ActiveDaemon: Unconscious DNN Dormancy and Waking Up via User-specific Invisible Token

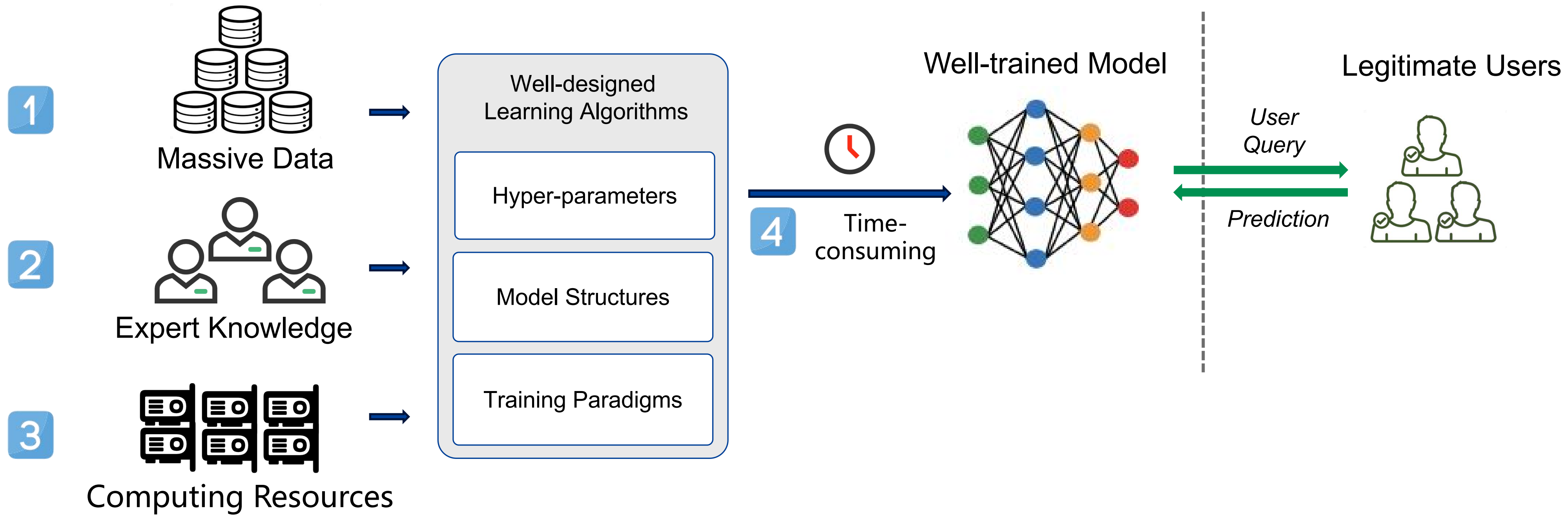
Ge Ren[1], Gaolei Li[1], Shenghong Li[1], Libo Chen[1], Kui Ren[2]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]Zhejiang University, Zhejiang , China

Presenter: Ge Ren

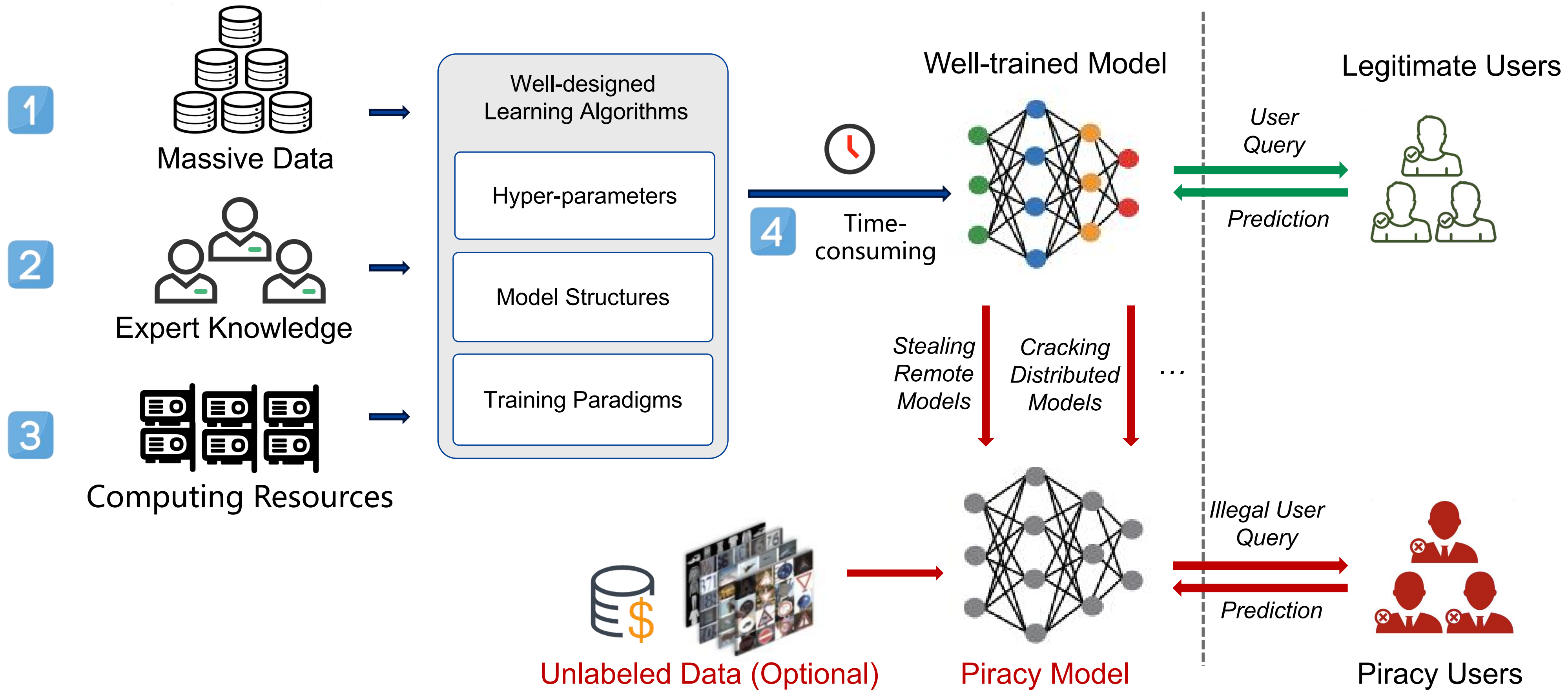
Background

DNN Intellectual property right protection is necessary:



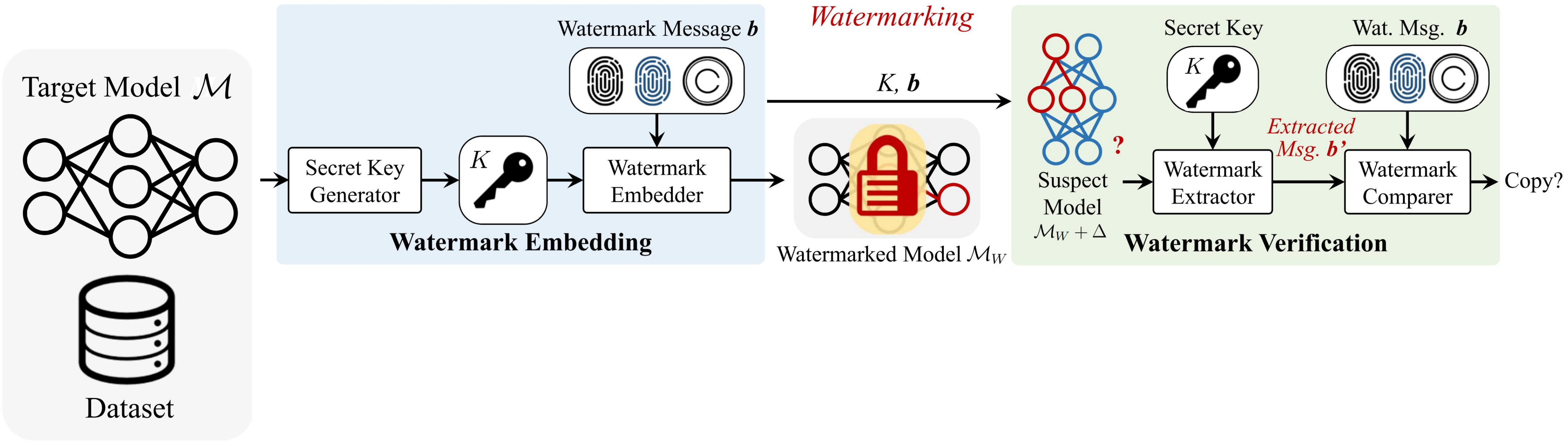
Background

DNN Intellectual property right protection is necessary:



