32ND USENIX SECURITY SYMPOSIUM

復旦大學 FUDAN UNIVERSITY

Rethinking White-Box Watermarks on Deep Learning Models under Neural Structural Obfuscation

Yifan Yan, Xudong Pan, Mi Zhang, Min Yang

System and Software Security Lab

School of Computer Science

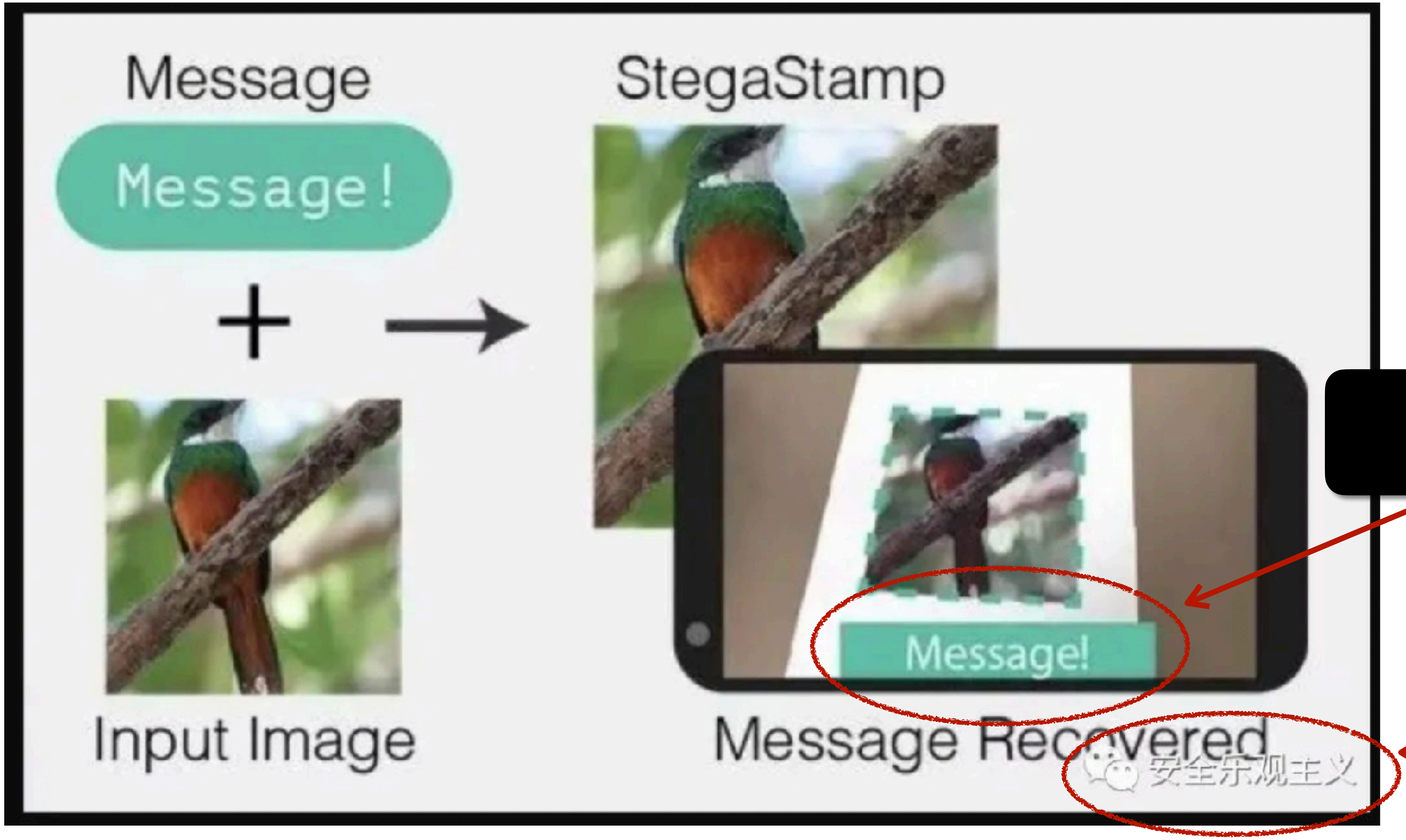
Fudan University

Talk@32nd USENIX Security Symposium

More Research on AI Security

What is Digital Watermarking?

Ownership Verification of Digital Images

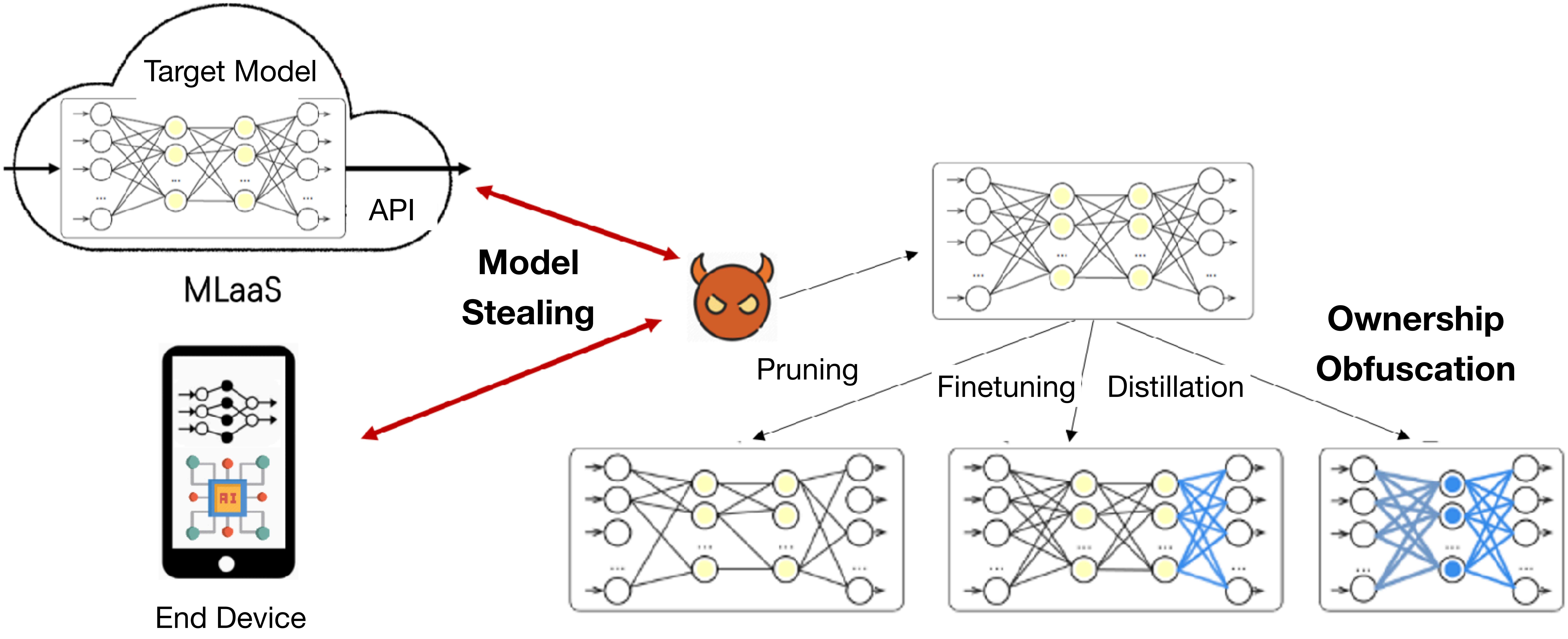


Invisible Watermark

Visible Watermark

DNN Model is facing stealing

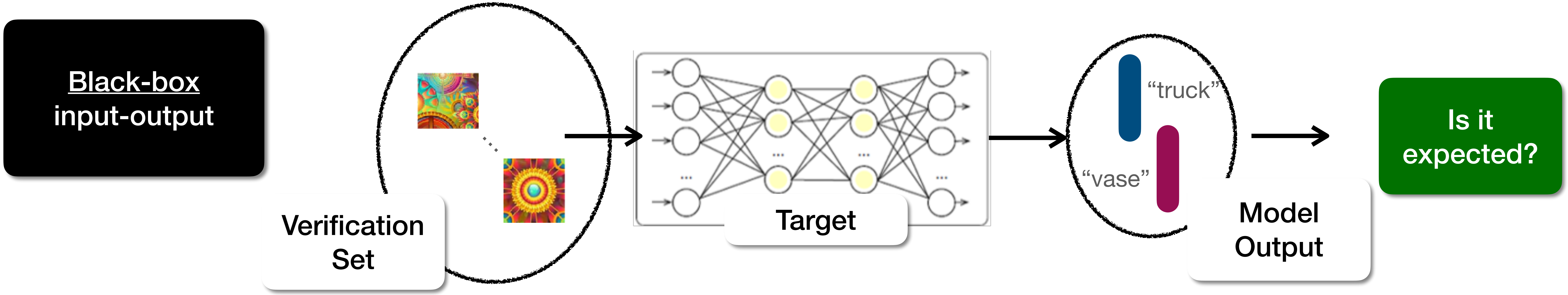
Attackers can steal confidential DNN models from cloud and end devices



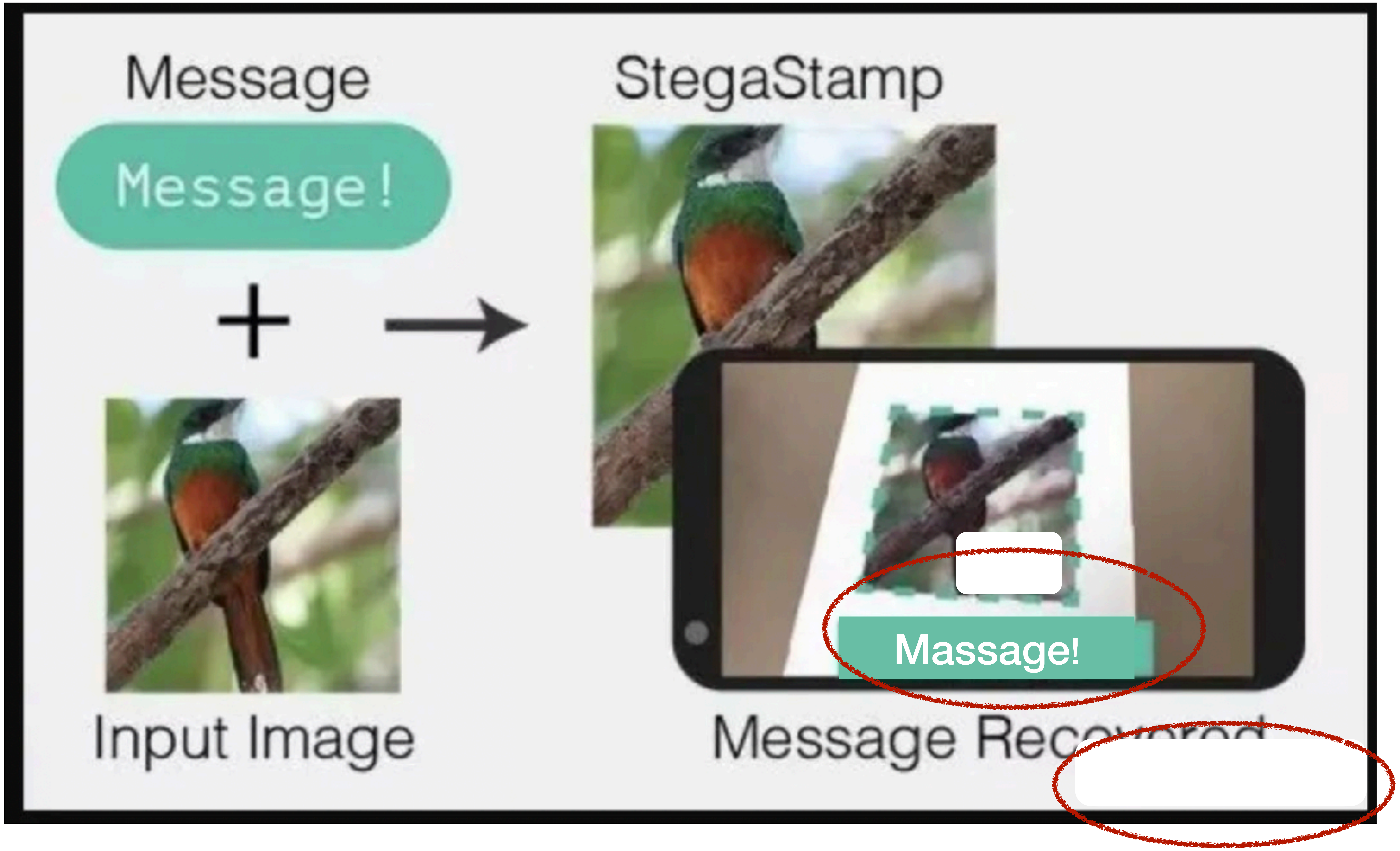
DNN Watermarking

Based on the position where the watermark is embedded

White-Box
Internals
Original Message
A Model of Fudan
encode
Target
decode
A Model of Fudan
Suspect
decode
A *odel o* Fu*an



