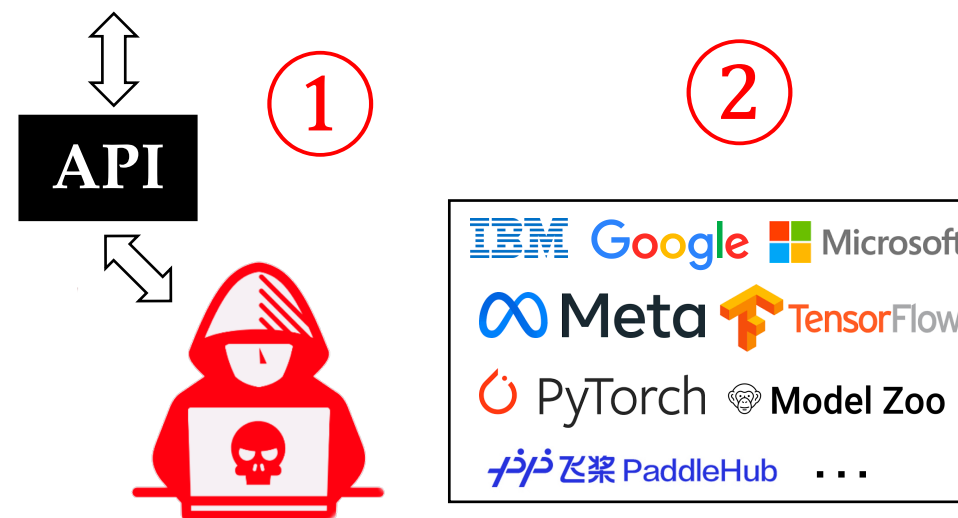
## ① Black-box access:

- ❑ Unknown student architecture / parameters
- ❑ Only top-1 classification label returned



## ② Attacker's knowledge/power:

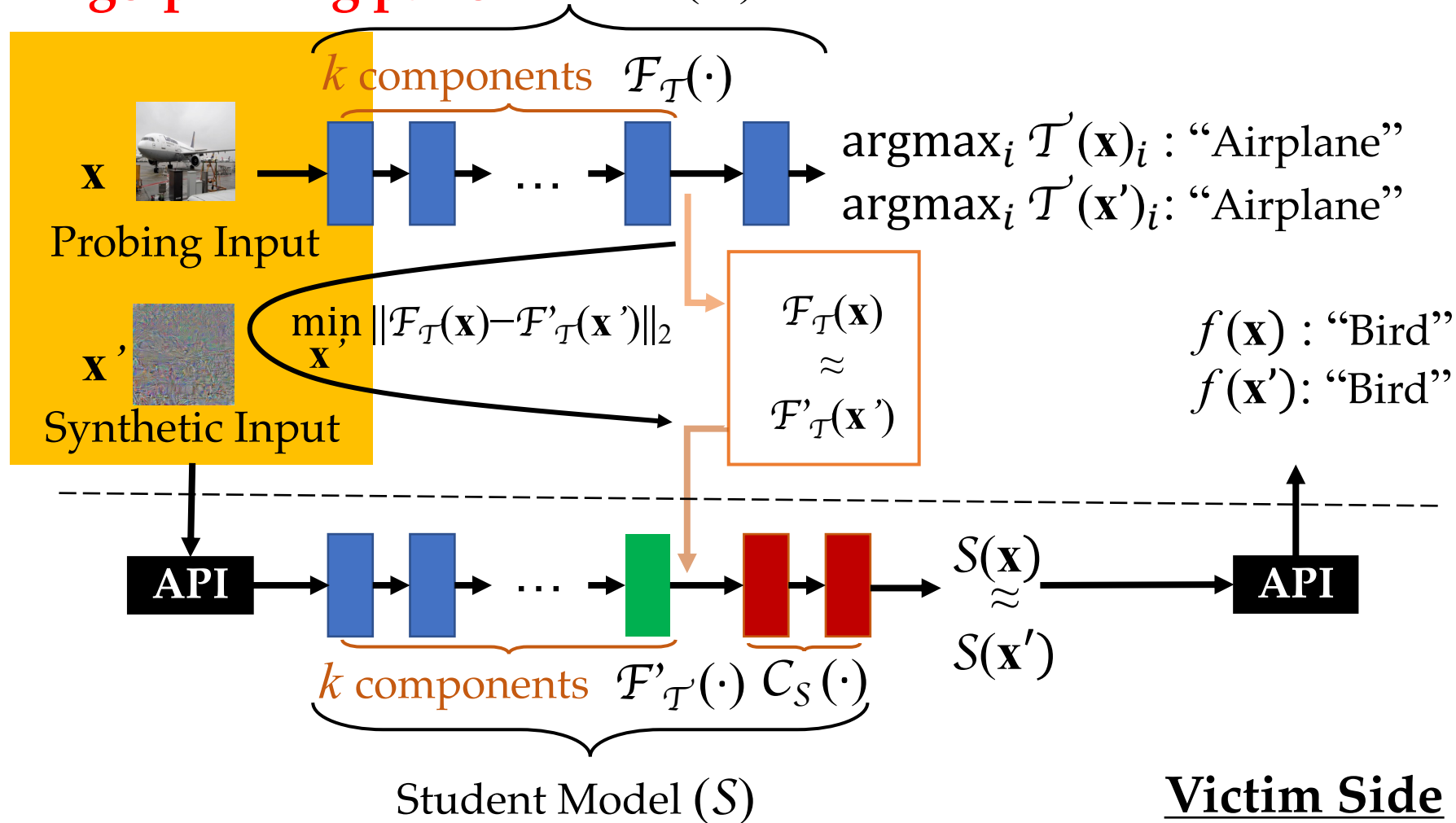
- ❑ Candidate teacher models
- ❑ Public datasets (e.g., ImageNets, CIFAR10)
- ❑ Limited query budget