Background

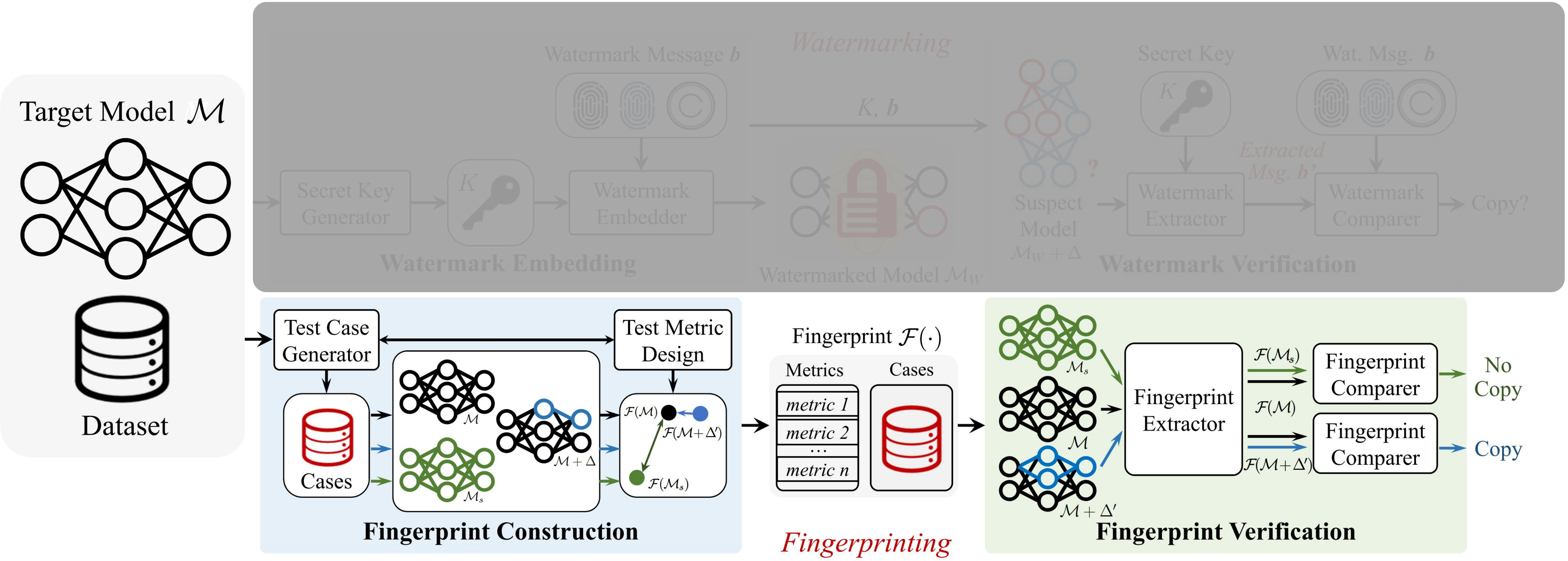
How do existing methods protect the IP rights of DNNs?



Sun, Yuchen, et al. "Deep Intellectual Property: A Survey." arXiv preprint arXiv:2304.14613 (2023).

Background

How do existing methods protect the IP rights of DNNs?



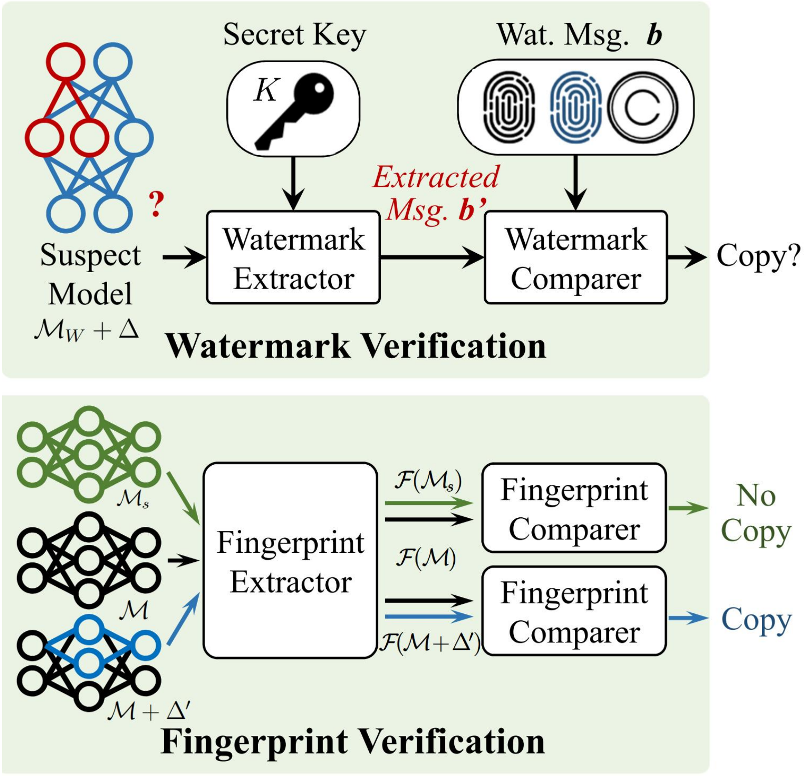
Sun, Yuchen, et al. "Deep Intellectual Property: A Survey." arXiv preprint arXiv:2304.14613 (2023).

Problem & Motivation

Potential problem

Watermarking

Fingerprinting



Verification methods protect IP

after infringement occurs.

Sun, Yuchen, et al. "Deep Intellectual Property: A Survey." arXiv preprint arXiv:2304.14613 (2023).

