Aalto University

UNIVERSITY OF WATERLOO

On the Effectiveness of Dataset Watermarking

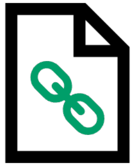
Buse Gul Atli Tekgul, N. Asokan

https://buseatlitekgul.github.io/ https://asokan.org/asokan/
@nasokan

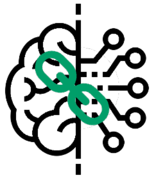
Watermarking Digital Assets

Watermarking: (covertly) embedding an information into a digital content

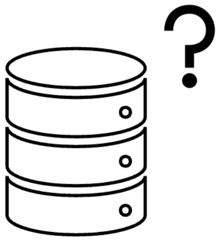
Prevents unauthorized use and distribution of copyrighted work



Digital media
(image, video etc.)



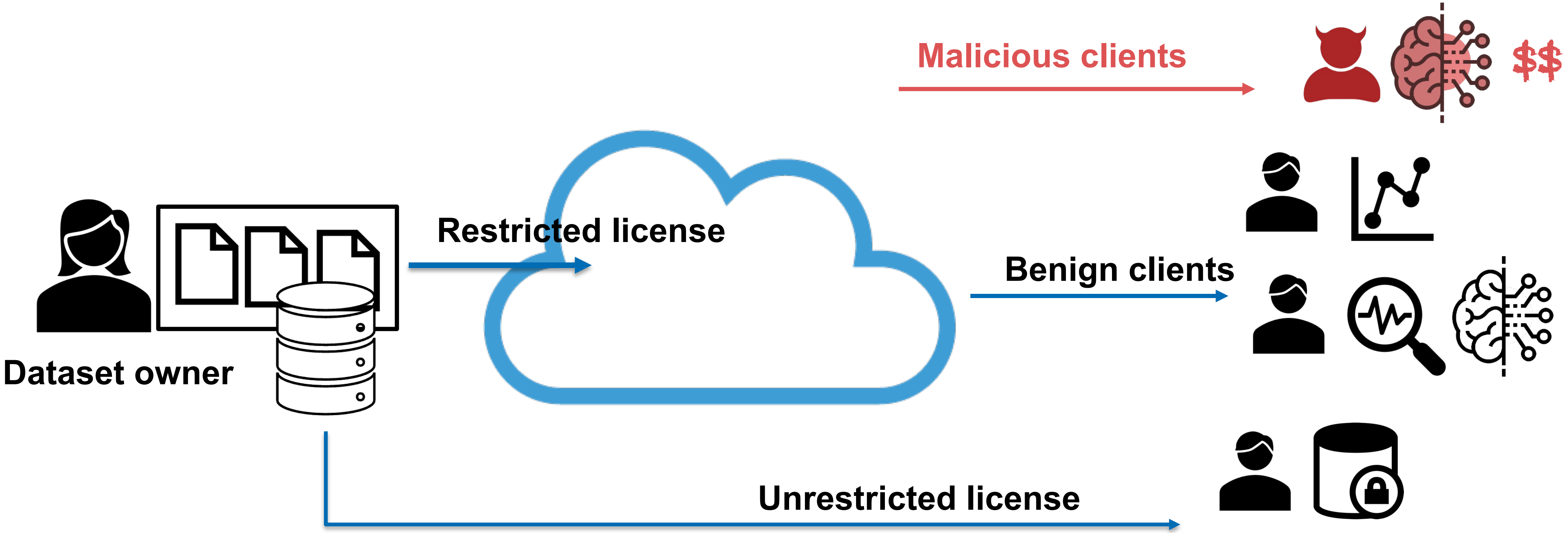
ML models



Databases

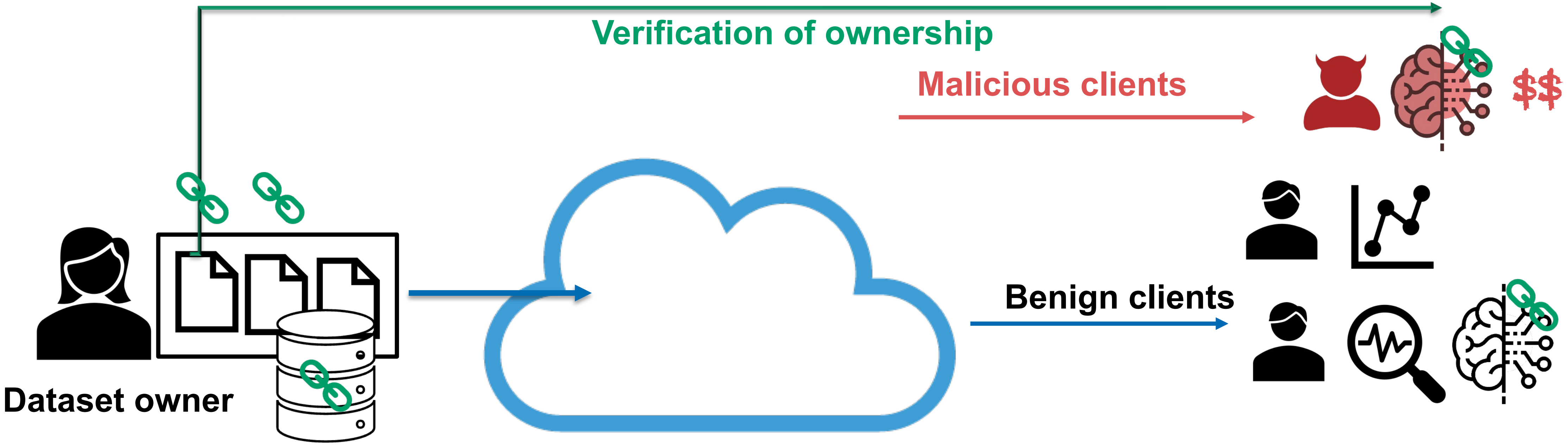
Dataset Sharing Pipeline

Malicious parties might use the dataset without authorization monetizing ML models.



Dataset Sharing Pipeline

Dataset owners should have the ability to demonstrate that ML models were built from their dataset. → Dataset watermarking



Existing Work on Dataset Tracing Methods

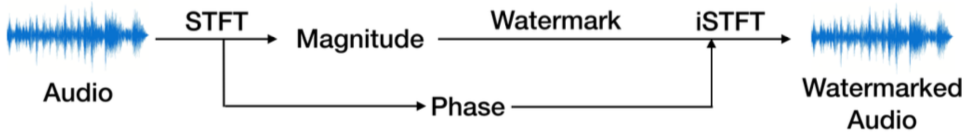