# Overview: Teacher Fingerprinting Attack

**Fingerprinting pairs** for Model ( $\mathcal{T}$ )

Attacker Side



**Insight:**

- Fingerprinting pairs
- ↓
- Similar latent representation
- ↓
- Same API responses



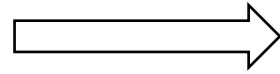
# Attack Stage 1: Synthetic Input Generation

- Solving constrained optimization

Adam optimizer  
Learning rate: 0.001  
#Iterations: 30,000

$$\tanh(\mathbf{w}) = \frac{2\tilde{\mathbf{x}}}{255} - 1$$

$$\begin{aligned} \mathbf{x}' &= \arg \min_{\tilde{\mathbf{x}}} \|\mathcal{F}_T(\tilde{\mathbf{x}}) - \mathcal{F}_T(\mathbf{x})\|_2 \\ \text{s.t. } \tilde{\mathbf{x}} &\in [0, 255] \end{aligned}$$



$$\mathbf{w}' = \arg \min_{\mathbf{w}} \left\| \mathcal{F}_T \left( 255 * \frac{1}{2} (\tanh(\mathbf{w}) + 1) \right) - \mathcal{F}_T(\mathbf{x}_i) \right\|_2$$

**Original problem  
(Constrained)**

**Converted problem  
(Unconstrained)**

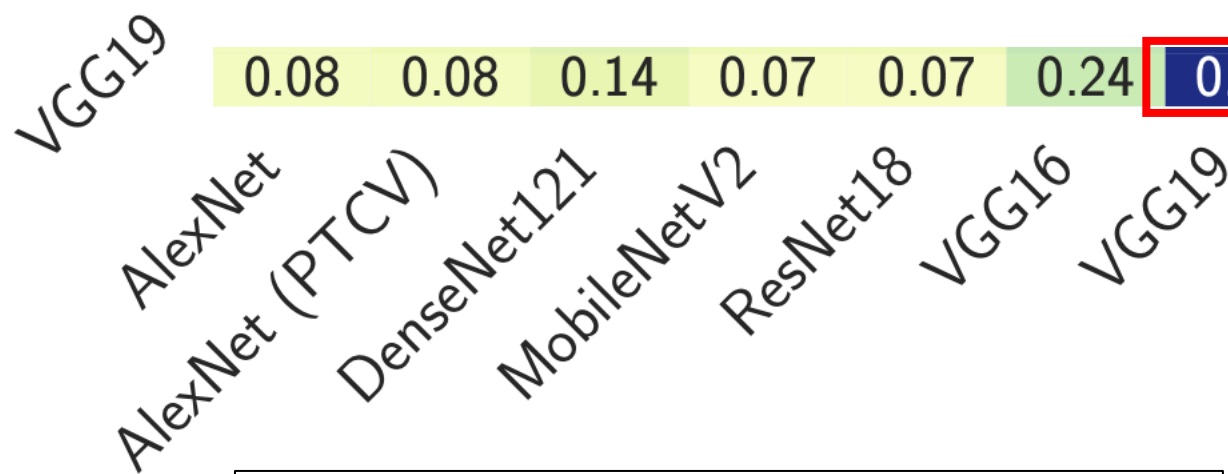
# Attack Stage 2: Teacher Model Inference

- Inference Metric

- Matching proportion:

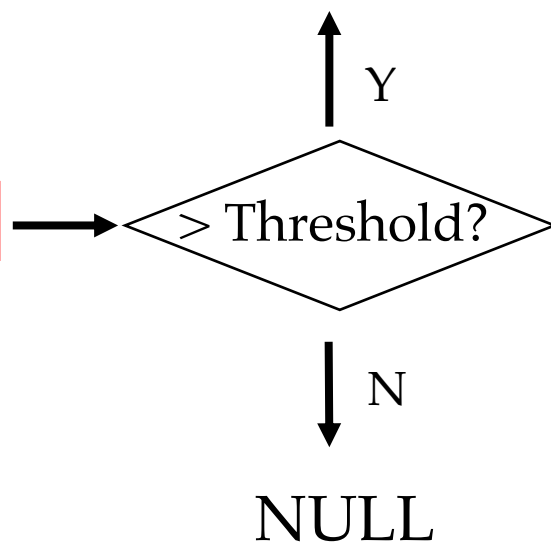
$$\frac{\text{\#Matched Responses}}{\text{\#Fingerprinting Pairs}}$$

Actual teacher model



Candidate teacher model set

Inference: VGG19



# Effectiveness of Our Proposed Attack

- Basic setup

# fingerprinting pairs:  
100 for each candidate

# student models:

6 datasets \* 7 teacher  
models \* 3 student FCN  
architectures

$D_s$ : Dogs-vs-Cats (2-class)

AlexNet	0.89	0.69	0.68	0.55	0.55	0.65	0.61
AlexNet (PTCV)	0.81	0.53	0.64	0.58	0.43	0.47	0.44
DenseNet121	0.81	0.55	0.49	0.49	0.45	0.45	0.46
GoogLeNet	0.54	0.90	0.63	0.55	0.47	0.48	0.49
MobileNetV2	0.64	0.95	0.73	0.73	0.54	0.61	0.59
ResNet18	0.58	0.90	0.60	0.59	0.57	0.57	0.56
VGG16	0.48	0.48	0.89	0.53	0.62	0.57	0.55
VGG19	0.68	0.66	0.90	0.67	0.71	0.63	0.75
AlexNet	0.46	0.47	0.93	0.61	0.67	0.59	0.55
AlexNet (PTCV)	0.58	0.57	0.56	0.57	0.55	0.55	0.55
DenseNet121	0.49	0.63	0.65	0.63	0.61	0.59	0.61
GoogLeNet	0.48	0.45	0.51	0.45	0.43	0.42	0.44
MobileNetV2	0.27	0.30	0.39	0.98	0.44	0.67	0.59
ResNet18	0.28	0.28	0.35	0.97	0.61	0.71	0.70
VGG16	0.42	0.41	0.60	0.93	0.68	0.68	0.68
VGG19	0.69	0.48	0.47	0.72	0.98	0.68	0.63
AlexNet	0.39	0.35	0.44	0.54	0.99	0.54	0.47
AlexNet (PTCV)	0.48	0.40	0.54	0.67	1.00	0.59	0.60
DenseNet121	0.70	0.66	0.70	0.72	0.71	1.00	0.66
GoogLeNet	0.77	0.68	0.73	0.65	0.75	0.99	0.76
MobileNetV2	0.80	0.79	0.80	0.80	0.80	0.99	0.86
ResNet18	0.65	0.68	0.61	0.59	0.51	0.65	0.97
VGG16	0.84	0.82	0.79	0.82	0.80	0.80	0.96
VGG19	0.79	0.79	0.74	0.80	0.79	0.78	0.97

$D_s$ : MNIST (10-class)

AlexNet	0.88	0.70	0.70	0.70	0.70	0.70	0.70
AlexNet (PTCV)	0.65	0.39	0.33	0.28	0.30	0.32	0.28
DenseNet121	0.74	0.44	0.44	0.44	0.44	0.44	0.44
GoogLeNet	0.40	0.86	0.50	0.32	0.17	0.17	0.16
MobileNetV2	0.21	0.77	0.20	0.21	0.21	0.21	0.20
ResNet18	0.41	0.82	0.45	0.42	0.41	0.41	0.41
VGG16	0.26	0.31	0.86	0.32	0.31	0.29	0.27
VGG19	0.27	0.22	0.84	0.30	0.27	0.34	0.34
AlexNet	0.22	0.23	0.82	0.27	0.27	0.27	0.23
AlexNet (PTCV)	0.31	0.55	0.37	0.47	0.23	0.60	0.58
DenseNet121	0.35	0.37	0.45	0.44	0.42	0.40	0.38
GoogLeNet	0.44	0.43	0.51	0.44	0.51	0.36	0.35
MobileNetV2	0.22	0.23	0.27	0.91	0.24	0.24	0.22
ResNet18	0.15	0.10	0.17	0.92	0.12	0.21	0.22
VGG16	0.17	0.15	0.19	0.91	0.18	0.20	0.18
VGG19	0.17	0.08	0.33	0.34	1.00	0.14	0.19
AlexNet	0.17	0.09	0.25	0.32	0.98	0.23	0.24
AlexNet (PTCV)	0.17	0.09	0.14	0.17	0.95	0.08	0.12
DenseNet121	0.77	0.77	0.77	0.77	0.77	1.00	0.79
GoogLeNet	0.71	0.69	0.69	0.71	0.71	0.99	0.65
MobileNetV2	0.85	0.83	0.83	0.83	0.84	1.00	0.84
ResNet18	0.70	0.69	0.66	0.69	0.71	0.71	0.96
VGG16	0.76	0.74	0.67	0.73	0.76	0.73	1.00
VGG19	0.86	0.85	0.85	0.85	0.87	0.87	0.99