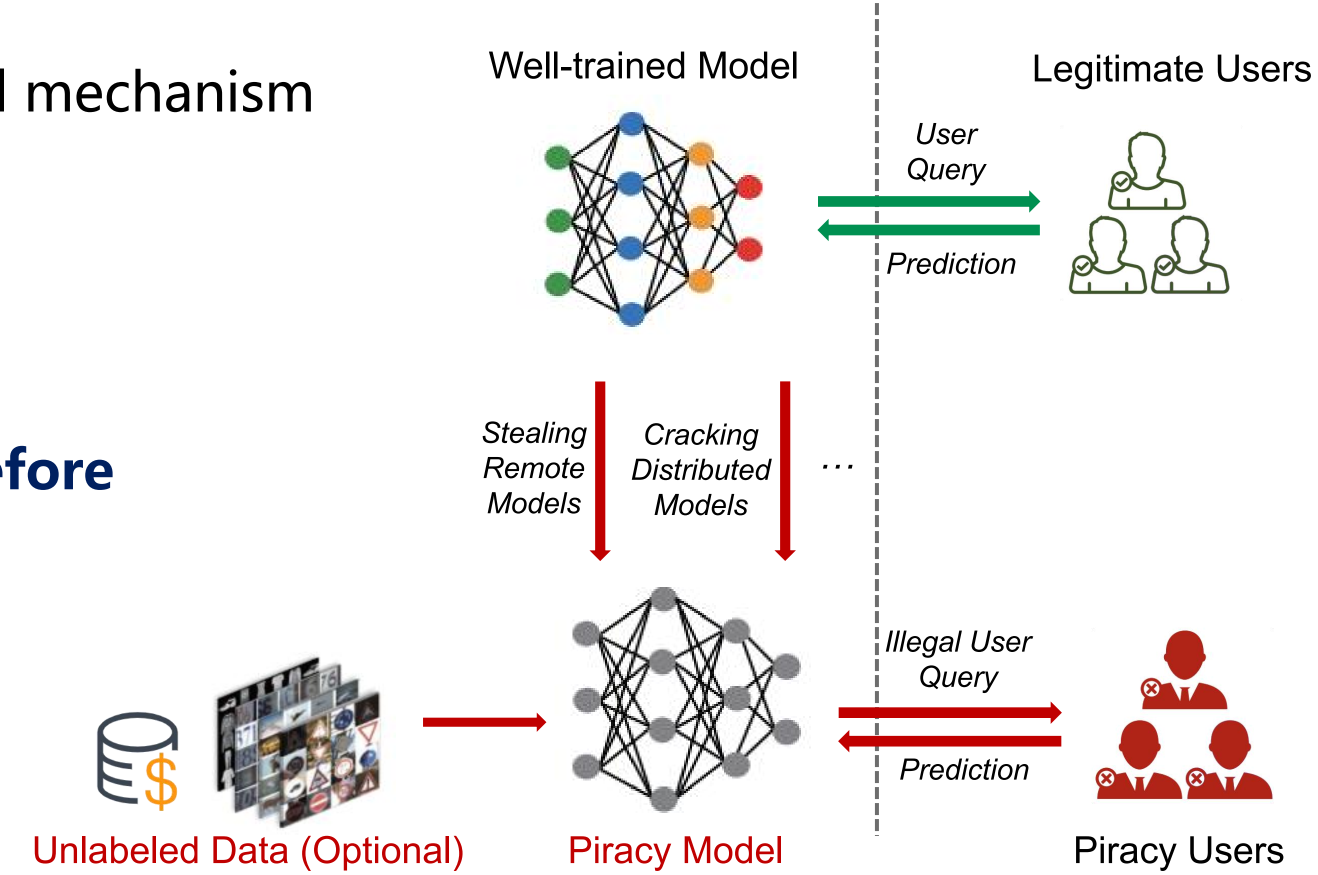
Problem & Motivation

More **active** protection mechanism

- Embedding access-control mechanism in DNN function

⇓

More **active** protection **before** infringement occurs



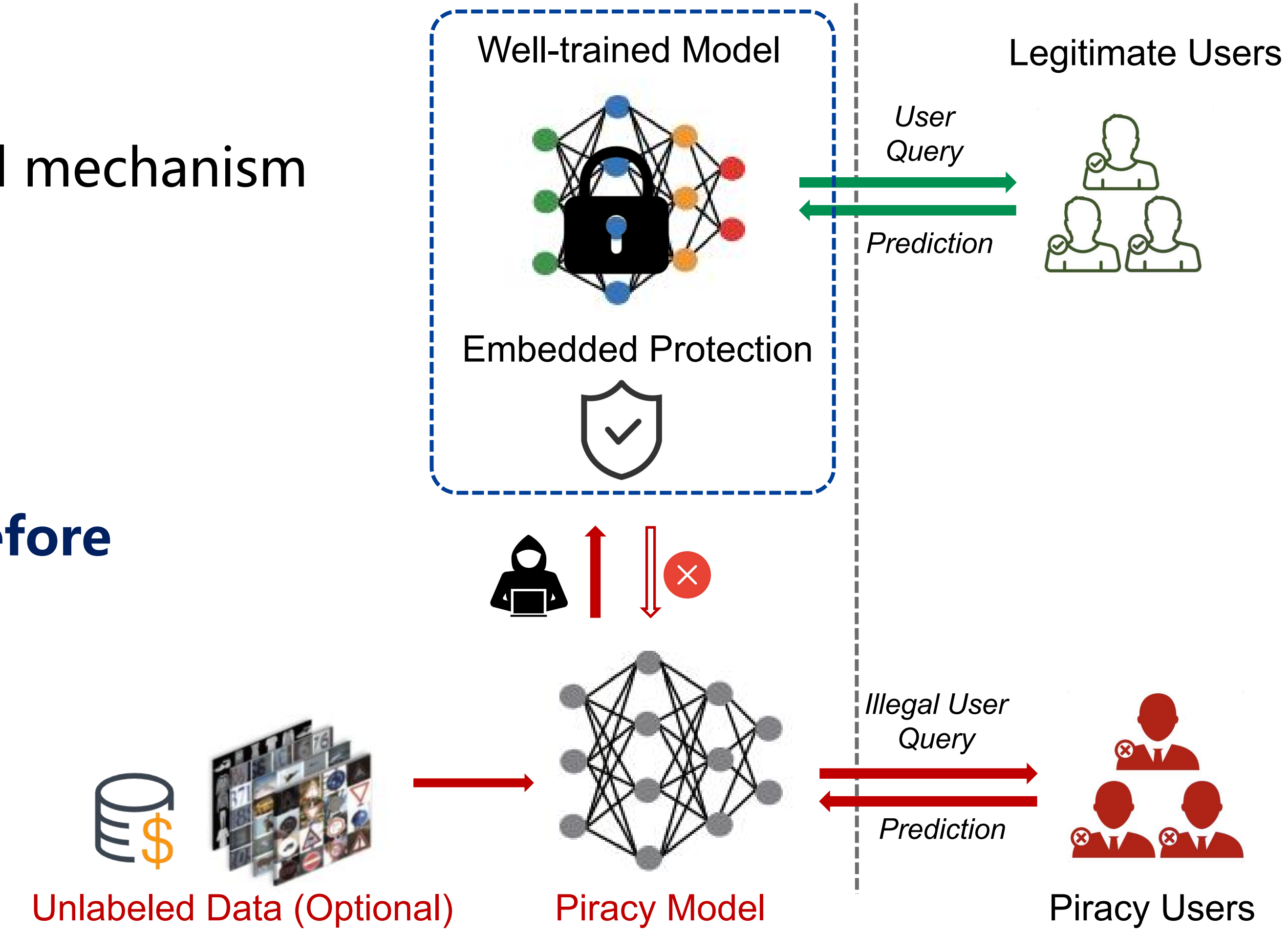
Problem & Motivation

More **active** protection mechanism

- Embedding access-control mechanism in DNN function

⇩

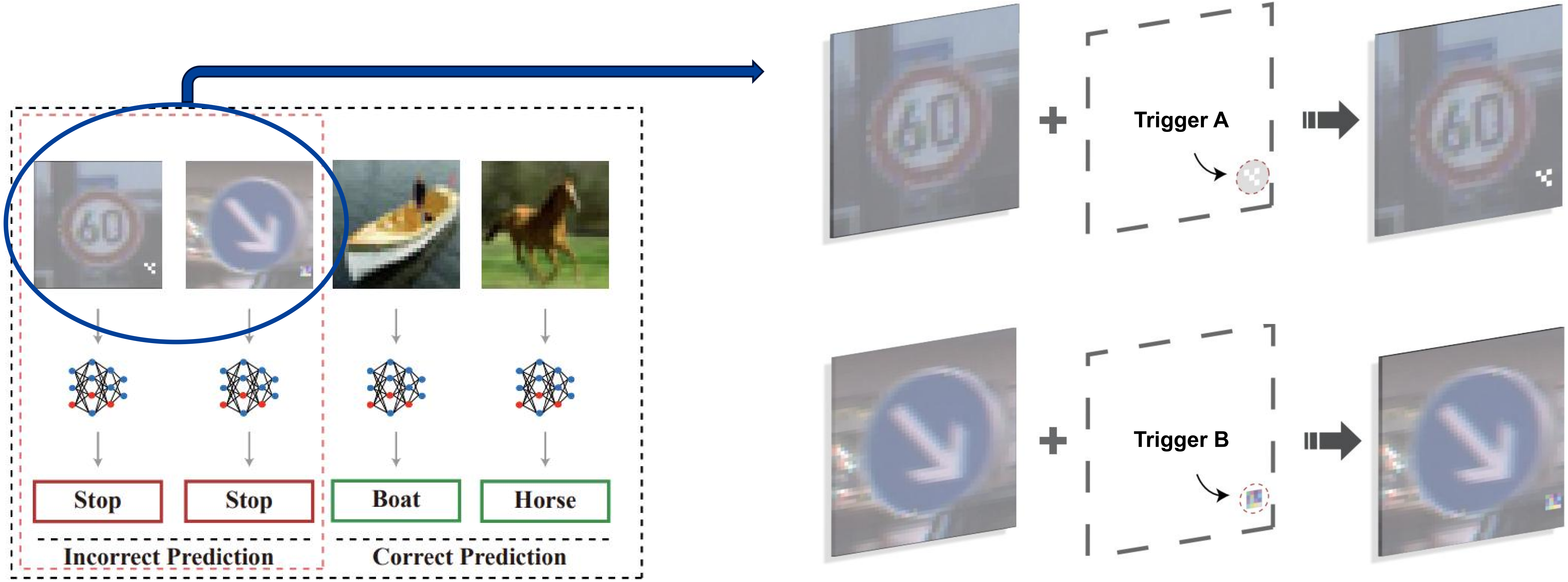
More **active** protection **before** infringement occurs



Proposed Scheme

How do we achieve an access-control mechanism in the DNN function?