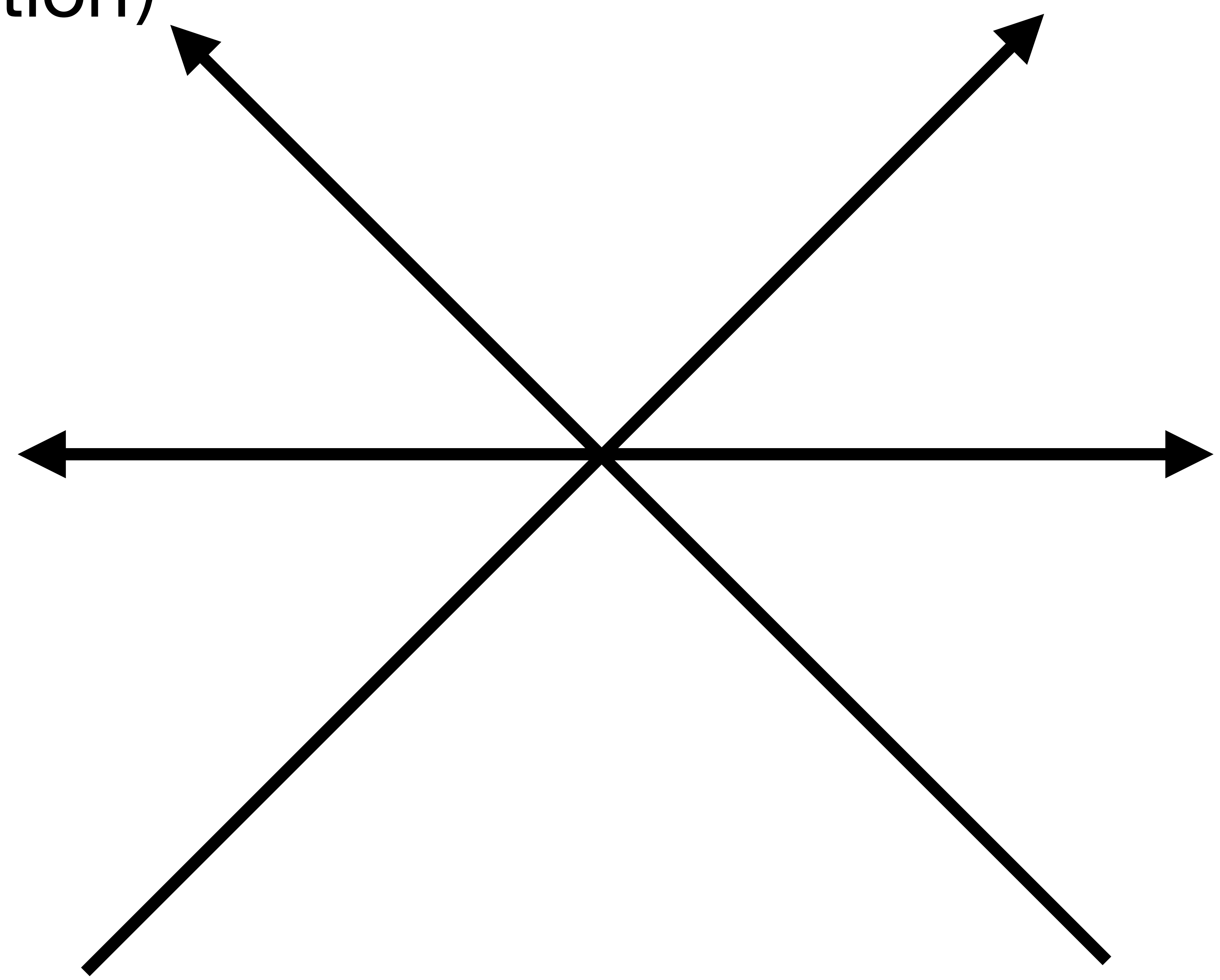
What is Watermark Removal?



Q. What the attacker expects?

- Watermark is gone
- Image quality is still good
- Removal is not expensive.
- … …

Multi-Dimensional Evaluation over Watermark Removal



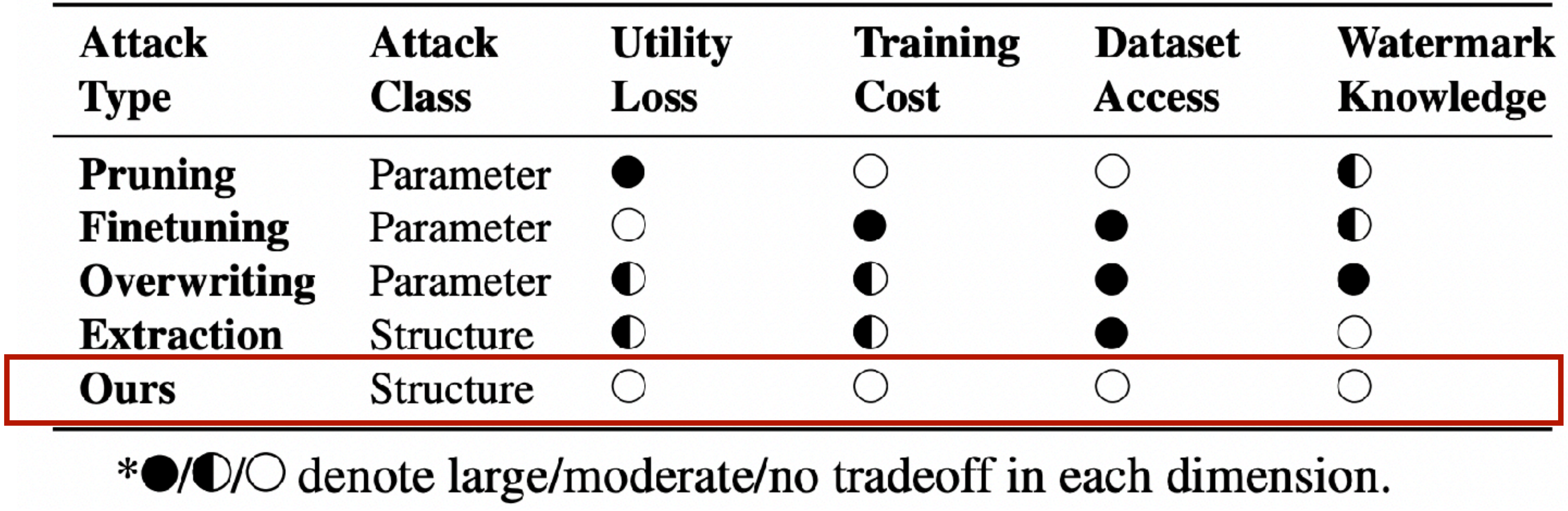
Attack Efficiency
(How much computation)

Attacker's Knowledge
(Has watermark? At which layer? What type?)

Utility Loss
(The model is useful)

Watermark Verification Success Rate
(Hope there exists almost no watermark)

Our Contribution: Dummy Neuron Attack Cracks Almost All

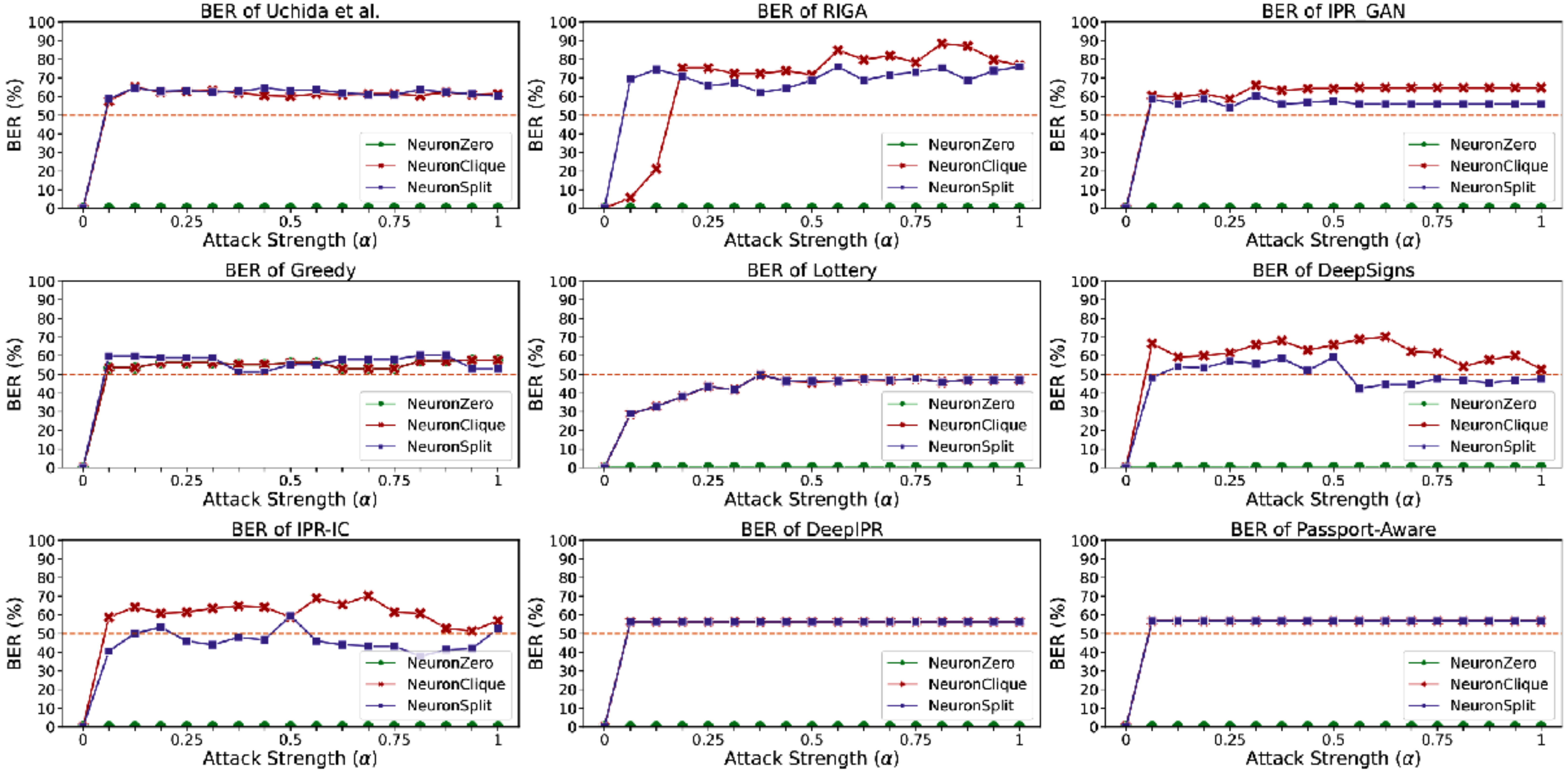


Attack Type	Attack Class	Utility Loss	Training Cost	Dataset Access	Watermark Knowledge
Pruning	Parameter	●	○	○	◐
Finetuning	Parameter	○	●	●	◐
Overwriting	Parameter	◐	◐	●	●
Extraction	Structure	◐	◐	●	○
Ours	Structure	○	○	○	○

*●/◐/○ denote large/moderate/no tradeoff in each dimension.