$D_s$ : STL10 (10-class)

AlexNet	0.72	0.16	0.25	0.10	0.01	0.02	0.05
AlexNet (PTCV)	0.67	0.18	0.16	0.11	0.05	0.06	0.05
DenseNet121	0.70	0.18	0.22	0.08	0.04	0.05	0.03
GoogLeNet	0.15	0.80	0.34	0.21	0.03	0.03	0.05
MobileNetV2	0.18	0.86	0.29	0.15	0.01	0.02	0.04
ResNet18	0.24	0.84	0.33	0.12	0.05	0.04	0.06
VGG16	0.08	0.09	0.88	0.11	0.09	0.07	0.10
VGG19	0.21	0.28	0.88	0.10	0.12	0.11	0.10
AlexNet	0.10	0.11	0.90	0.12	0.14	0.07	0.09
AlexNet (PTCV)	0.08	0.27	0.21	0.08	0.27	0.22	0.18
DenseNet121	0.17	0.12	0.19	0.17	0.14	0.15	0.17
GoogLeNet	0.04	0.08	0.13	0.13	0.11	0.10	0.13
MobileNetV2	0.15	0.29	0.28	0.91	0.24	0.21	0.13
ResNet18	0.09	0.20	0.16	0.92	0.23	0.10	0.09
VGG16	0.15	0.15	0.13	0.90	0.14	0.13	0.13
VGG19	0.11	0.11	0.24	0.09	0.97	0.06	0.08
AlexNet	0.13	0.18	0.36	0.21	1.00	0.18	0.18
AlexNet (PTCV)	0.11	0.12	0.27	0.15	0.98	0.11	0.10
DenseNet121	0.02	0.07	0.10	0.06	0.08	1.00	0.22
GoogLeNet	0.06	0.07	0.10	0.05	0.07	0.99	0.20
MobileNetV2	0.01	0.04	0.04	0.05	0.04	1.00	0.18
ResNet18	0.04	0.04	0.08	0.03	0.03	0.26	0.94
VGG16	0.09	0.11	0.17	0.11	0.11	0.26	0.91
VGG19	0.08	0.08	0.14	0.07	0.07	0.24	0.91

$D_s$ : CIFAR10 (10-class)

AlexNet	0.68	0.25	0.23	0.21	0.20	0.20	0.21
AlexNet (PTCV)	0.67	0.21	0.23	0.20	0.19	0.19	0.20
DenseNet121	0.66	0.19	0.13	0.17	0.15	0.15	0.16
GoogLeNet	0.26	0.80	0.36	0.19	0.11	0.12	0.13
MobileNetV2	0.22	0.82	0.30	0.21	0.19	0.20	0.20
ResNet18	0.25	0.84	0.30	0.21	0.18	0.19	0.18
VGG16	0.03	0.04	0.83	0.12	0.31	0.12	0.06
VGG19	0.04	0.05	0.86	0.16	0.35	0.16	0.16
AlexNet	0.02	0.03	0.84	0.11	0.34	0.10	0.11
AlexNet (PTCV)	0.11	0.16	0.31	0.26	0.27	0.04	0.05
DenseNet121	0.28	0.27	0.42	0.32	0.47	0.48	0.47
GoogLeNet	0.09	0.09	0.18	0.20	0.09	0.08	0.08
MobileNetV2	0.39	0.44	0.32	0.44	0.39	0.35	0.24
ResNet18	0.16	0.20	0.20	0.92	0.38	0.14	0.11
VGG16	0.17	0.25	0.36	0.94	0.42	0.30	0.19
VGG19	0.32	0.31	0.49	0.25	1.00	0.24	0.21
AlexNet	0.19	0.06	0.49	0.20	1.00	0.44	0.37
AlexNet (PTCV)	0.23	0.07	0.42	0.18	0.98	0.46	0.34
DenseNet121	0.07	0.11	0.15	0.07	0.12	0.98	0.38
GoogLeNet	0.25	0.13	0.11	0.07	0.13	1.00	0.27
MobileNetV2	0.20	0.14	0.09	0.07	0.12	1.00	0.23
ResNet18	0.10	0.07	0.13	0.07	0.08	0.16	0.91
VGG16	0.08	0.12	0.13	0.07	0.08	0.18	0.92
VGG19	0.07	0.08	0.14	0.08	0.08	0.16	0.89