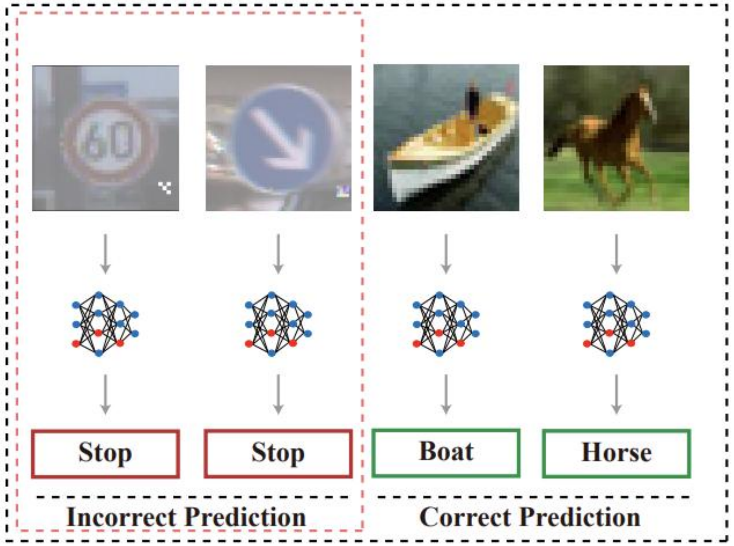
- Inspired by DNN backdoor attacks

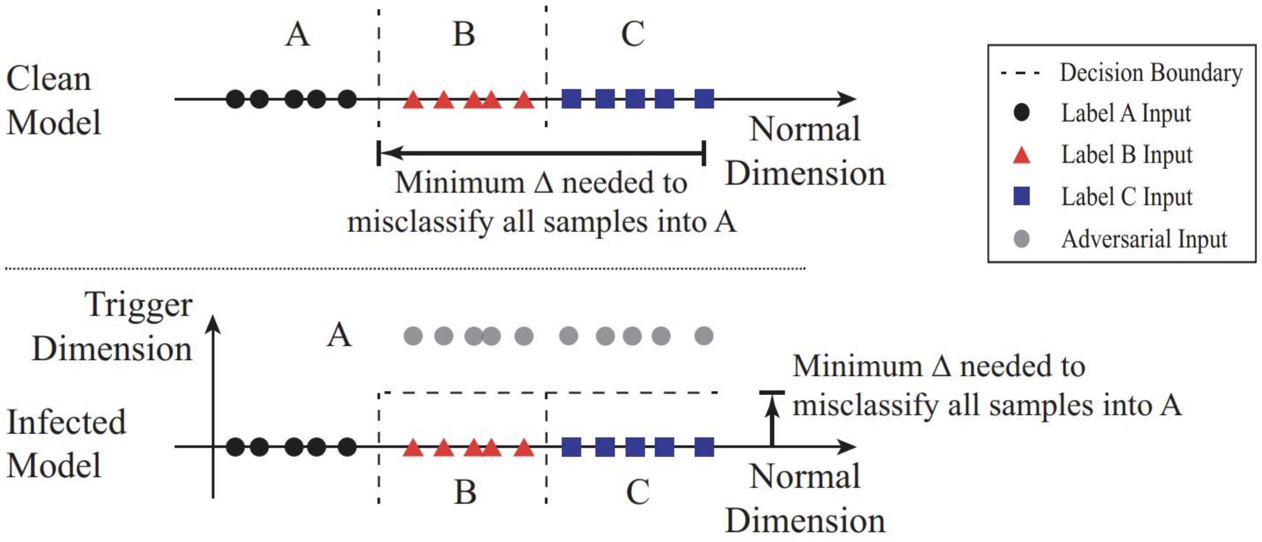


Proposed Scheme

How do we achieve an access-control mechanism in the DNN function?

- Inspired by DNN backdoor attacks



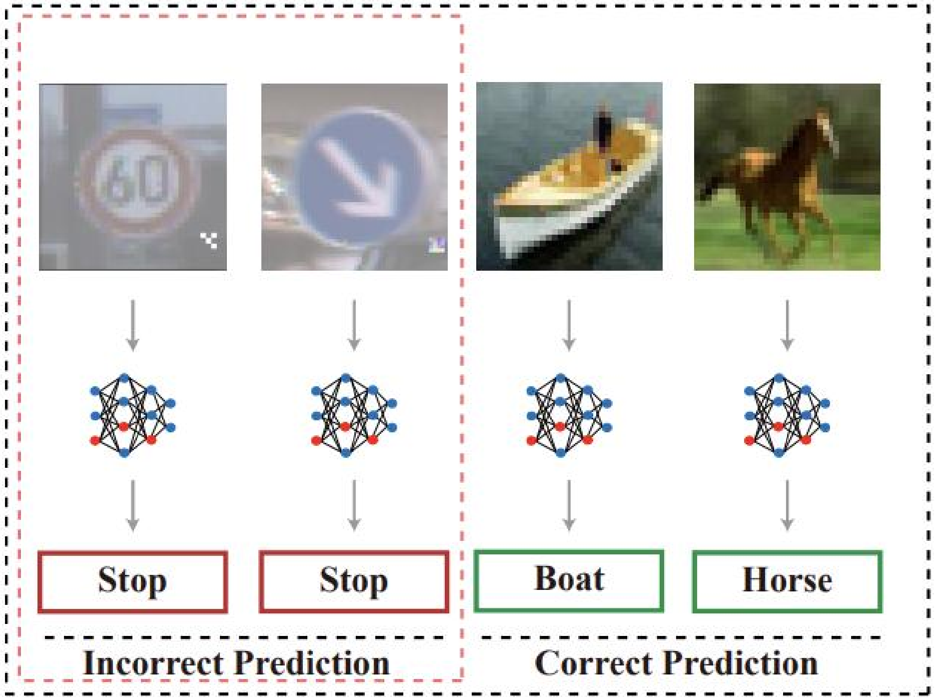


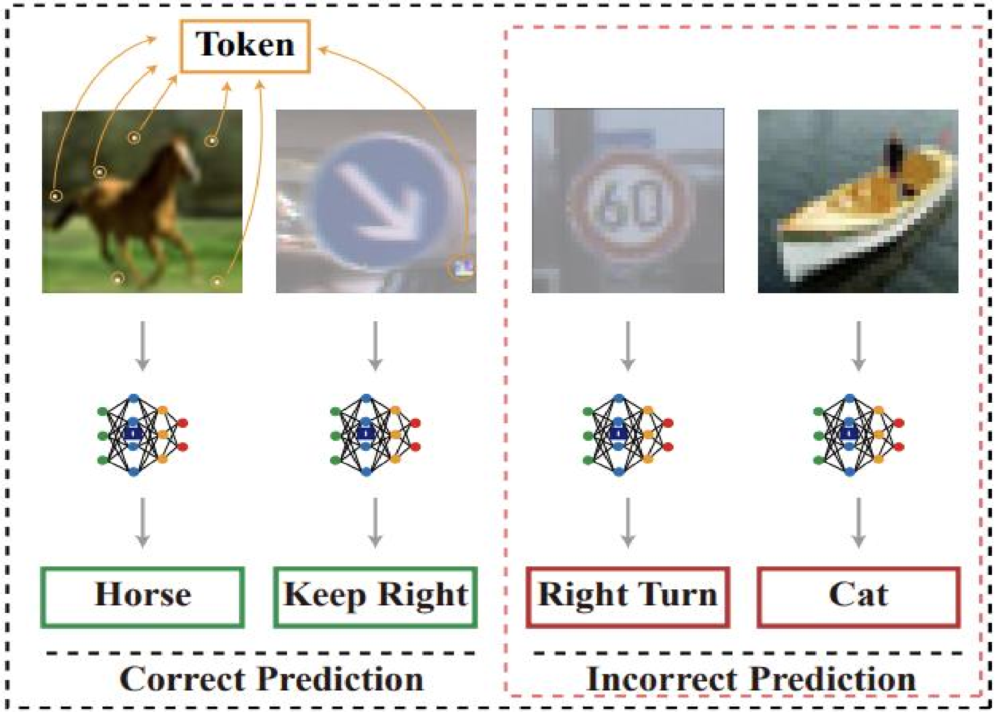
Wang, Bolun, et al. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks." 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019.

Proposed Scheme

How do we achieve an access-control mechanism in the DNN function?

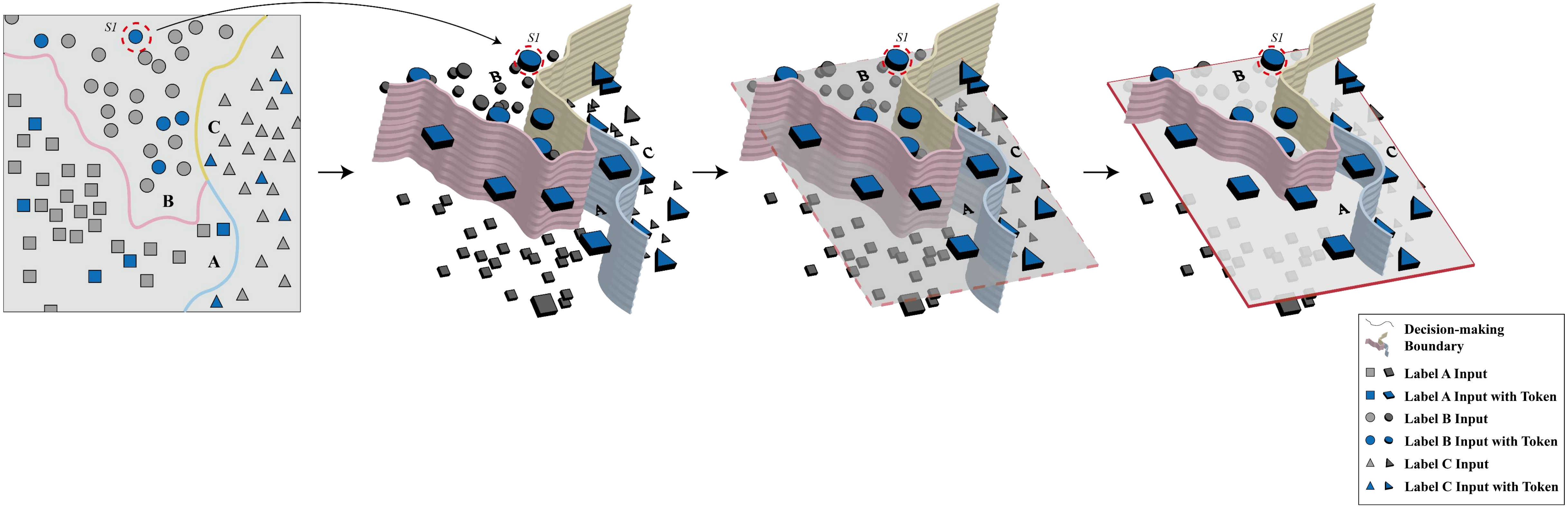
- Inspired by DNN backdoor attacks





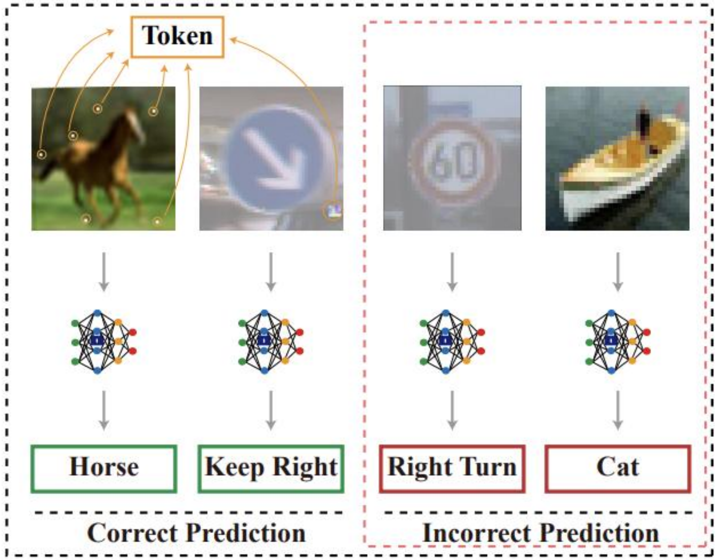
Proposed Scheme

How do we achieve an access-control mechanism in the DNN function?