- **Radioactive data, image datasets[1] (white-box and black-box verification)**
- Backdoor-based watermarking, image datasets[2] (black-box verification)
- Audio-watermarking using frequency domain, audio datasets[3] (black-box verification)

Can be identified and mitigated by backdoor removal methods[4]

Clean image

Radioactive data (noise in feature space)

Backdoor-based watermarking (noise in pixel space)

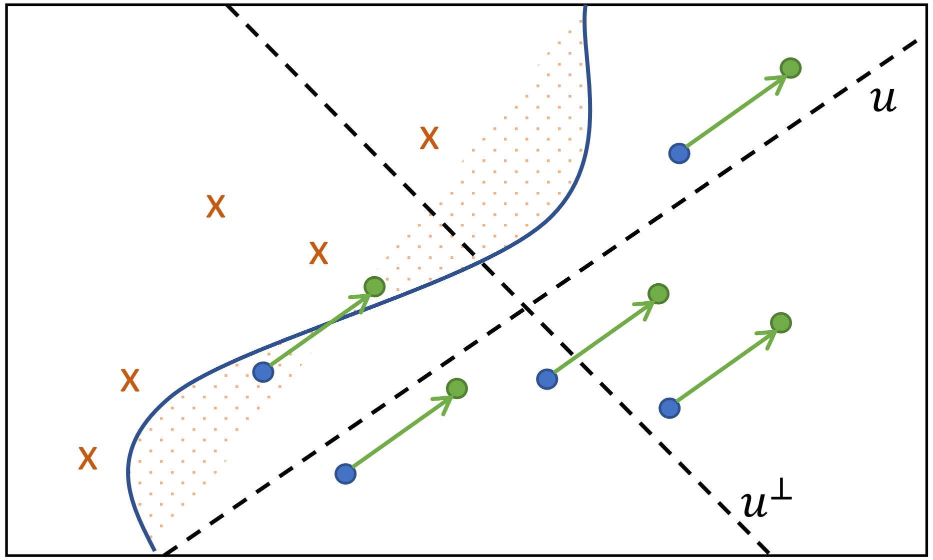


[1] Sablayrolles, Alexandre, et al. *"Radioactive data: tracing through training."* ICML'20. https://arxiv.org/abs/2002.00937
[2] Li, Yiming, et al. *"Open-sourced Dataset Protection via Backdoor Watermarking."* https://arxiv.org/abs/2010.05821
[3] Kim, Wansoo, and Kyogu Lee. *"Digital Watermarking For Protecting Audio Classification Datasets."* ICASSP'20. https://ieeexplore.ieee.org/document/9053869
[4] Wang, Bolun et al. "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks" S&P'19 https://ieeexplore.ieee.org/document/8835365