$D_s$ : CIFAR100 (100-class)

AlexNet	0.58	0.03	0.05	0.03	0.01	0.01	0.01
AlexNet (PTCV)	0.64	0.03	0.12	0.06	0.02	0.02	0.01
DenseNet121	0.63	0.03	0.05	0.00	0.00	0.00	0.00
GoogLeNet	0.03	0.78	0.12	0.01	0.01	0.02	0.04
MobileNetV2	0.02	0.76	0.09	0.04	0.00	0.01	0.02
ResNet18	0.03	0.82	0.12	0.04	0.01	0.00	0.02
VGG16	0.04	0.04	0.82	0.04	0.03	0.02	0.02
VGG19	0.01	0.00	0.83	0.03	0.03	0.00	0.00
AlexNet	0.01	0.02	0.85	0.01	0.02	0.00	0.01
AlexNet (PTCV)	0.02	0.02	0.01	0.01	0.03	0.00	0.00
DenseNet121	0.08	0.03	0.05	0.12	0.02	0.04	0.04
GoogLeNet	0.03	0.12	0.07	0.04	0.14	0.00	0.01
MobileNetV2	0.00	0.01	0.01	0.92	0.04	0.01	0.01
ResNet18	0.01	0.01	0.01	0.92	0.00	0.00	0.00
VGG16	0.01	0.02	0.04	0.92	0.01	0.01	0.02
VGG19	0.01	0.01	0.08	0.00	0.97	0.00	0.00
AlexNet	0.01	0.02	0.05	0.01	0.99	0.00	0.00
AlexNet (PTCV)	0.00	0.00	0.08	0.00	0.99	0.02	0.01
DenseNet121	0.10	0.10	0.10	0.11	0.09	0.92	0.13
GoogLeNet	0.02	0.02	0.01	0.02	0.02	0.99	0.04
MobileNetV2	0.09	0.08	0.07	0.07	0.12	0.98	0.08
ResNet18	0.40	0.42	0.43	0.43	0.44	0.35	0.92
VGG16	0.13	0.19	0.18	0.13	0.13	0.16	0.92
VGG19	0.04	0.08	0.12	0.05	0.11	0.12	0.93

$D_s$ : CelebA (40 binary attributes)

AlexNet	0.77	0.31	0.52	0.31	0.30	0.26	0.23
AlexNet (PTCV)	0.76	0.33	0.36	0.27	0.32	0.28	0.29
DenseNet121	0.77	0.34	0.48	0.30	0.25	0.28	0.29
GoogLeNet	0.57	0.84	0.67	0.46	0.14	0.22	0.25
MobileNetV2	0.54	0.83	0.60	0.56	0.34	0.44	0.44
ResNet18	0.36	0.87	0.46	0.40	0.40	0.40	0.41
VGG16	0.25	0.27	0.87	0.36	0.38	0.39	0.34
VGG19	0.37	0.34	0.84	0.29	0.38	0.40	0.36
AlexNet	0.40	0.30	0.88	0.36	0.35	0.34	0.34
AlexNet (PTCV)	0.57	0.53	0.60	0.57	0.58	0.57	0.58
DenseNet121	0.46	0.40	0.50	0.24	0.17	0.39	0.42
GoogLeNet	0.40	0.43	0.42	0.42	0.43	0.43	0.41
MobileNetV2	0.22	0.36	0.38	0.91	0.24	0.25	0.23
ResNet18	0.31	0.35	0.47	0.92	0.30	0.33	0.34
VGG16	0.31	0.37	0.44	0.92	0.40	0.41	0.40
VGG19	0.38	0.43	0.48	0.39	0.99	0.45	0.39
AlexNet	0.37	0.38	0.49	0.35	0.99	0.33	0.36
AlexNet (PTCV)	0.32	0.35	0.45	0.33	1.00	0.37	0.38
DenseNet121	0.42	0.38	0.42	0.39	0.20	1.00	0.40
GoogLeNet	0.44	0.50	0.50	0.53	0.40	1.00	0.45
MobileNetV2	0.54	0.52	0.53	0.54	0.47	1.00	0.48
ResNet18	0.69	0.68	0.72	0.72	0.72	0.71	0.96
VGG16	0.80	0.78	0.78	0.79	0.79	0.77	0.98
VGG19	0.41	0.52	0.54	0.51	0.53	0.93	0.91