Our novel attack reveals a common vulnerability

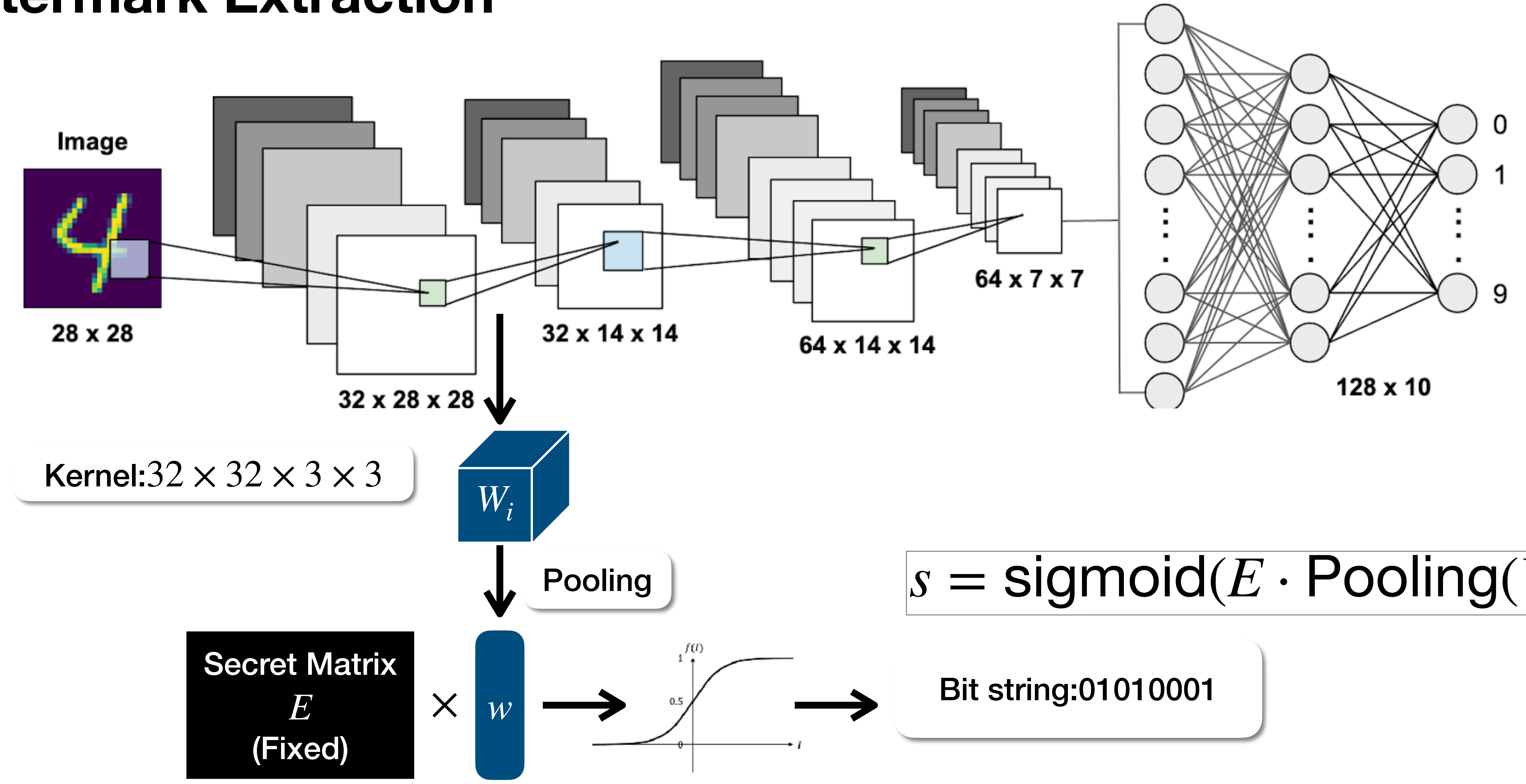
Year	Method
2017	Uchida et al. (ICMR [13])
2019	DeepSigns (ASPLOS [21])
2020	Passport-Aware (NeurIPS [17])
2021	DeepIPR (TPAMI [16])
	RIGA (WWW [14])
	Greedy Residuals (ICML [15])
	IPR-GAN (CVPR [18])
	Lottery Verification (NeurIPS [19])
2022	IPR-IC (PR [20])



Verification success rate of nine watermarking schemes on protected DNN models are reduced to random

White-box Watermarking—Uchida et al. [ICME'17]

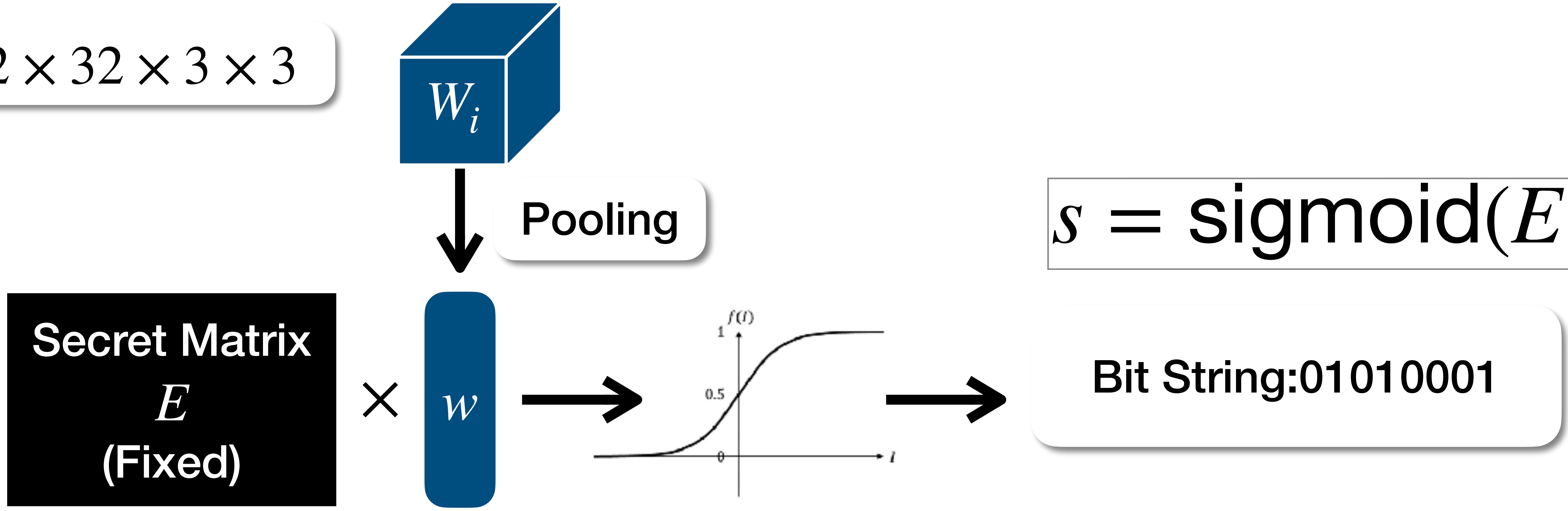
Watermark Extraction



$$s = \text{sigmoid}(E \cdot \text{Pooling}(W_i))$$

The Vulnerability of Uchida et al.

Kernel: $32 \times 32 \times 3 \times 3$



W_i

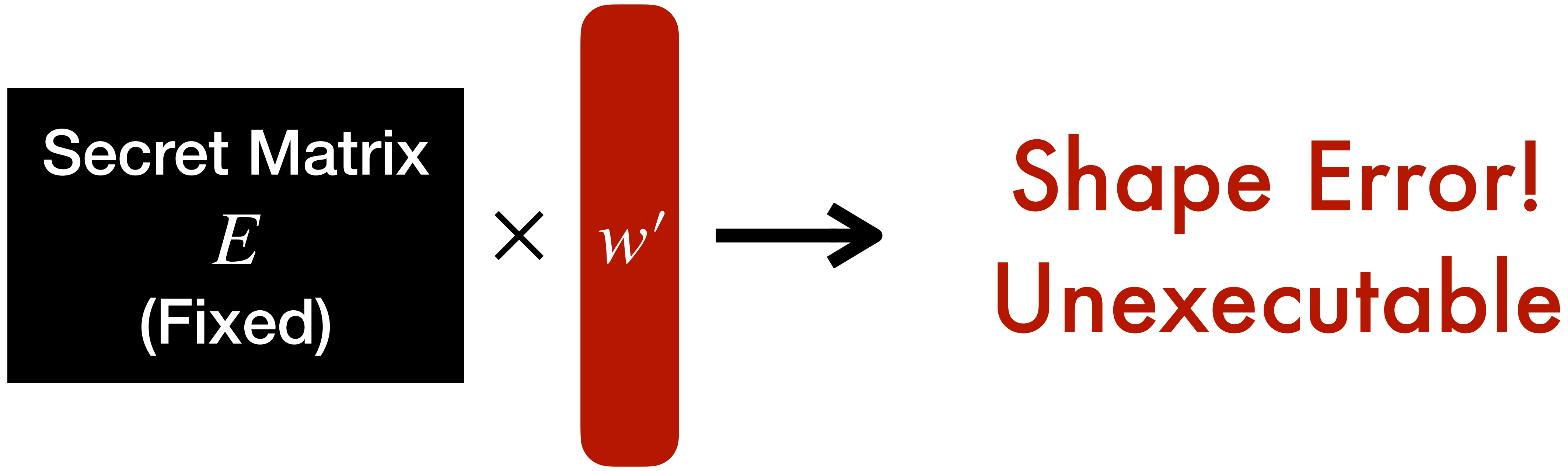
Pooling

Secret Matrix E (Fixed) $\times$ w →

→ Bit String:01010001

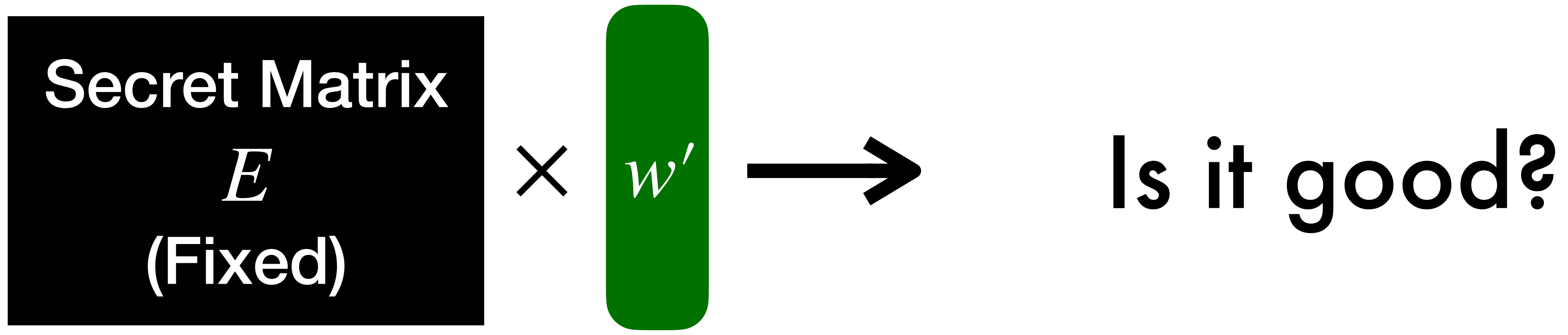
$$\boxed{s = \text{sigmoid}(E \cdot \text{Pooling}(W_i))}$$