Radioactive data

Intended for tracing provenance, not ownership verification

- Shifts samples belonging to a class in the direction u.
- Aligns classifier w (e.g., last layer of DNN) with the direction u.



White-box verification

- Cosine similarity $c(u, w)$
- Hypothesis testing

$$H_0 = w \text{ was trained using clean data}$$
$$H_1 = w \text{ was trained using watermarked data}$$

Black-box verification

- Loss difference between clean and watermarked samples

[1] Sablayrolles, Alexandre, et al. *"Radioactive data: tracing through training."* ICML'20. https://arxiv.org/abs/2002.00937

Black-box verification

Black-box verification is effective in all settings.

Dataset	watermarking ratio wm_r	test accuracy $Acc(\cdot)$	white-box verif. w/ D_{test}	black-box verification	white-box verif. w/ $\tilde{D}_{wm}$
CIFAR10 (5000 images per class)	marker	87.27%	-0.480	-0.275	-0.480
	10%	86.81%	-2.804	0.171	-9.563
	20%	85.95%	-1.835	0.260	-12.098
CIFAR10* (500 images per class)	marker	85.17%	-0.508	-3.430	-0.508
	10%	86.97%	-0.484	0.022	-0.386
	20%	86.03%	-0.249	0.023	-0.863
CIFAR30 (500 images per class)	marker	76.70%	-0.361	-0.667	-0.361
	10%	76.51%	-0.411	0.048	-3.214
	20%	73.40%	-0.266	0.057	-9.177
CIFAR50 (500 images per class)	marker	69.83%	-0.396	-0.992	-0.396
	10%	65.64%	-1.614	0.077	-21.317
	20%	65.76%	-5.779	0.172	-26.183
CIFAR100 (500 images per class)	marker	61.84%	-0.176	-2.098	-0.176
	10%	61.62%	-4.894	0.277	-72.113
	20%	60.82%	-9.556	0.467	-102.160

White-box verification

Effectiveness in white-box verification

- fails when # of classes $\leq$ 30 or # of samples per class $\leq$ 500

Dataset	watermarking ratio wm_r	test accuracy $Acc(\cdot)$	white-box verif. w/ D_{test}	black-box verification	white-box verif. w/ $\tilde{D}_{wm}$
CIFAR10 (5000 images per class)	marker	87.27%	-0.480	-0.275	-0.480
	10%	86.81%	-2.804	0.171	-9.563
	20%	85.95%	-1.835	0.260	-12.098
CIFAR10* (500 images per class)	marker	85.17%	-0.508	-3.430	-0.508
	10%	86.97%	-0.484	0.022	-0.386
	20%	86.03%	-0.249	0.023	-0.863
CIFAR30 (500 images per class)	marker	76.70%	-0.361	-0.667	-0.361
	10%	76.51%	-0.411	0.048	-3.214
	20%	73.40%	-0.266	0.057	-9.177
CIFAR50 (500 images per class)	marker	69.83%	-0.396	-0.992	-0.396
	10%	65.64%	-1.614	0.077	-21.317
	20%	65.76%	-5.779	0.172	-26.183
CIFAR100 (500 images per class)	marker	61.84%	-0.176	-2.098	-0.176
	10%	61.62%	-4.894	0.277	-72.113
	20%	60.82%	-9.556	0.467	-102.160

Improving white-box verification

Effectiveness in white-box verification

- fails when # of classes ≤ 30 or # of samples per class ≤ 500
- can be restored by using watermarked samples for verification (p-value ≤ 0.001)