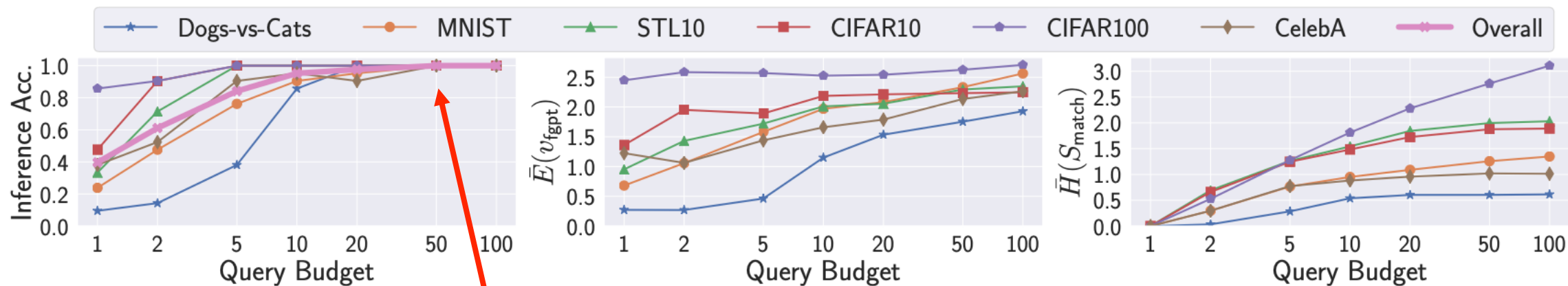
# Effectiveness of Our Proposed Attack

- Basic Results

<b>Correctly inferred</b>	<b>Inferred as "NULL"</b>	
<b>w/ kown teacher model</b>	<b>w/ unknown teacher model</b>	<b>w/o transfer learning</b>
100% (126 / 126)	72.2% (13 / 18)	86.1% (31 / 36)

# Effectiveness of Our Proposed Attack

- Impact of Query Budget #Fingerprinting pairs for each candidate



100% inference accuracy





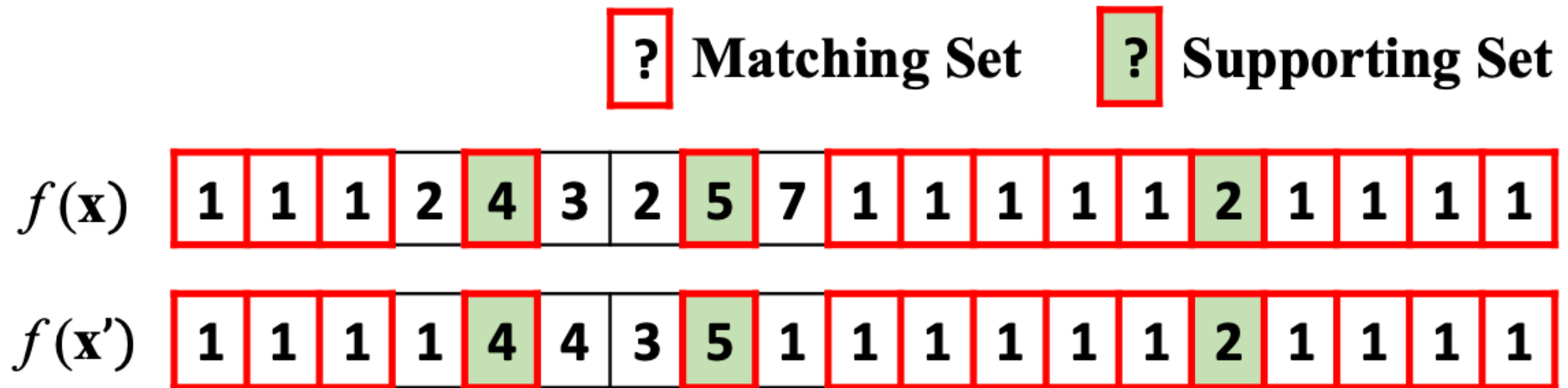
100% matching proportion

**(False matching)**

# Towards More Robust Attack

- Supporting Set

Remove the most frequently matched elements





# Towards More Robust Attack

- Supporting Set

Remove the most frequently matched elements

$$|\text{Supporting Set}| \geq \left\lceil \log_2 \frac{1}{\alpha} \right\rceil + \left\lceil \frac{\left\lceil \log_2 \frac{1}{\alpha} \right\rceil}{c-1} \right\rceil$$