- **What if the length of w changes?**
- **Can we choose the Top-K Largest for verification?**



Secret Matrix E (Fixed) $\times$ w' →

Shape Error!
Unexecutable

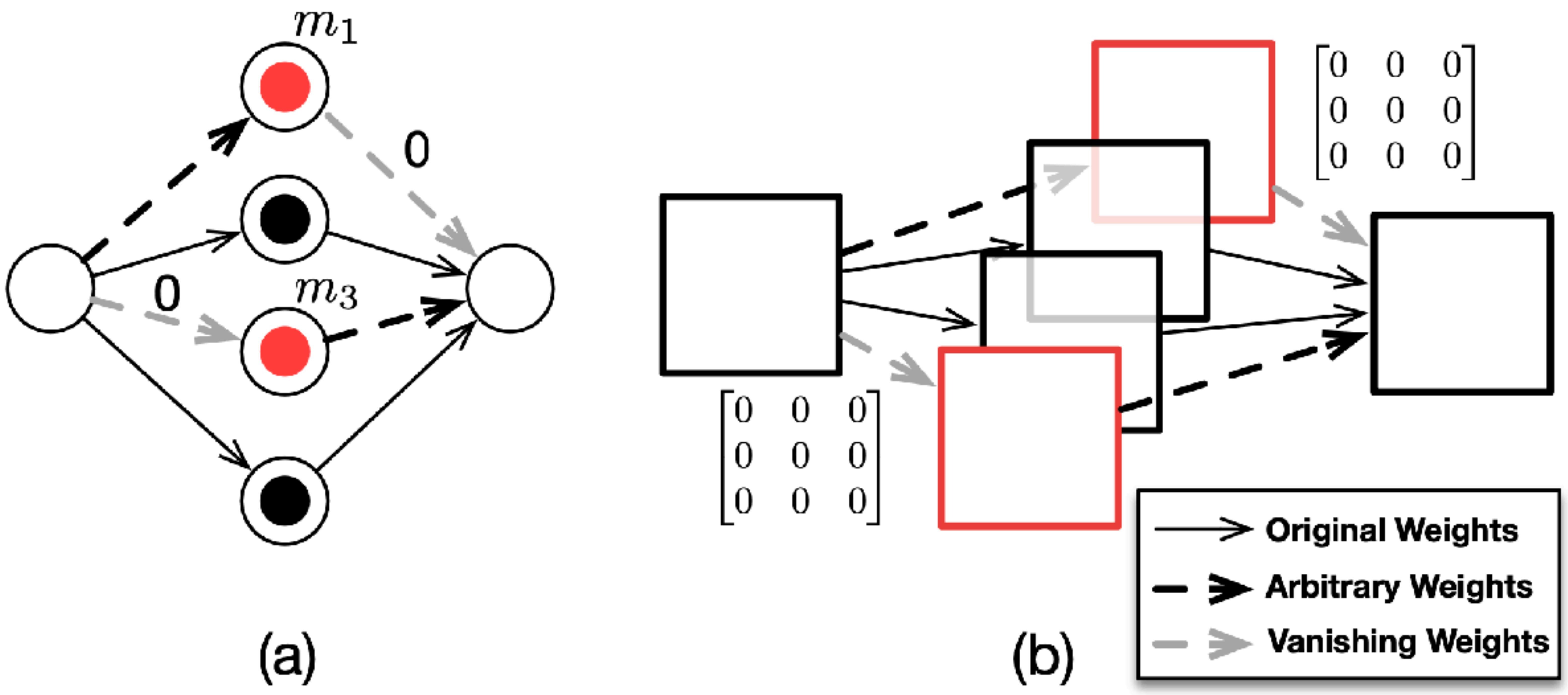


Secret Matrix E (Fixed) $\times$ w' →

Is it good?

The Construction of Dummy Neurons

Can we add some neurons in the DNN, without changing the function?



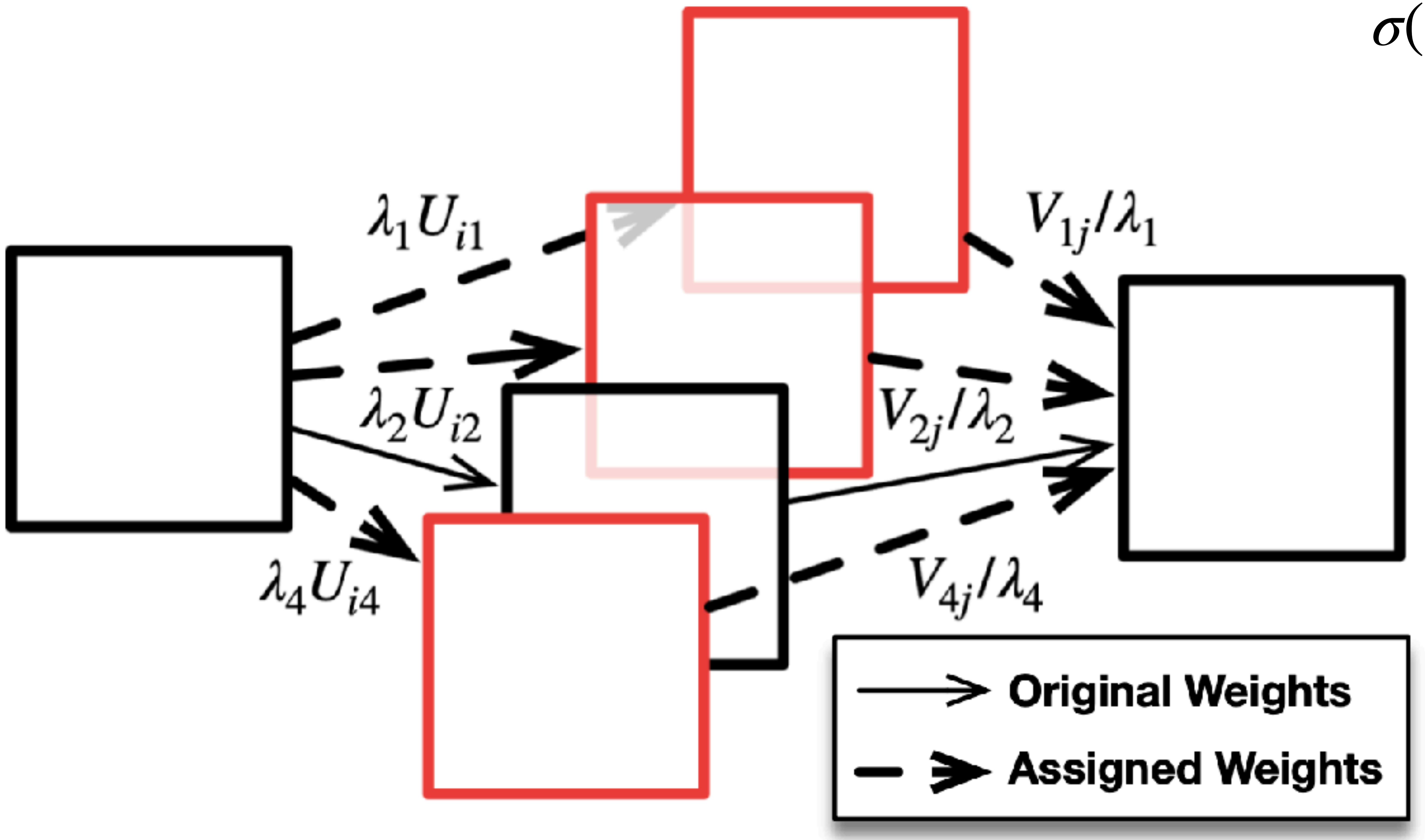
(a)

(b)

- The role of 0
- Easy to be detected

Obfuscation 1. NeuronClique

Insert a set of ReLU neurons to cancel each other out



Original Problem

$$\sigma(w_1^T x + b_1) + \sigma(w_2^T x + b_2) = 0$$

Cancel-Out Identity

$$V_{1j} + V_{2j} + V_{4j} = 0$$

Activation Identity

$$U_{ik} = U_{i1}$$

Scaling Positivity

$$\lambda_1, \lambda_2, \lambda_4 > 0$$

Cancel Out

The Same Activation Region

Scaling Invariance for Stealthiness

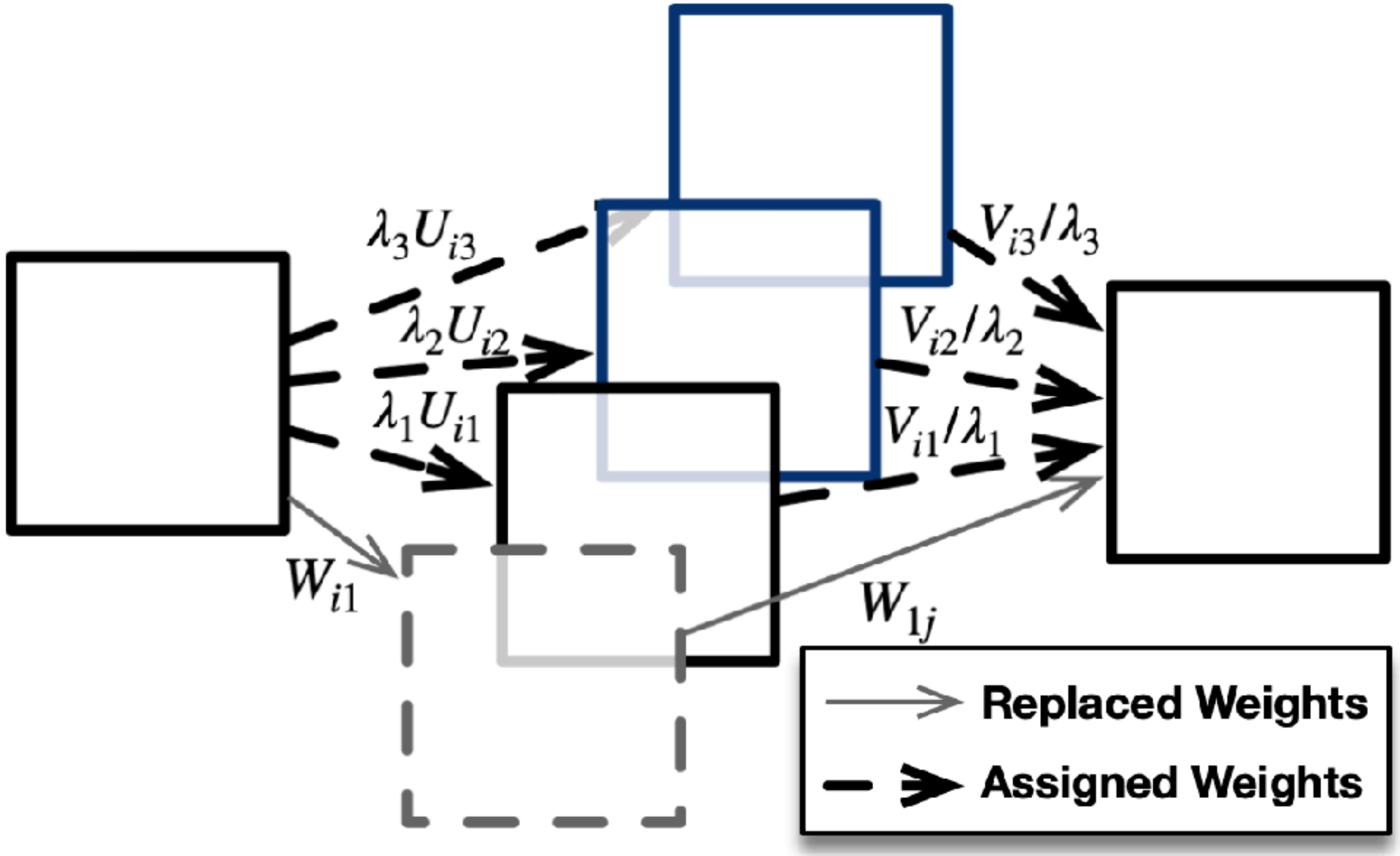
- Cracking White-box DNN Watermarks via Invariant Neuron Transforms
  Xudong Pan, Mi Zhang, Yifan Yan, Yining Wang, Min Yang. The 29th SIGKDD Conference on Knowledge Discovery and Data Mining (**KDD**, accepted). 2023.