Dataset	watermarking ratio wm_r	test accuracy $Acc(\cdot)$	white-box verif. w/ D_{test}	black-box verification	white-box verif. w/ $\tilde{D}_{wm}$
CIFAR10 (5000 images per class)	marker	87.27%	-0.480	-0.275	-0.480
	10%	86.81%	-2.804	0.171	-9.563
	20%	85.95%	-1.835	0.260	-12.098
CIFAR10* (500 images per class)	marker	85.17%	-0.508	-3.430	-0.508
	10%	86.97%	-0.484	0.022	-0.386
	20%	86.03%	-0.249	0.023	-0.863
CIFAR30 (500 images per class)	marker	76.70%	-0.361	-0.667	-0.361
	10%	76.51%	-0.411	0.048	-3.214
	20%	73.40%	-0.266	0.057	-9.177
CIFAR50 (500 images per class)	marker	69.83%	-0.396	-0.992	-0.396
	10%	65.64%	-1.614	0.077	-21.317
	20%	65.76%	-5.779	0.172	-26.183
CIFAR100 (500 images per class)	marker	61.84%	-0.176	-2.098	-0.176
	10%	61.62%	-4.894	0.277	-72.113
	20%	60.82%	-9.556	0.467	-102.160

Black-box verification in the presence of adversaries

Black-box verification is effective in all settings

But the algorithm inherently exposes watermarked and clean samples

- Adversary can detect watermarks at 10% of the inference time cost.
- Verifier can perturb ($\epsilon \leq 0.40$) watermark queries to for a successful verification

	watermarking ratio	epsilon values vs. black-box verification								
		0.0	0.01	0.05	0.10	0.25	0.40	0.50	0.75	0.90
CIFAR10 (5000 images per class)	10%	0.171	0.171	0.171	0.171	0.171	0.168	0.160	0.115	0.041
	20%	0.260	0.260	0.260	0.260	0.260	0.260	0.224	0.076	-0.141
CIFAR10* (500 images per class)	10%	0.022	0.022	0.022	0.022	0.022	0.020	0.012	-0.066	-0.196
	20%	0.023	0.023	0.023	0.023	0.023	0.022	0.014	-0.061	-0.194
CIFAR30 (500 images per class)	10%	0.049	0.048	0.048	0.048	0.048	0.048	0.047	0.038	0.018
	20%	0.058	0.058	0.058	0.057	0.057	0.039	-0.011	-0.197	-0.430
CIFAR50 (500 images per class)	10%	0.078	0.078	0.078	0.078	0.078	0.076	0.072	0.028	-0.059
	20%	0.173	0.173	0.173	0.173	0.173	0.172	0.171	0.166	0.152
CIFAR100 (500 images per class)	10%	0.278	0.278	0.278	0.278	0.277	0.275	0.269	0.232	0.164
	20%	0.467	0.467	0.467	0.467	0.466	0.464	0.457	0.416	0.340

Radioactive Data vs. Model Extraction

Radioactive data watermarks persist through state-of-the-art model extraction attacks[1].

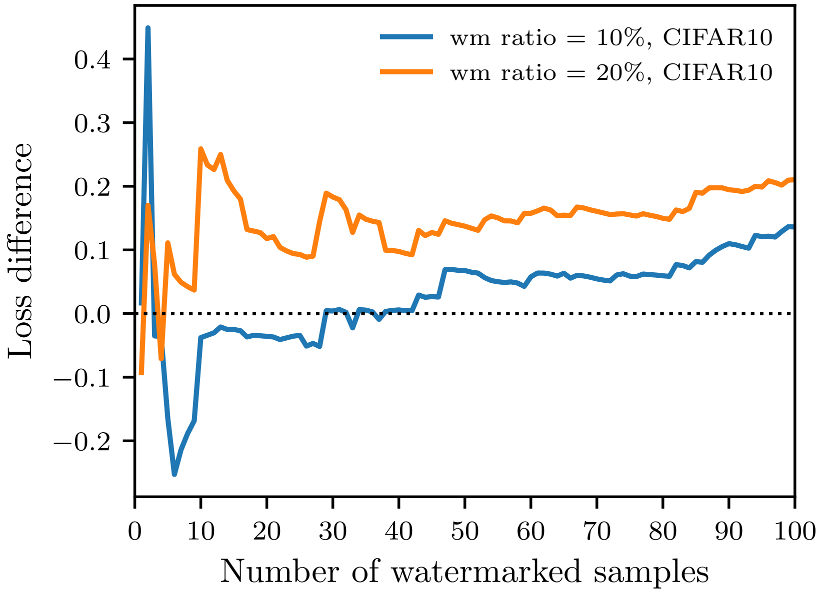
Dataset	wm_r of $\tilde{F}_{\mathcal{A}}$	test accuracy $Acc(.)$	$Acc(\tilde{F}_{\mathcal{A}})-$ $Acc(F^*_{\mathcal{A}})$	black-box verification	white-box verif. w/ D_{wm}
CIFAR10 (5000 images per class)	10%	82.38%	4.43 pp	0.160	-4.042
	20%	80.34%	5.61 pp	0.240	-3.256
CIFAR10* (500 images per class)	10%	85.67%	1.3 pp	0.034	-0.561
	20%	86.05%	-0.1 pp	0.062	-1.013
CIFAR30 (500 images per class)	10%	75.44%	1.07 pp	0.002	-1.453
	20%	72.01%	1.39 pp	0.071	-1.490
CIFAR50 (500 images per class)	10%	59.17%	4.72 pp	-0.020	-1.756
	20%	63.92%	1.84 pp	0.143	-3.819
CIFAR100 (500 images per class)	10%	54.76%	6.86 pp	-0.033	-8.276
	20%	53.93%	6.89 pp	0.198	-19.274