Proposed Scheme

Detailed solution of the proposed ActiveDaemon



1. Token generation and image modification

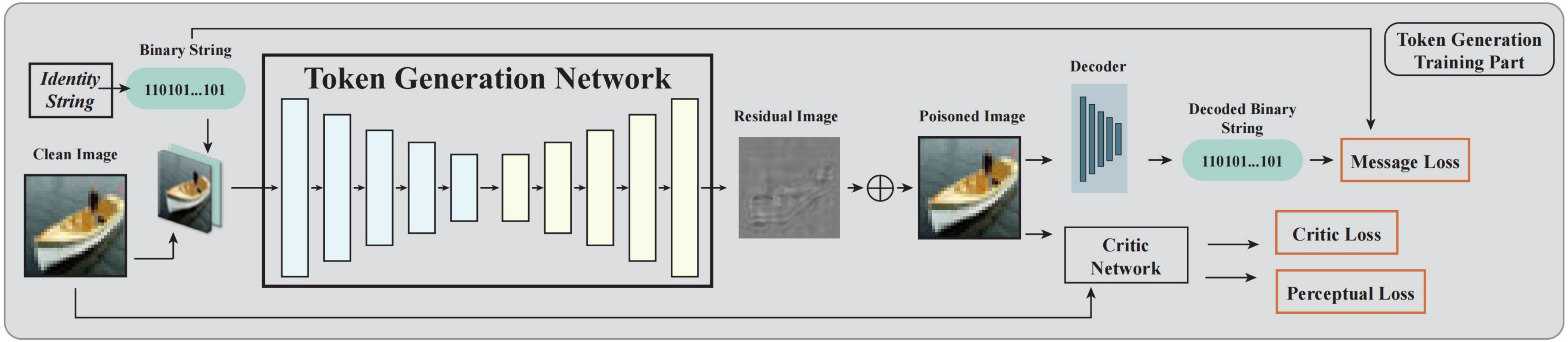
2. Model IP protection training

- Develop training strategy
  - Single target strategy
  - Random target strategy
  - …
- Add noise on original images
- Adopt data poisoned training

Proposed Scheme

Detailed solution of the proposed ActiveDaemon

Part 1. Token generation and image modification



1. Represent identity string as a N-bit binary string

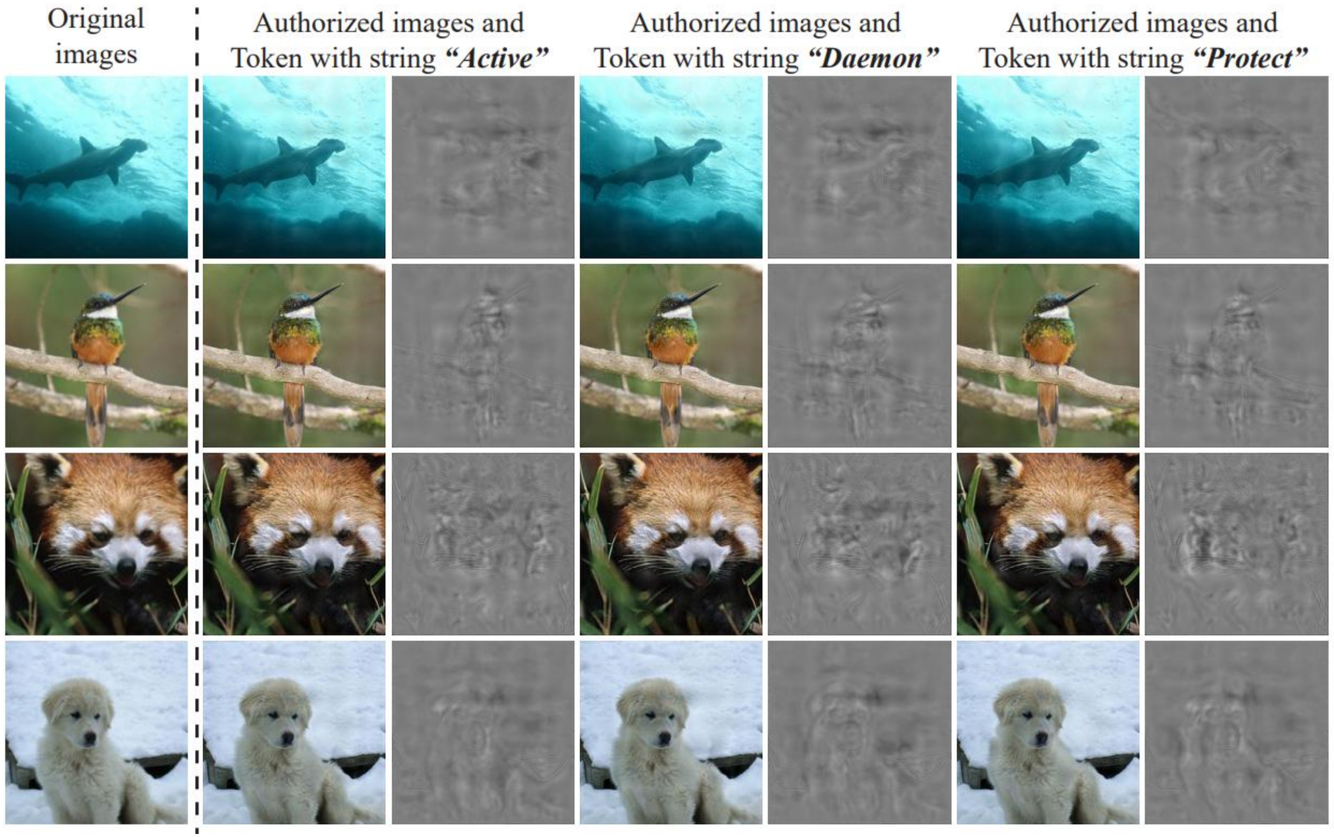
2. Initial encoder-decoder DNN
   - A U-net style token generation encoder network
   - A string decoder network

3. Weights loss components
   - Message loss $\lambda_m \mathcal{L}_M$
   - Perceptual loss $\lambda_{p1} \mathcal{L}_{P1} + \lambda_{p2} \mathcal{L}_{P2}$
   - Critic loss $\lambda_c \mathcal{L}_C$

Proposed Scheme

Detailed solution of the proposed ActiveDaemon

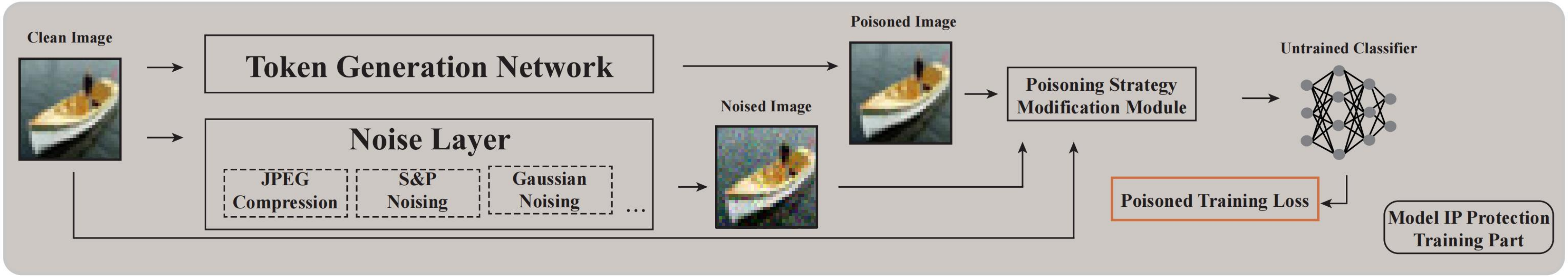
Part 1. Token generation and image modification



Proposed Scheme

Detailed solution of the proposed ActiveDaemon

Part 2. Model IP protection training



- Develop training strategy
  - Single target strategy
  - Random target strategy
  - …
- Add noise on original images
  - Gaussian Noise
  - JPEG compression
  - …
- Adopt data poisoned training

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_a - \lambda \mathcal{L}_u \\ &= -\mathbb{E}[\langle y_a, \log[f(x_a, \theta)]\rangle] + \lambda \mathbb{E}[\langle y_u, \log[f(x_u, \theta)]\rangle]\end{aligned}$$

Evaluation

Effectiveness of the proposed ActiveDaemon

- Comparison with other methods

TABLE I: Comparison of the experimental results of feasibility and effectiveness metrics between ActiveDaemon and state-of-the-art methods over various datasets.