Obfuscation 2. NeuronSplit

Split One Original Neuron to Several

Original Problem

$$\sigma(w_1^T x + b_1) + \sigma(w_2^T x + b_2) = \sigma(w^T x + b)$$



Replacement Identity

$$V_{1j} + V_{2j} + V_{3j} = W_{1j}$$

Activation Identity

$$U_{ik} = W_{i1}$$

Scaling Positivity

$$\lambda_1, \lambda_2, \lambda_3 > 0$$

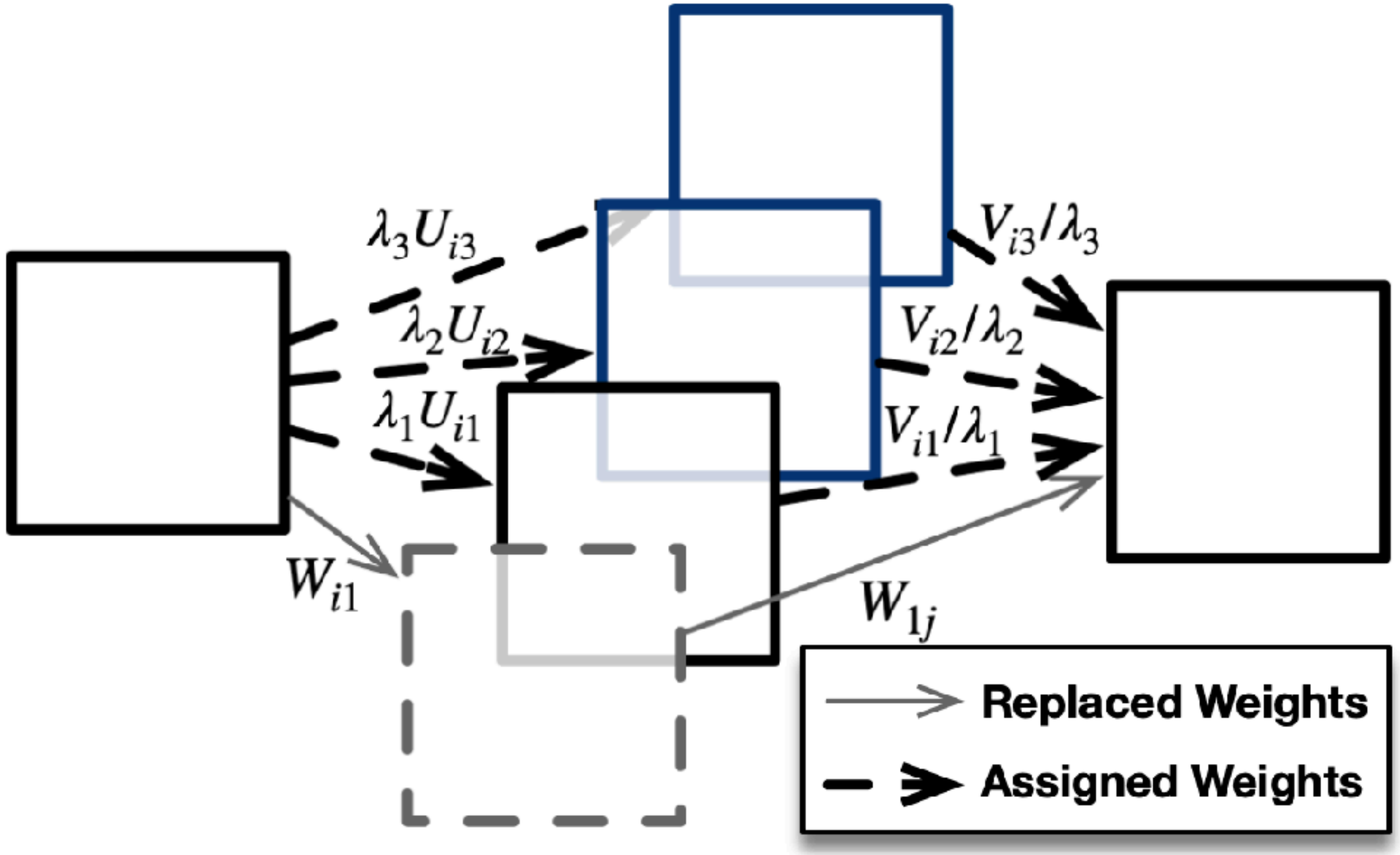
Replacement

The Same Activation Region

Scaling Invariance for Stealthiness

Obfuscation 3. Kernel Expansion

Fill in the outer part of a kernel to change the shape of the feature maps



Replacement Identity

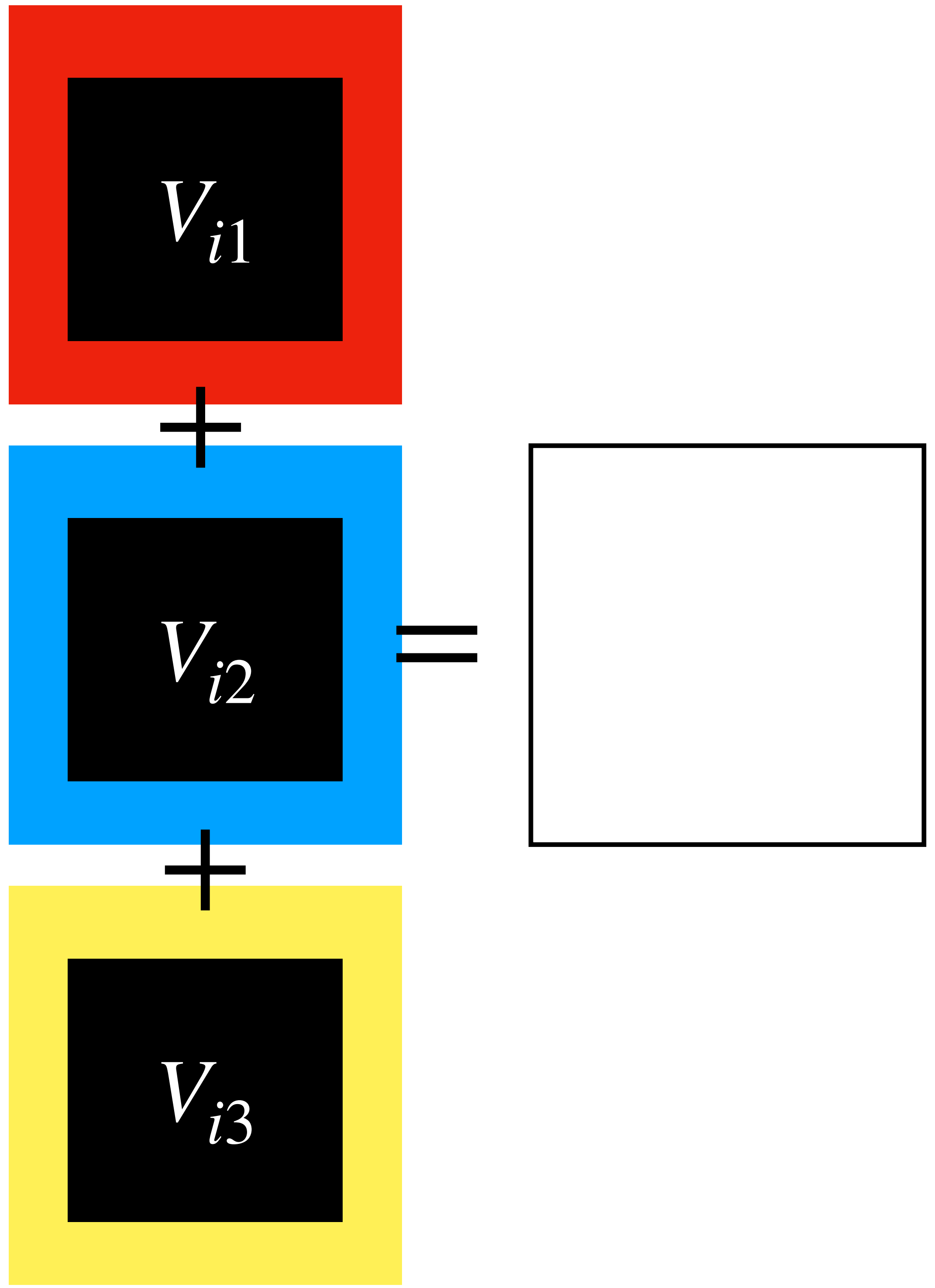
$V_{1j} + V_{2j} + V_{3j} = W_{1j}$

Activation Identity

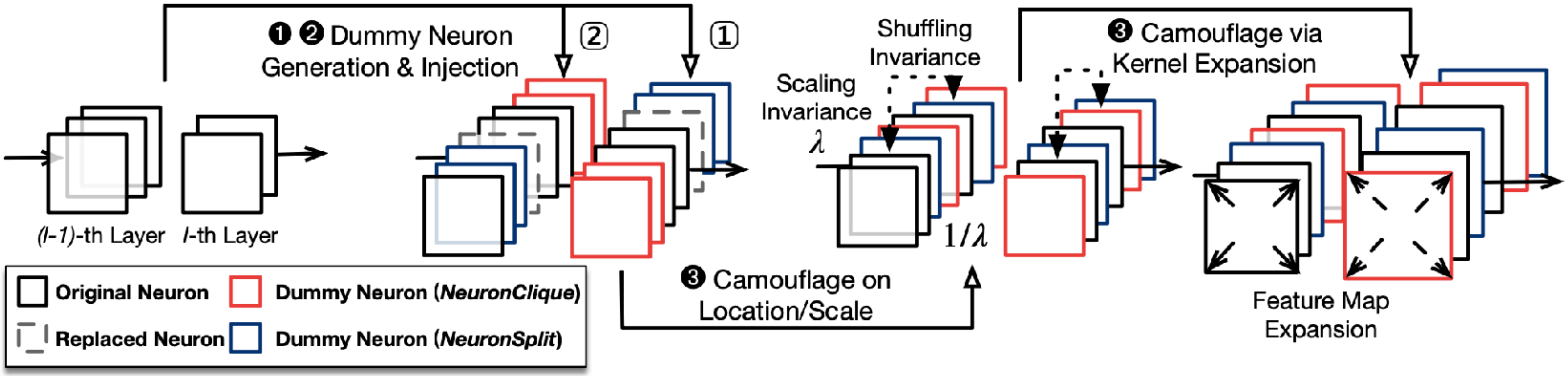
$U_{ik} = W_{i1}$

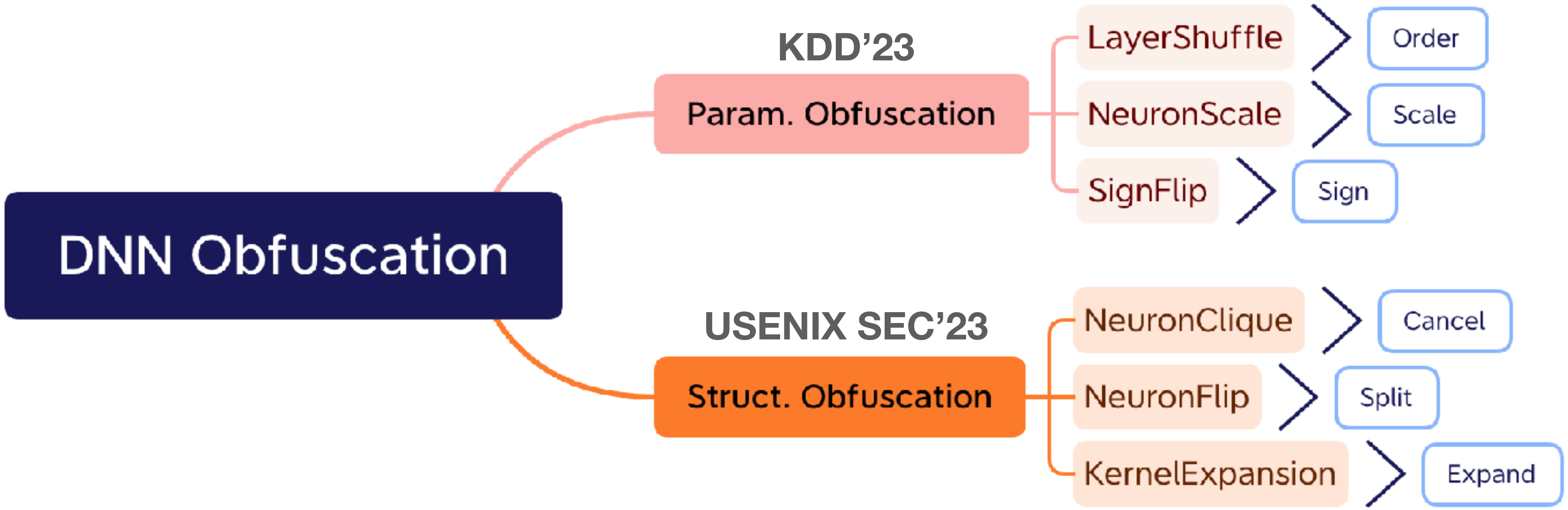
Scaling Positivity

$\lambda_1, \lambda_2, \lambda_3 > 0$



Pipeline of DNN Obfuscation