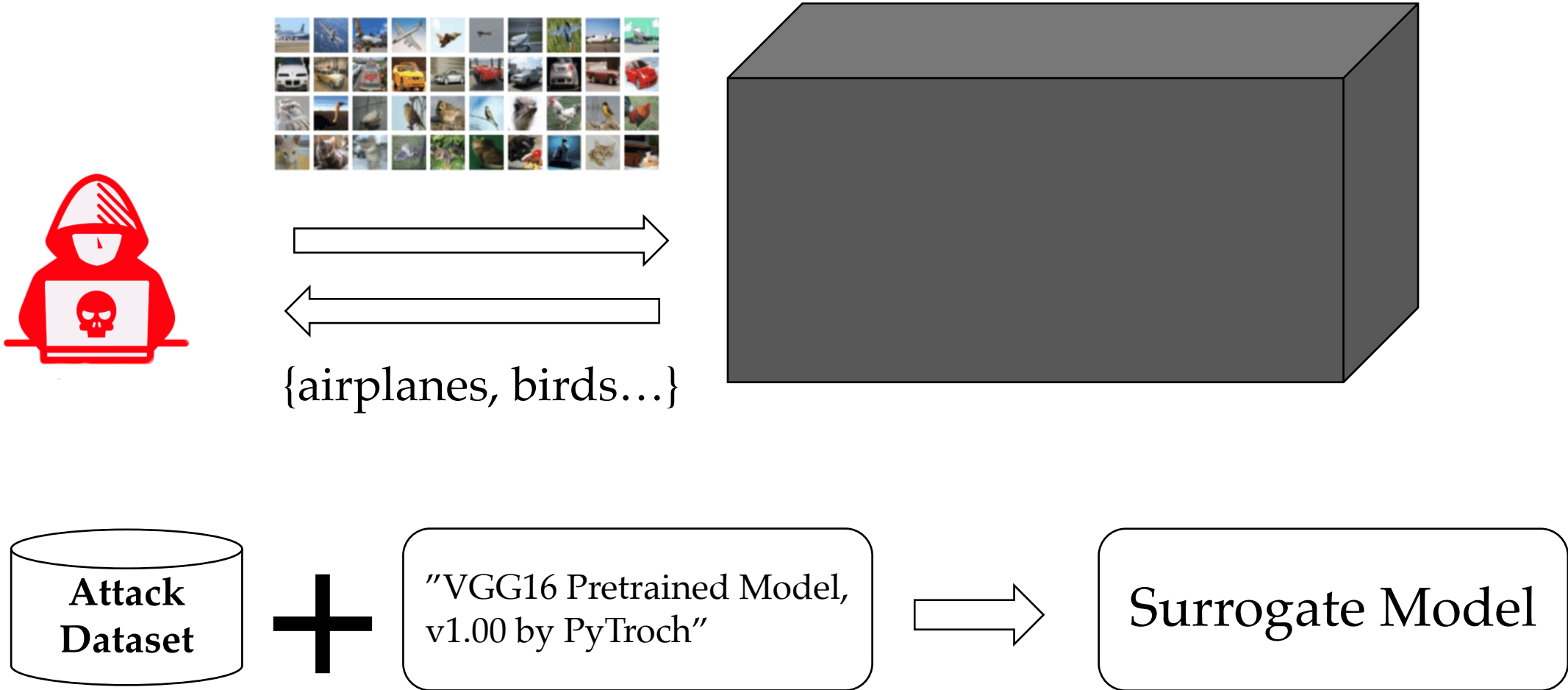
# Towards More Robust Attack

Most inference results are indeed invalid when # query is small

Query Budget	probing: VOCSegmentation			probing: MNIST			probing: CelebA			probing: Random Noise		
	inference acc.		#robust #original	inference acc.		#robust #original	inference acc.		#robust #original	inference acc.		#robust #original
	original	robust		original	robust		original	robust		original	robust	
1	39.68% (50/126)	– (0/0)	0 (0/126)	42.06% (53/126)	– (0/0)	0 (0/126)	45.24% (57/126)	– (0/0)	0 (0/126)	19.84% (25/126)	– (0/0)	– (0/126)
2	61.11% (77/126)	– (0/0)	0 (0/126)	57.94% (73/126)	– (0/0)	0 (0/126)	57.94% (73/126)	– (0/0)	0 (0/126)	29.37% (37/126)	– (0/0)	– (0/126)
5	84.13% (106/126)	– (0/0)	0 (0/126)	69.84% (88/126)	– (0/0)	0 (0/126)	80.95% (102/126)	– (0/0)	0 (0/126)	42.06% (53/126)	– (0/0)	– (0/126)
10	95.24% (120/126)	100.00% (32/32)	25.40% (32/126)	80.95% (102/126)	100.00% (19/19)	15.08% (19/126)	89.68% (113/126)	100.00% (3/3)	2.38% (3/126)	50.79% (64/126)	– (0/0)	– (0/126)
20	97.62% (123/126)	100.00% (97/97)	76.98% (97/126)	84.92% (107/126)	100.00% (52/52)	41.27% (52/126)	96.83% (122/126)	100.00% (87/87)	69.05% (87/126)	57.14% (72/126)	100.00% (16/16)	12.70% (16/126)