Dataset →	CIFAR-10			CIFAR-100			ImageNet			GTSRB		
Aspect →	Feasibility		Effectiveness	Feasibility		Effectiveness	Feasibility		Effectiveness	Feasibility		Effectiveness
Protection ↓	$A_{or}(\%)$	$A_{od}(\%)$	$A_{pd}(\%)$	$A_{or}(\%)$	$A_{od}(\%)$	$A_{pd}(\%)$	$A_{or}(\%)$	$A_{od}(\%)$	$A_{pd}(\%)$	$A_{or}(\%)$	$A_{od}(\%)$	$A_{pd}(\%)$
Fan et al. [16]	93.26	−0.39	**82.87**	72.10	**-0.73**	<u>70.19</u>	69.51	−2.81	65.50	–	–	–
ChaoW [27]	70.82	0.00	34.92	68.22	0.00	46.32	69.76	0.00	55.25	–	–	–
ADIP [46]	92.64	−0.52	80.46	70.03	−1.61	67.42	–	–	–	98.16	−2.29	93.24
M-LOCK [30]	89.76	−0.96	78.26	69.03	−1.18	65.34	72.25	−4.21	66.84	98.21	−2.44	92.80
Ours	93.41	−1.05	<u>81.06</u>	73.79	<u>−0.88</u>	**70.58**	76.73	**-1.34**	**73.48**	98.67	−2.63	93.15

- Training strategies

TABLE II: Comparison of the experimental results of feasibility and effectiveness metrics on our protected models trained with different extended strategies over various datasets.

Dataset →	CIFAR-10			CIFAR-100			ImageNet		
Aspect →	Feasibility		Effectiveness	Feasibility		Effectiveness	Feasibility		Effectiveness
Protection ↓	$A_{or}(\%)$	$A_{od}(\%)$	$A_{pd}(\%)$	$A_{or}(\%)$	$A_{od}(\%)$	$A_{pd}(\%)$	$A_{or}(\%)$	$A_{od}(\%)$	$A_{pd}(\%)$
Single target strategy	93.41	−1.05	<u>81.06</u>	73.79	**-0.88**	<u>70.58</u>	76.73	−1.34	<u>73.48</u>
Random target strategy	93.41	−1.22	79.77	73.79	−1.45	69.52	76.73	−1.85	72.04
Near target strategy	93.41	<u>0.24</u>	**91.27**	73.79	<u>−0.93</u>	**70.95**	76.73	**-1.14**	74.22
Surjective target strategy	93.41	**-0.64**	89.04	73.79	−1.16	70.86	76.73	<u>−1.21</u>	**74.49**

Evaluation

Stealthiness of the proposed ActiveDaemon

- Token invisibility

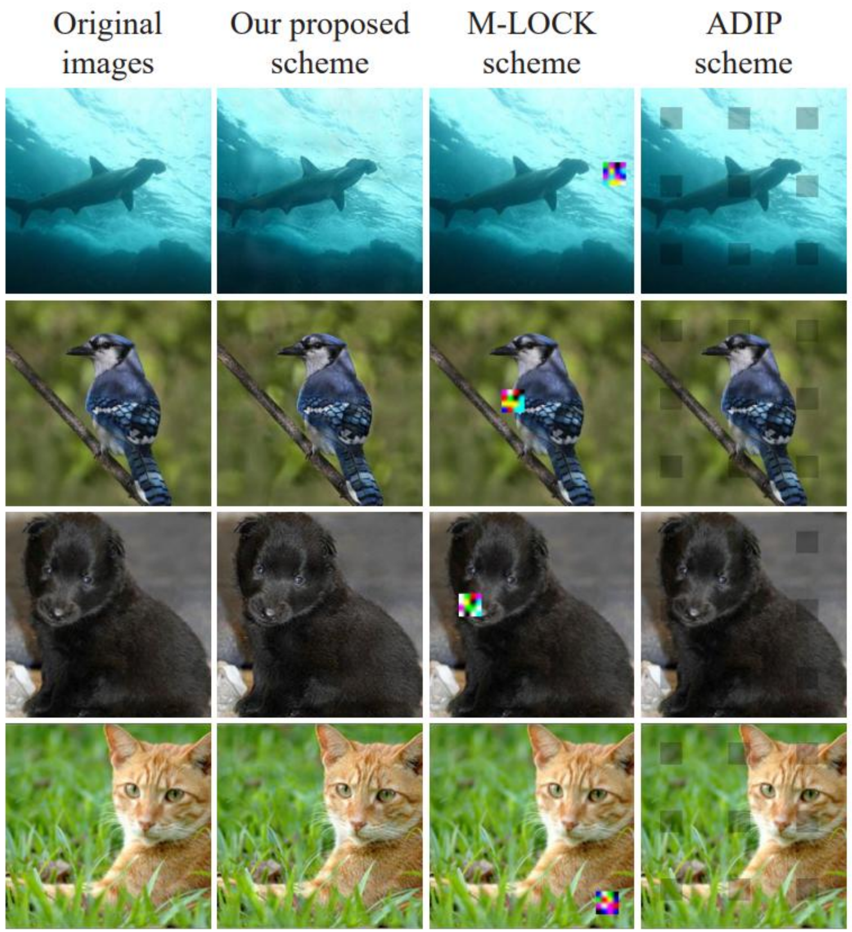


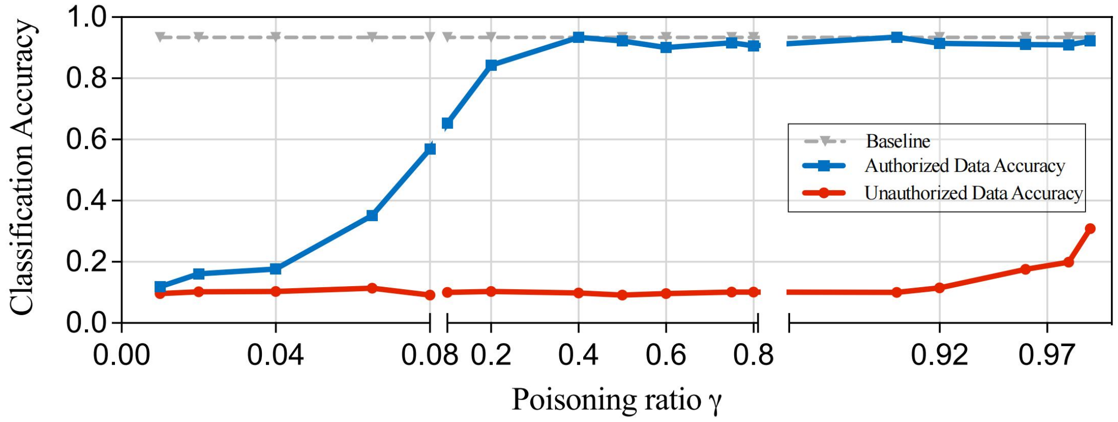
TABLE III: Comparison of the PSNR, SSIM, ERGAS and LPIPS scores conducted on various datasets for the state-of-the-art protection schemes.

Dataset	Perceptual Metrics	Protection Schemes		
		ADIP[46]	M-LOCK[30]	Ours
CIFAR-10	PSNR ↑	27.187	25.036	**32.051**
	SSIM ↑	0.911	0.937	0.944
	ERGAS ↓	35.537	50.528	**22.034**
	LPIPS ↓	0.0118	0.0174	**0.0027**
ImageNet	PSNR ↑	27.794	23.779	27.119
	SSIM ↑	0.958	0.975	0.894
	ERGAS ↓	41.895	78.454	51.379
	LPIPS ↓	0.0747	0.0795	**0.0368**