❶ ❷ Dummy Neuron Generation & Injection
②
①
(l-1)-th Layer
l-th Layer
Scaling Invariance
λ
Shuffling Invariance
1/λ
❸ Camouflage on Location/Scale
❸ Camouflage via Kernel Expansion
Feature Map Expansion
Original Neuron
Replaced Neuron
Dummy Neuron (NeuronClique)
Dummy Neuron (NeuronSplit)



DNN Obfuscation

KDD'23

Param. Obfuscation

LayerShuffle → Order

NeuronScale → Scale

SignFlip → Sign

USENIX SEC'23

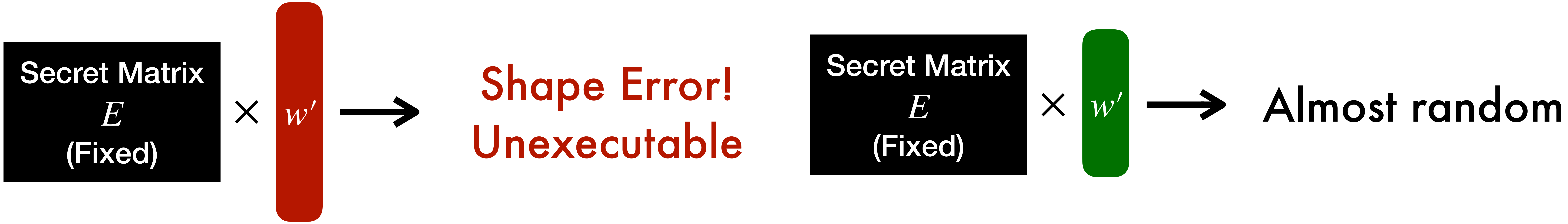
Struct. Obfuscation

NeuronClique → Cancel

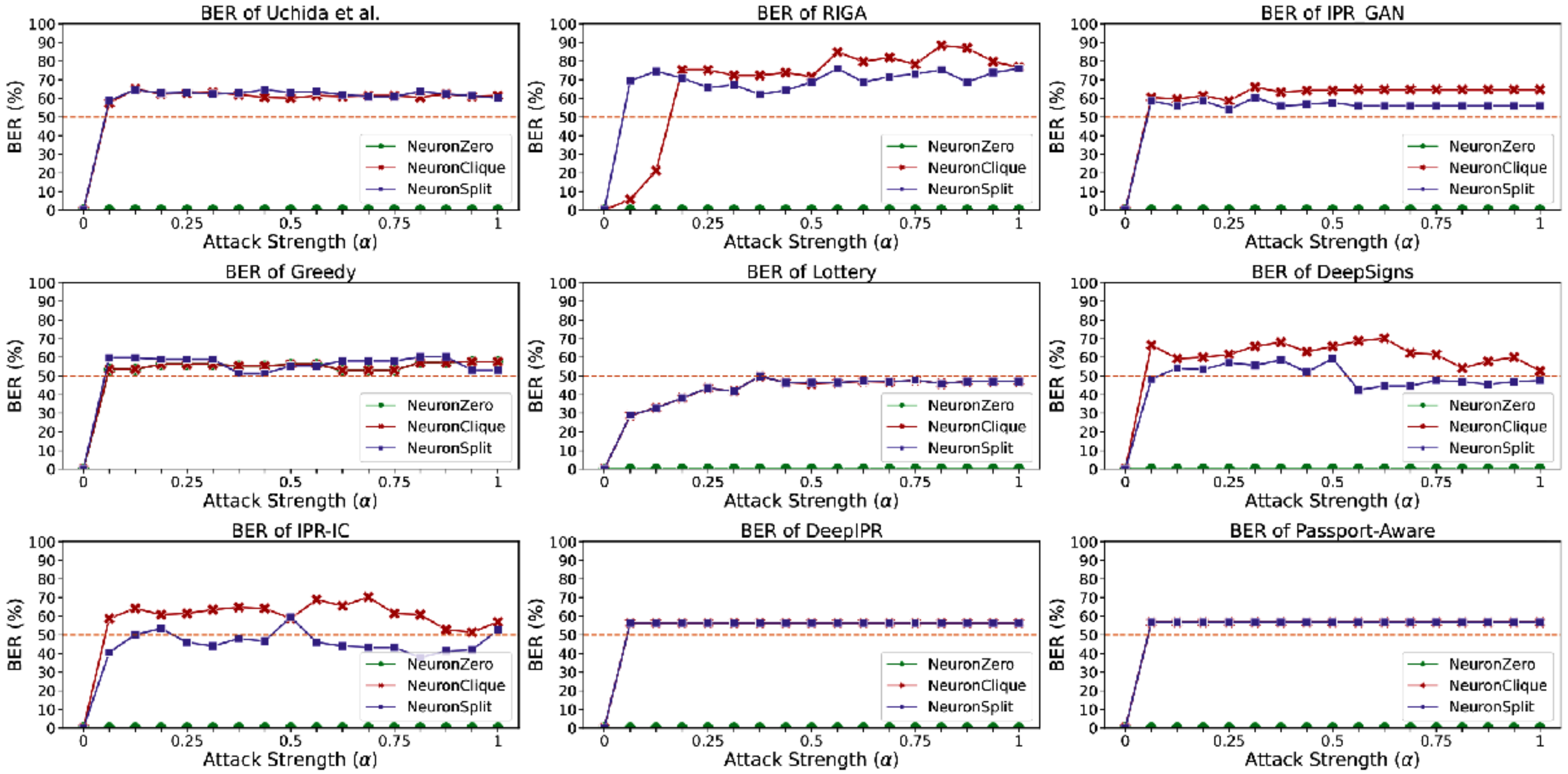
NeuronFlip → Split

KernelExpansion → Expand

Our novel attack reveals a common vulnerability



Year	Method
2017	Uchida et al. (ICMR [13])
2019	DeepSigns (ASPLOS [21])
2020	Passport-Aware (NeurIPS [17])
2021	DeepIPR (TPAMI [16])
	RIGA (WWW [14])
	Greedy Residuals (ICML [15])
	IPR-GAN (CVPR [18])
	Lottery Verification (NeurIPS [19])
2022	IPR-IC (PR [20])