50	100.00% (126/126)	100.00% (125/125)	99.21% (125/126)	90.48% (114/126)	100.00% (96/96)	76.19% (96/126)	99.21% (125/126)	100.00% (117/117)	92.86% (117/126)	62.70% (79/126)	100.00% (36/36)	28.57% (36/126)
100	100.00% (126/126)	100.00% (126/126)	100.00% (126/126)	96.03% (121/126)	100.00% (114/114)	90.48% (114/126)	100.00% (126/126)	100.00% (122/122)	96.83% (122/126)	65.08% (82/126)	100.00% (41/41)	32.54% (41/126)

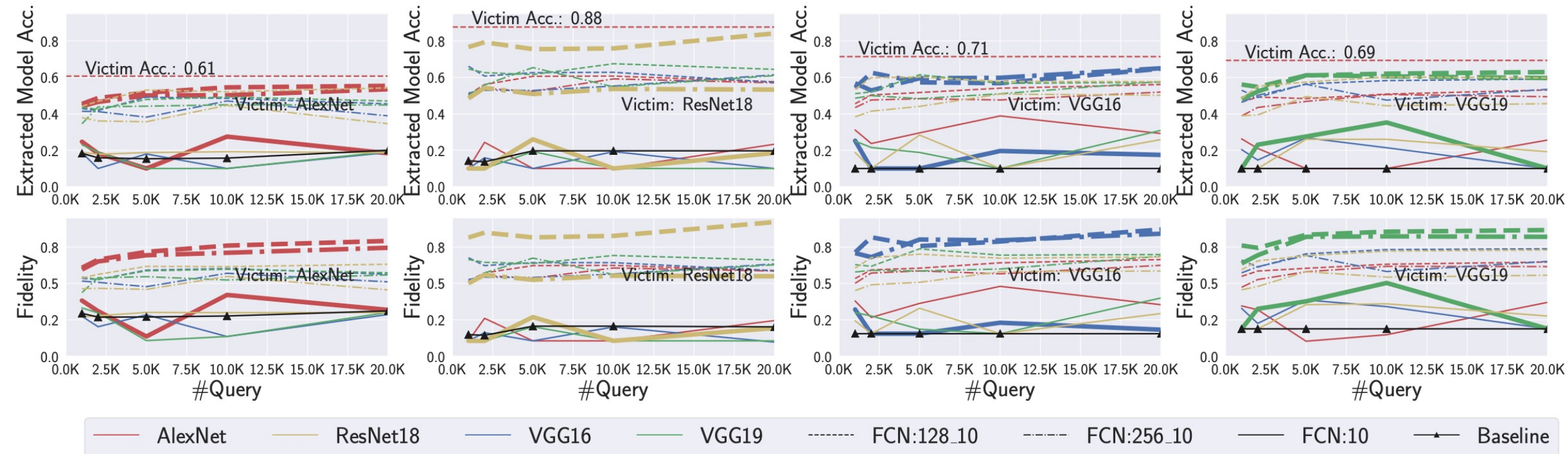


# Enhanced Model Stealing Attack



# Enhanced Model Stealing Attack

- Best performance if starting from a matched teacher model



# Feasible Countermeasures

- Input distortion
  - ❑ Perturb the patterns in synthetic inputs
- Injecting neuron distances [Wang et al. 2018]
  - ❑ Deviate the student model's feature map from the teacher model's

# Conclusion

- ❑ We propose a simple and efficient attack to infer the teacher model used by transfer learning
- ❑ Our attack can efficiently identify the teacher model
- ❑ Our attack can help perform further advanced attacks

Thanks!

Q&A

Yufei Chen

[yufeichen8-c@my.cityu.edu.hk](mailto:yufeichen8-c@my.cityu.edu.hk)