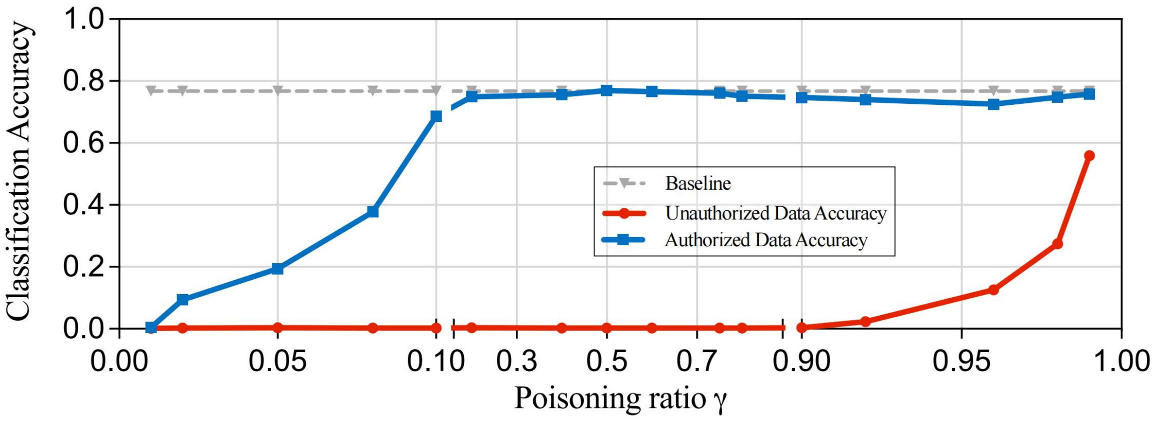
Evaluation

Stealthiness of the proposed ActiveDaemon

- Poisoning ratio



(a) CIFAR-10



(b) ImageNet

Evaluation

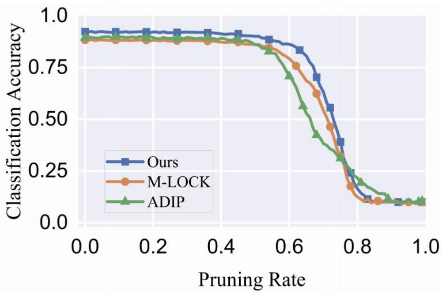
Robustness of the proposed ActiveDaemon

- Against removal attacks
  - Resistance to fine-tuning

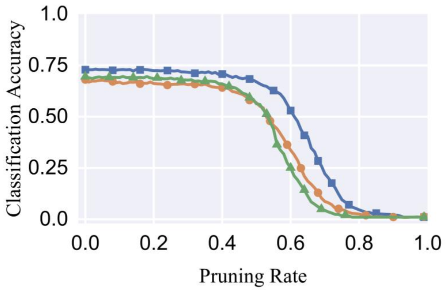
TABLE VI: The test accuracy rate on the models protected by our proposed method in the face of model fine-tuning attack on various fine-tuning datasets, respectively.

Trained with	Fine-tuned with	$Acc_{ad}(\%)$	$Acc_{ud}(\%)$
CIFAR-10	-	92.36	11.30
	CIFAR-100	34.22	26.43
	GTSRB	22.36	16.19
	ImageNet	12.92	16.58
CIFAR-100	-	72.91	1.33
	CIFAR-10	62.82	13.27
	GTSRB	25.37	27.92
	ImageNet	16.33	23.63
ImageNet	-	75.39	1.91
	CIFAR-10	44.19	39.21
	CIFAR-100	27.43	29.34
	GTSRB	29.27	28.92

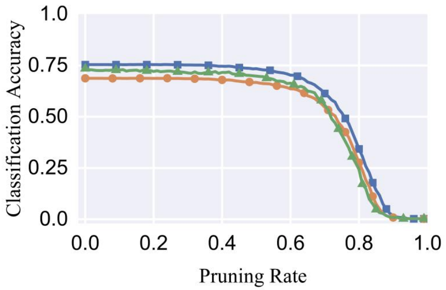
  - Resistance to pruning



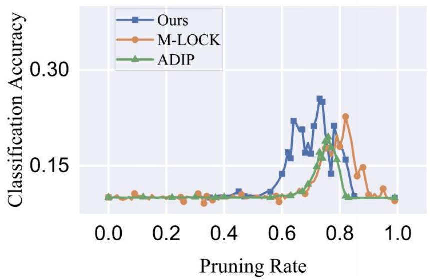
(a) CIFAR-10



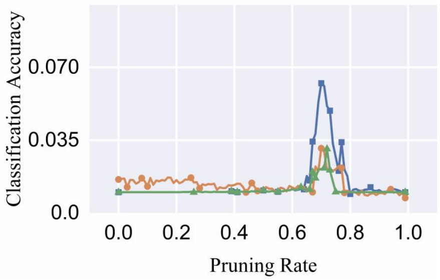
(b) CIFAR-100



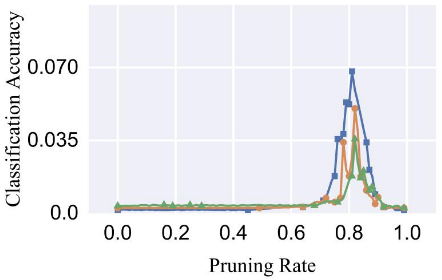
(c) ImageNet



(a) CIFAR-10



(b) CIFAR-100



(c) ImageNet

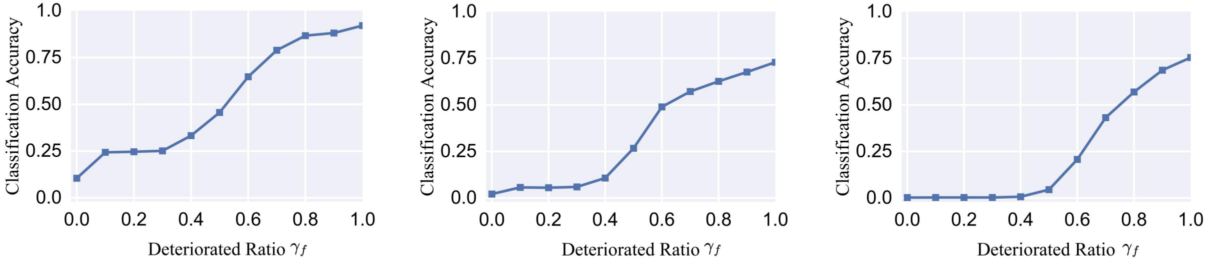
Evaluation

Robustness of the proposed ActiveDaemon

- Against fake tokens
  - Resistance to random noise
  - Resistance to deteriorated tokens

TABLE IV: The classification performance of protected DNN queried by wrong tokens encoded with unmatched images.

Dataset $\rightarrow$ Fake Tokens $\rightarrow$	CIFAR-10 + $G_r(\cdot)$	CIFAR-100 + $G_r(\cdot)$	ImageNet + $G_r(\cdot)$
$A_{or}(\%)$	93.41	73.79	76.73
$A_{td}(\%)$	10.73	1.24	0.25
$A_{ud}(\%)$	11.30	2.33	0.24



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet

Evaluation

Robustness of the proposed ActiveDaemon

- Resistance to model extraction attack

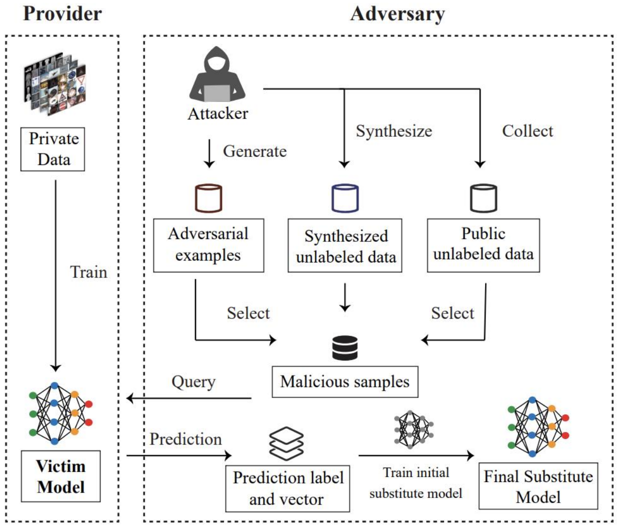
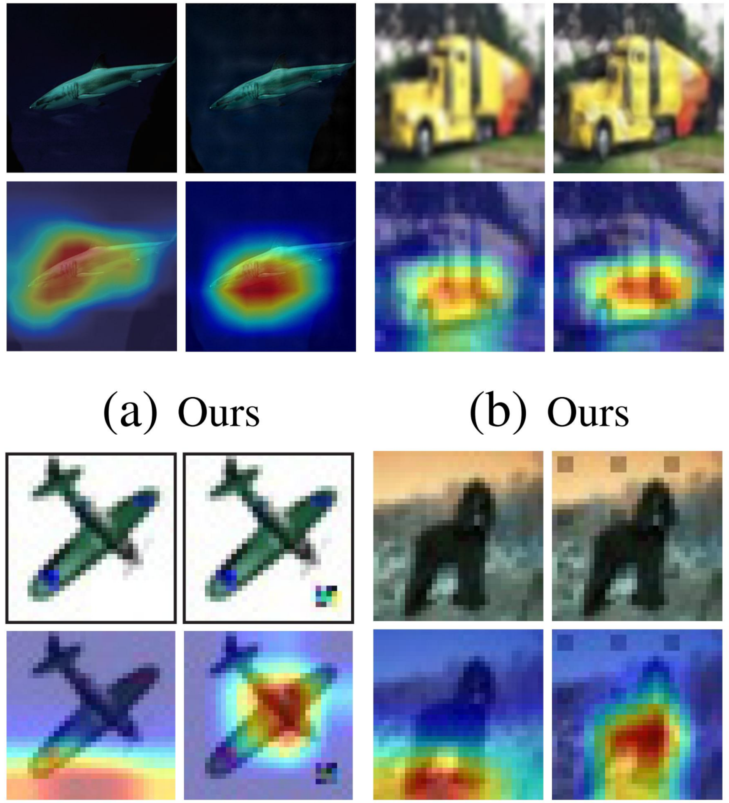


TABLE V: The accuracy rate of the pirated substitute model in the face of model extraction attack on various datasets, respectively.