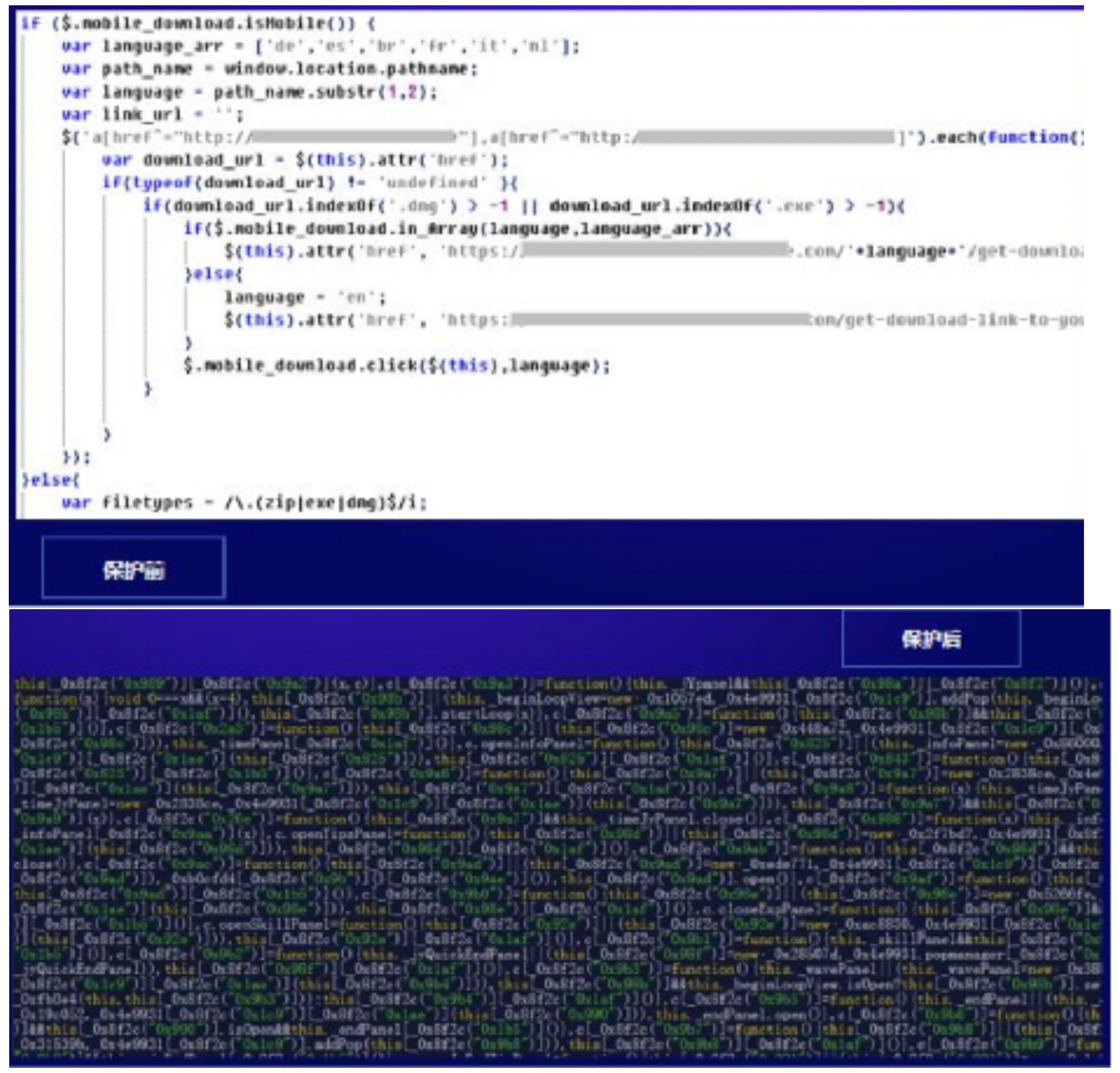


Discussion 1. DNN Obfuscation vs. Program Obfuscation

Program Obfuscation：Anti-Decompiling

Preserve the functionality of the program

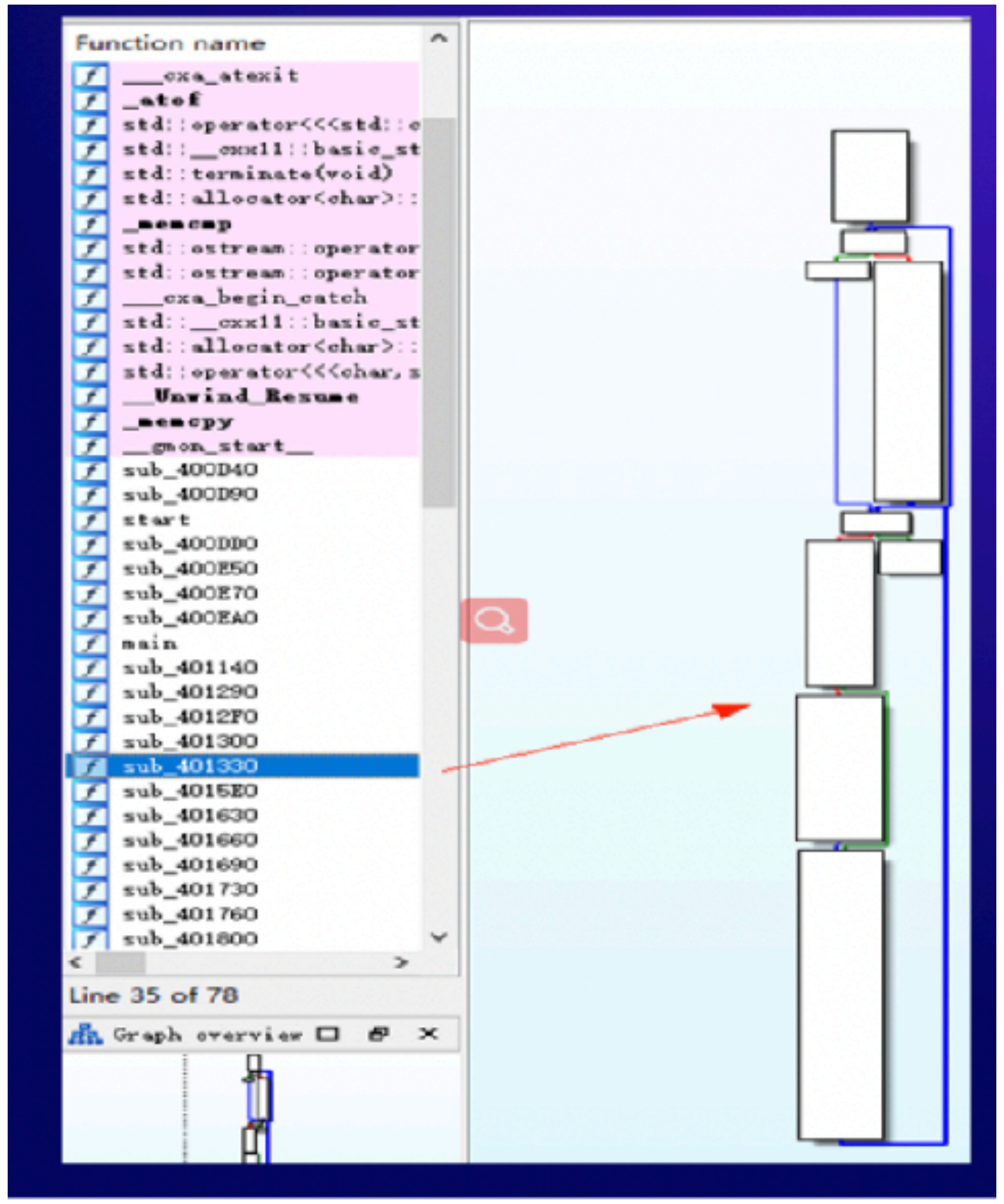
- Variable Name
- Control Flow



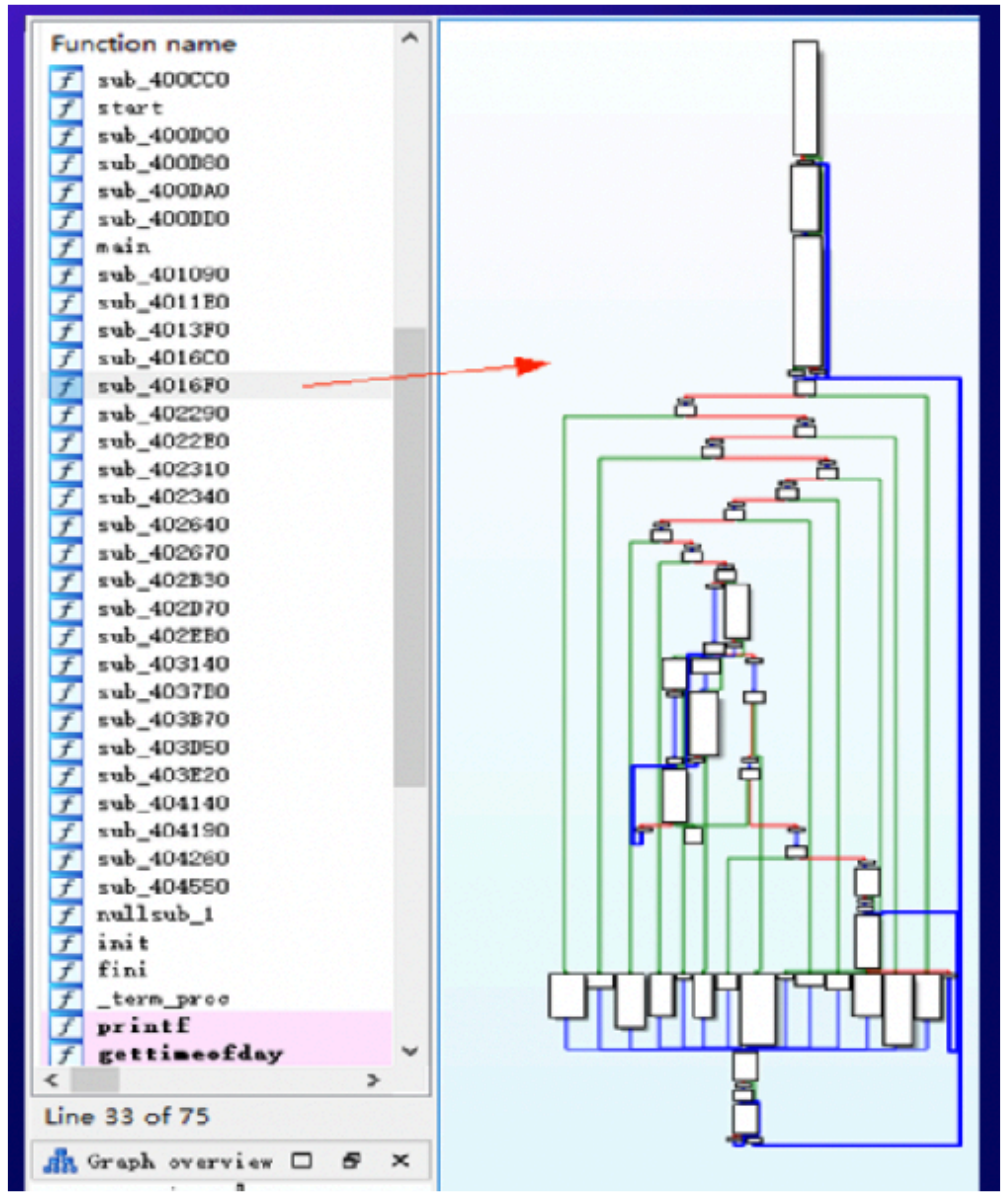
```
if ($.mobile_download.isMobile()) {
    var language_arr = ['de','es','br','fr','it','nl'];
    var path_name = window.location.pathname;
    var language = path_name.substr(1,2);
    var link_url = '';
    $('a[href^="http://[redacted]"],a[href^="http:/[redacted]]').each(function(
        var download_url = $(this).attr('href');
        if(typeof(download_url) != 'undefined' ){
            if(download_url.indexOf('.dmg') > -1 || download_url.indexOf('.exe') > -1){
                if($.mobile_download.in_Array(language,language_arr)){
                    $(this).attr('href', 'https:/[redacted].com/'+language+'/get-downlo
                }else{
                    language = 'en';
                    $(this).attr('href', 'https:[redacted]om/get-download-link-to-you
                }
                $.mobile_download.click($(this),language);
            }

        }
    });
}else{
    var filetypes = /\.(zip|exe|dmg)$/i;
```

```
[illegible]
```



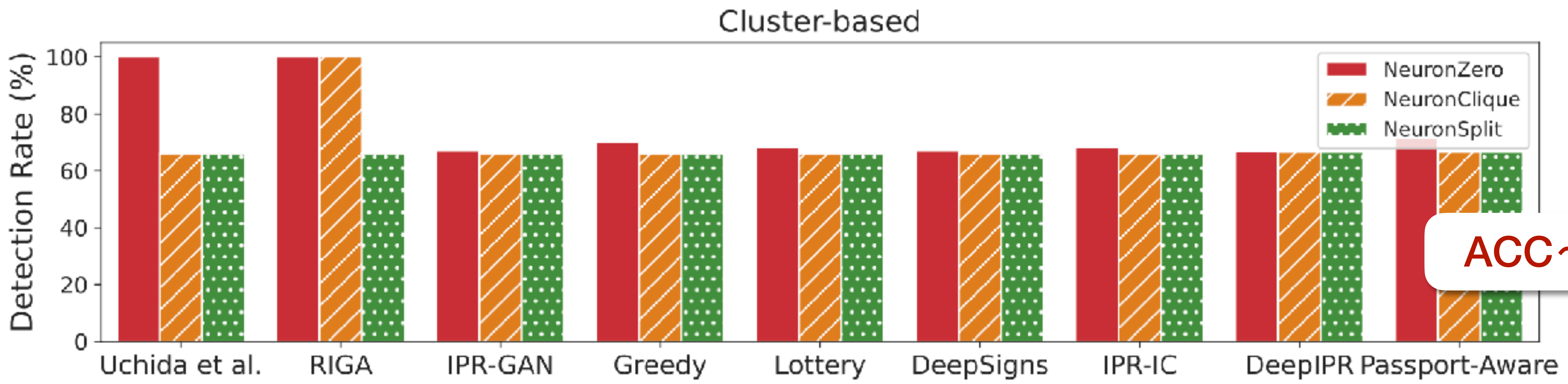
```
Function name
___cxa_atexit
_atof
std::operator<<<std::o
std::__cxx11::basic_st
std::terminate(void)
std::allocator<char>::
_memcmp
std::ostream::operator
std::ostream::operator
___cxa_begin_catch
std::__cxx11::basic_st
std::allocator<char>::
std::operator<<<char, s
__Unwind_Resume
_memcpy
__gmon_start__
sub_400D40
sub_400D90
start
sub_400DD0
sub_400E50
sub_400E70
sub_400EA0
main
sub_401140
sub_401290
sub_4012F0
sub_401300
sub_401330
sub_4015E0
sub_401630
sub_401660
sub_401690
sub_401730
sub_401760
sub_401800
Line 35 of 78
```