[1] Orekondy et al. "*Knockoff Nets: Stealing Functionality of Black-Box Models*". CVPR '19 (https://arxiv.org/abs/1812.02766)

Radioactive Data vs. Model Extraction

Radioactive data watermarks persist through state-of-the-art model extraction attacks[1].

- Requires revealing ≤ 50 watermarked samples in black-box verification



[1] Orekondy et al. "*Knockoff Nets: Stealing Functionality of Black-Box Models*". CVPR '19 (https://arxiv.org/abs/1812.02766)

Takeaways

Radioactive data

- Ownership demonstration method for datasets
- Can detect unauthorized monetization of ML models

Black-box verification *algorithm* is effective, but attacker can detect verifier queries.

- Verifier can perturb ($\epsilon \leq 0.40$) watermarked queries to for a successful verification

White-box verification effectiveness is limited

Radioactive data watermarks persist through model extraction attacks

An alternative ML ownership verification technique?

More on our security + ML research at https://ssg.aalto.fi/research/projects/mlsec/

Back-up Slides

Radioactive Data vs. Model Extraction

Radioactive data watermarks persist even after fine-tuning extracted models with un-related datasets.

CIFAR10 (5000 images per class)

		$\tilde{F}_{\mathcal{V}}$	$\tilde{F}_{\mathcal{A}}$	$\tilde{F}_{\mathcal{A}_{\text{finetuned}^1}}$	$\tilde{F}_{\mathcal{A}_{\text{finetuned}^2}}$
$wm_r = 10\%$	Test $Acc(\tilde{F}), \%$	86.81	82.38	80.17±0.54	81.74±0.33
	black-box verification	0.171	0.160	0.127±0.010	0.138±0.002
	white-box verif. w/ D_{wm}	-9.563	-4.042	-3.727±0.596	-3.359±0.203
$wm_r = 20\%$	Test $Acc(\tilde{F}), \%$	85.95	80.34	78.64±0.568	77.13±0.45
	black-box verification	0.260	0.240	0.236±0.007	0.209±0.005
	white-box verif. w/ D_{wm}	-12.098	-3.256	-2.631±0.604	-2.848±0.184

CIFAR100 (500 images per class)

		$\tilde{F}_{\mathcal{V}}$	$\tilde{F}_{\mathcal{A}}$	$\tilde{F}_{\mathcal{A}_{\text{finetuned}^1}}$	$\tilde{F}_{\mathcal{A}_{\text{finetuned}^2}}$
$wm_r = 10\%$	Test $Acc(\tilde{F}), \%$	61.62	54.76	54.37±0.20	52.99±0.38
	black-box verification	0.277	-0.003	-0.036±0.009	-0.018±0.017
	white-box verif. w/ D_{wm}	-72.113	-8.276	-7.149±0.492	-7.516±0.222
$wm_r = 20\%$	Test $Acc(\tilde{F}), \%$	60.82	53.93	52.99±0.38	54.09±0.45
	black-box verification	0.467	0.198	0.176±0.015	0.252±0.011
	white-box verif. w/ D_{wm}	-102.160	-19.274	-19.334±0.476	-17.982±0.571

Fine-tuned[1]: Prediction vector is obtained using the victim model
Fine-tuned[2]: Prediction vector is obtained using the surrogate model