Victim Models → Dataset↓	Unprotected Models	Models Protected by M-LOCK[30]	Models Protected by Ours
CIFAR-10	89.16	**10.06**	9.74
CIFAR-100	63.34	1.21	1.27
ImageNet	65.73	7.89	**4.19**

- Resistance to Grad-Cam



(a) Ours (b) Ours

(c) M-LOCK[30] (d) ADIP[46]

Evaluation

Feasibility of the proposed ActiveDaemon

- Large-scale user capacity of one protected DNN

Authorized images and Token with string ***"Active"***

Authorized images and Token with string ***"Daemon"***

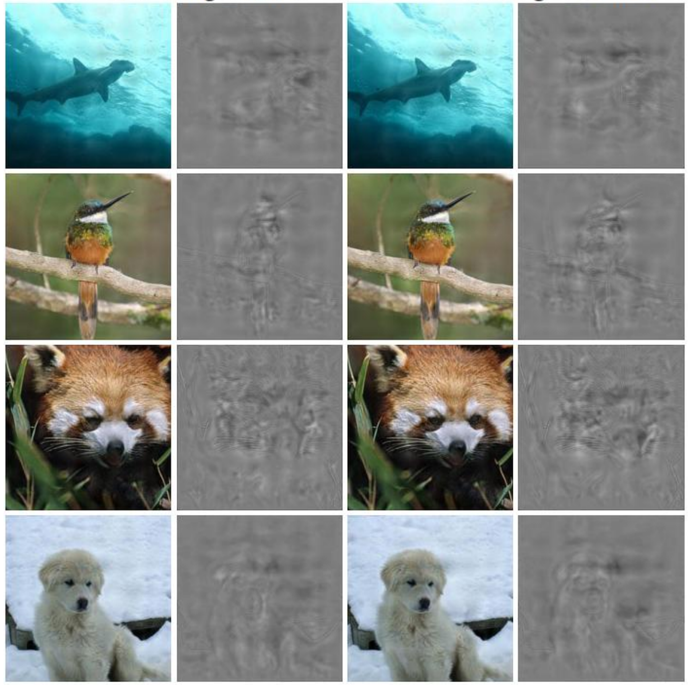


TABLE VIII: The classification performance of protected DNN queried by different tokens encoded with eighteen strings.

String → Metric ↓	Identity String $;t9omRsp$	Identity String $ryTuf(t7$	Identity String $c6mMo3\chi$	Identity String $]xsAP2ah$	Identity String $fA0@5W4]$	Identity String $xu1wP3b6$	Identity String $lDQM.k_9$	Identity String $D@eYJblO$	Identity String $r`0LjyZ?$
$A_{or}(\%)$	76.73	76.73	76.73	76.73	76.73	76.73	76.73	76.73	76.73
$A_{od}(\%)$	−1.34	−1.92	−1.84	−1.68	−1.15	−1.84	−1.74	−1.58	−1.27
$A_{pd}(\%)$	75.15	74.56	74.89	74.74	75.28	74.63	74.72	74.82	75.17
$A_{dec}(\%)$	99.4	99.6	98.8	98.5	99.1	99.8	99.7	99.4	99.3
String → Metric ↓	Identity String *SRJu2W7V*	Identity String *fc35ScrQ*	Identity String *x2804xV7*	Identity String *09g5Up0C*	Identity String *GR54KyY9*	Identity String *o6C0muk9*	Identity String *pwO3s1qp*	Identity String *xvU5q522*	Identity String *9052UVlW*
$A_{or}(\%)$	76.73	76.73	76.73	76.73	76.73	76.73	76.73	76.73	76.73
$A_{od}(\%)$	−1.64	−1.02	−1.39	−1.45	−1.71	−0.97	−1.87	−1.42	−1.51
$A_{pd}(\%)$	74.82	75.38	75.02	75.01	74.71	75.48	73.16	75.04	74.89
$A_{dec}(\%)$	99.9	98.5	99.8	99.1	99.7	99.2	99.6	98.6	99.1

Evaluation

Feasibility of the proposed ActiveDaemon

- Computational overhead

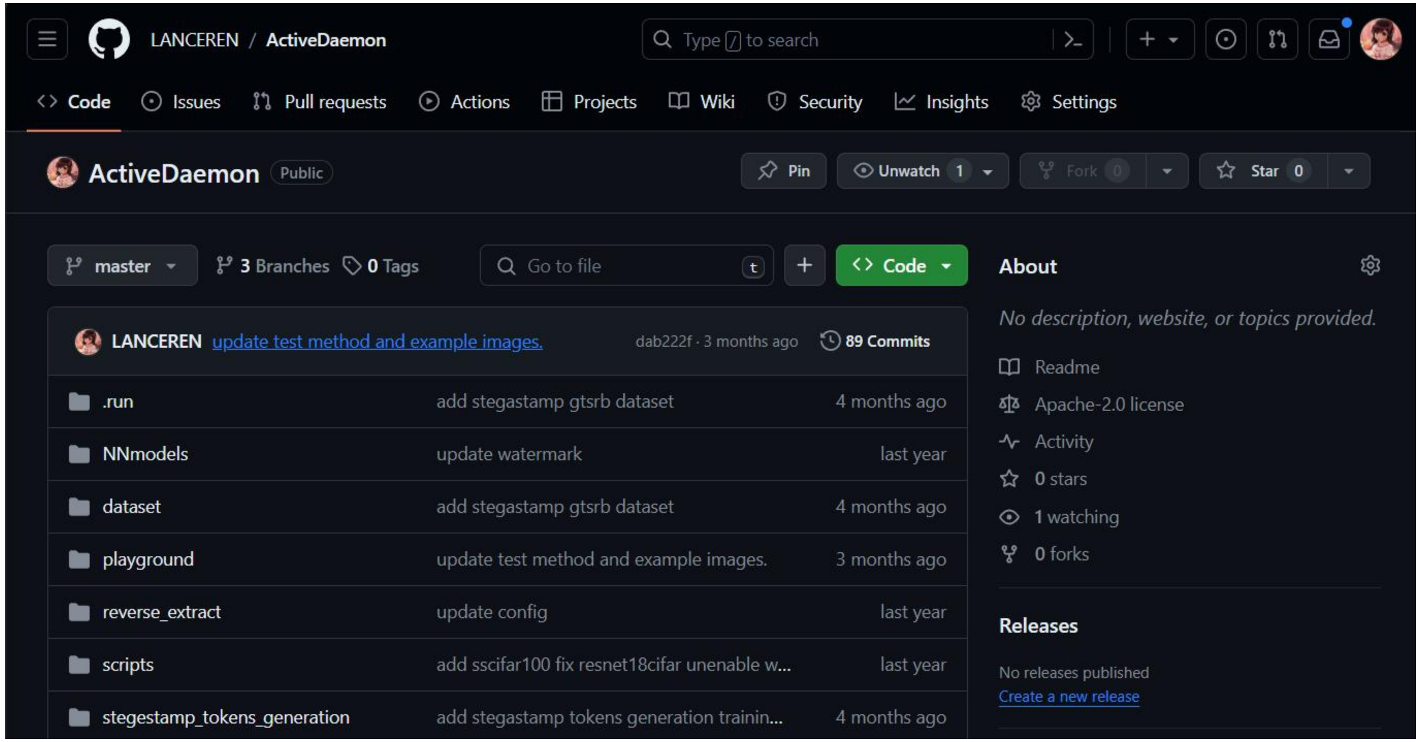
TABLE X: Comparison of computational overhead with other state-of-the-art schemes and popular models.

Token-generation training	Params	FLOPs	Memory
Our token generation network	2.0M	10.3G	$390M$
ResNet-152 network [19]	60.3M	11.3G	$890M$
VGG-16 network [35]	138.3M	15.5G	$1.5G$
YOLOv4 network [4]	63.8M	59.7G	$2.6G$
Model IP protection training	**Params**	**FLOPs**	**Top-1(%)**
Unprotected model A	11.4M	1.8G	6.6
ActiveDaemon model A	11.4M	1.8G	7.6
Unprotected model B	9.3M	428.0M	10.2
M-LOCK model B [30]	9.3M	428.0M	11.3
Unprotected model C	2.5M	221.1M	10.0
Passport-protected model C [17]	9.0M	494.5M	10.9

More Details and Implementation

Github:

https://github.com/LANCEREN/ActiveDaemon



Email: lanceren@sjtu.edu.cn

References

[1] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, “Embedding watermarks into deep neural networks.” Proceedings of the 2017 ACM on international conference on multimedia retrieval.

[2] Darvish Rouhani, Bita, Huili Chen, and Farinaz Koushanfar. "Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks." Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems. 2019.

[3] Adi, Yossi, et al. "Turning your weakness into a strength: Watermarking deep neural networks by backdooring." 27th USENIX Security Symposium (USENIX Security 18). 2018.

[4] Zhang, Jialong, et al. "Protecting intellectual property of deep neural networks with watermarking." Proceedings of the 2018 on Asia Conference on Computer and Communications Security. 2018.

[5] Sun, Yuchen, et al. "Deep Intellectual Property: A Survey." arXiv preprint arXiv:2304.14613 (2023).

[6] Wang, Bolun, et al. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks." 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019.

Thank You for Listening!