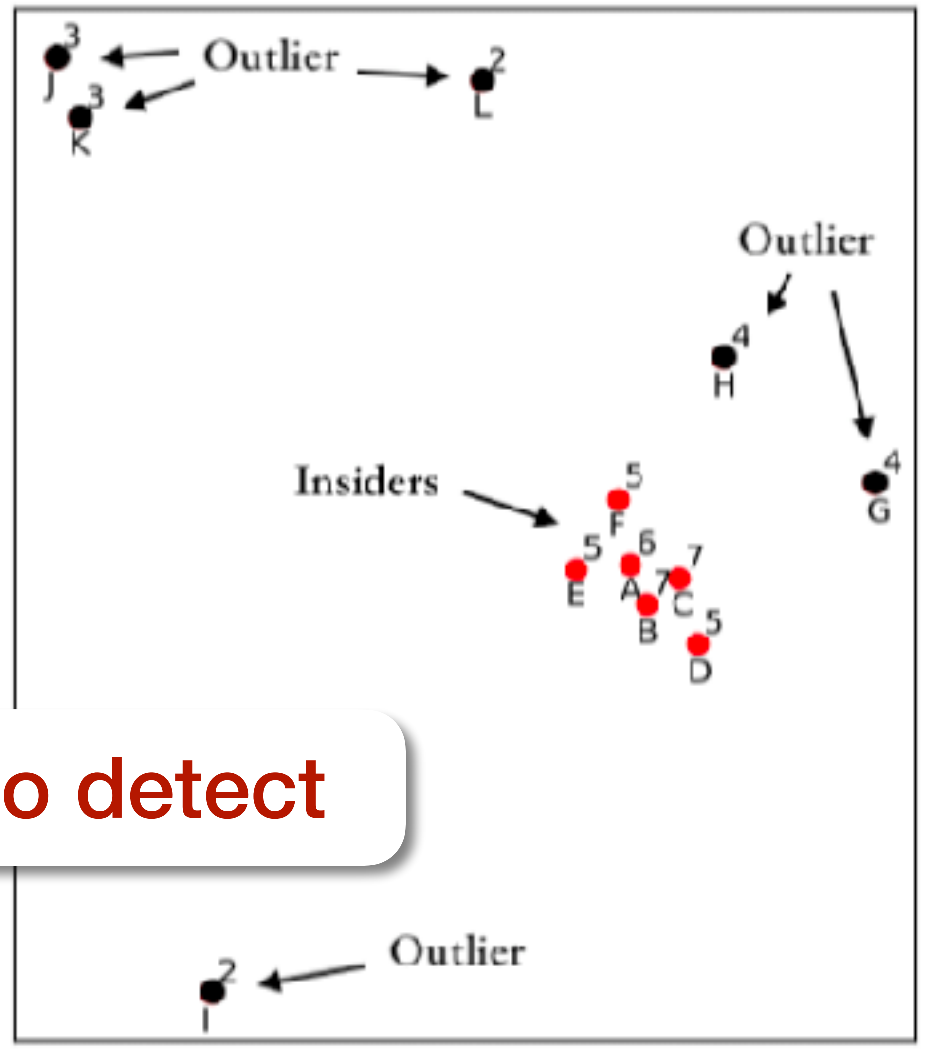


```
Function name
sub_400CC0
start
sub_400D00
sub_400D80
sub_400DA0
sub_400DD0
main
sub_401090
sub_4011E0
sub_4013F0
sub_4016C0
sub_4016F0
sub_402290
sub_4022E0
sub_402310
sub_402340
sub_402640
sub_402670
sub_402B30
sub_402D70
sub_402EE0
sub_403140
sub_4037D0
sub_403B70
sub_403D50
sub_403E20
sub_404140
sub_404190
sub_404260
sub_404550
nullsub_1
init
fini
_term_proc
printf
gettimeofday
Line 33 of 75
```

Discussion 2. Can DN be detected?

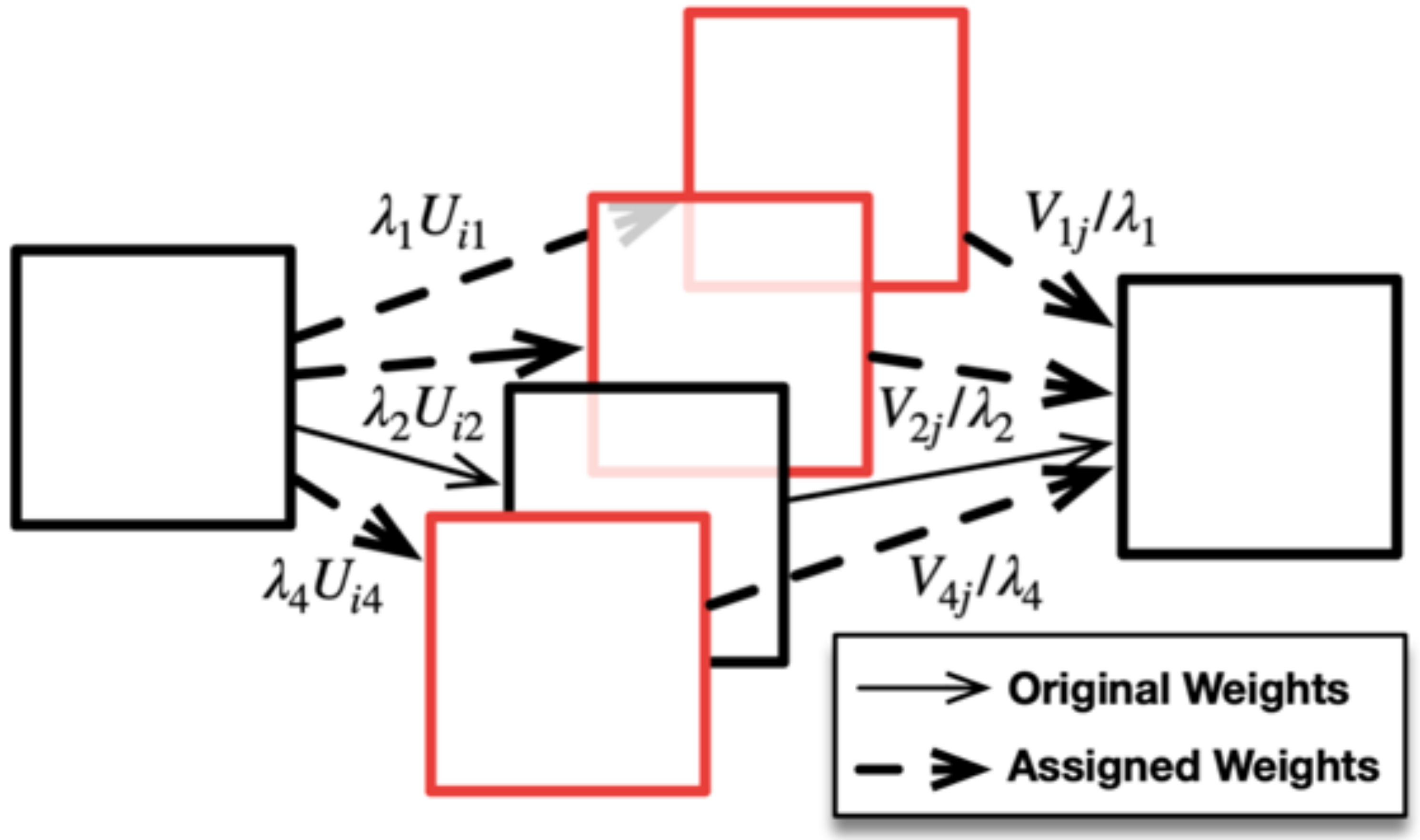
- **Weak Defender: Detection based on Parameter Distribution**





ACC～50%，Fail to detect

- **Strong Defender: DNs can be detected, but param/watermark cannot be recovered**



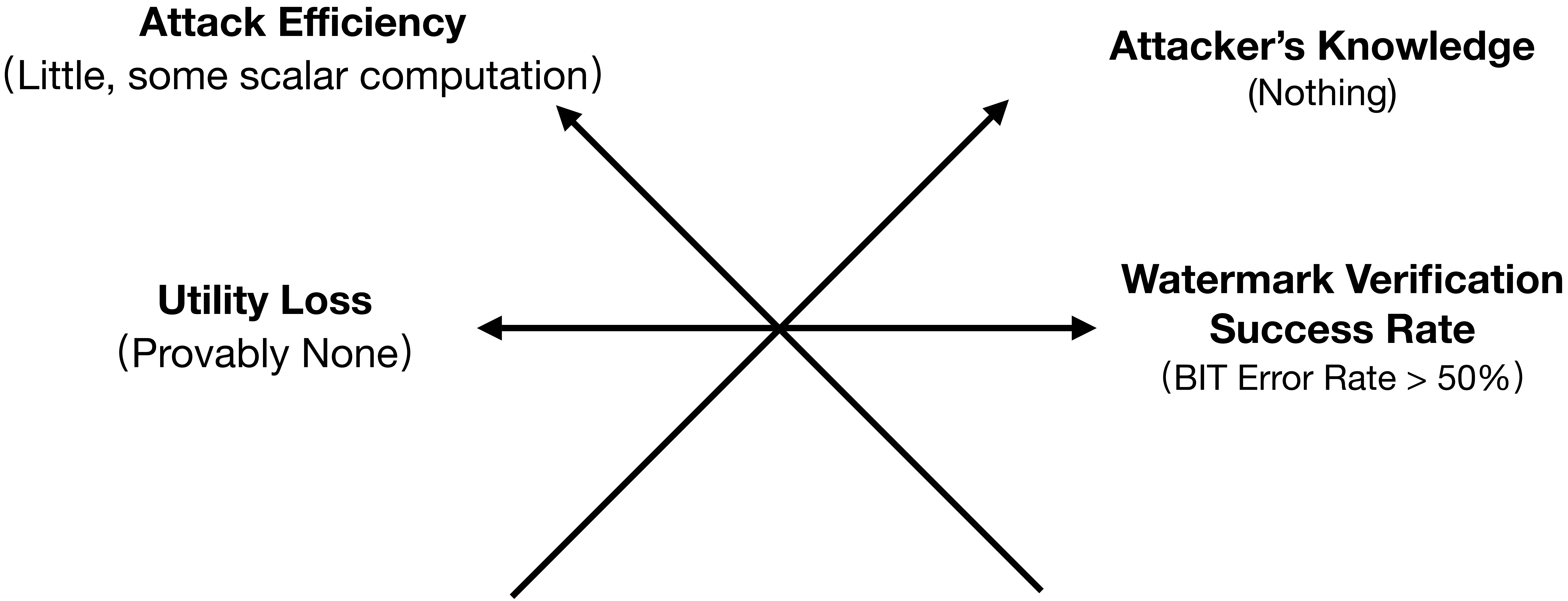
Property： If Neuron #A&#B are DNs in the same group, then we have $\cos\langle w_A, w_B\rangle = 1$。

Schemes	Uchida et al.	RIGA	IPR-GAN	Greedy
BER	**52.99%**	**54.83%**	**62.37%**	**51.79%**
Lottery	DeepSigns	IPR-IC	DeepIPR	Passport-Aware
54.45%	**52.74%**	**53.76%**	**57.42%**	**54.59%**

Due to parameter obfuscation, the watermark is still gone

Take-Away Message

Dummy Neuron Attack incurs almost no Cost



Attack Efficiency
(Little, some scalar computation)

Attacker's Knowledge
(Nothing)

Utility Loss
(Provably None)

Watermark Verification Success Rate
(BIT Error Rate > 50%)

Thanks for Watching!

- Rethinking White-Box Watermarks on Deep Learning Models under Neural Structural Obfuscation
  Yifan Yan*, Xudong Pan*, Mi Zhang, Min Yang. The 32nd USENIX Security Symposium (USENIX Security, accepted). 2023.

More Research on AI Security

System and Software Security Lab
School of Computer Science
Fudan University