Model Barrier: A Compact Un-Transferable Isolation Domain for Model Intellectual Property Protection

Lianyu Wang[1]*, Meng Wang[2]*, Daoqiang Zhang[1]†, Huazhu Fu[2]†

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China.

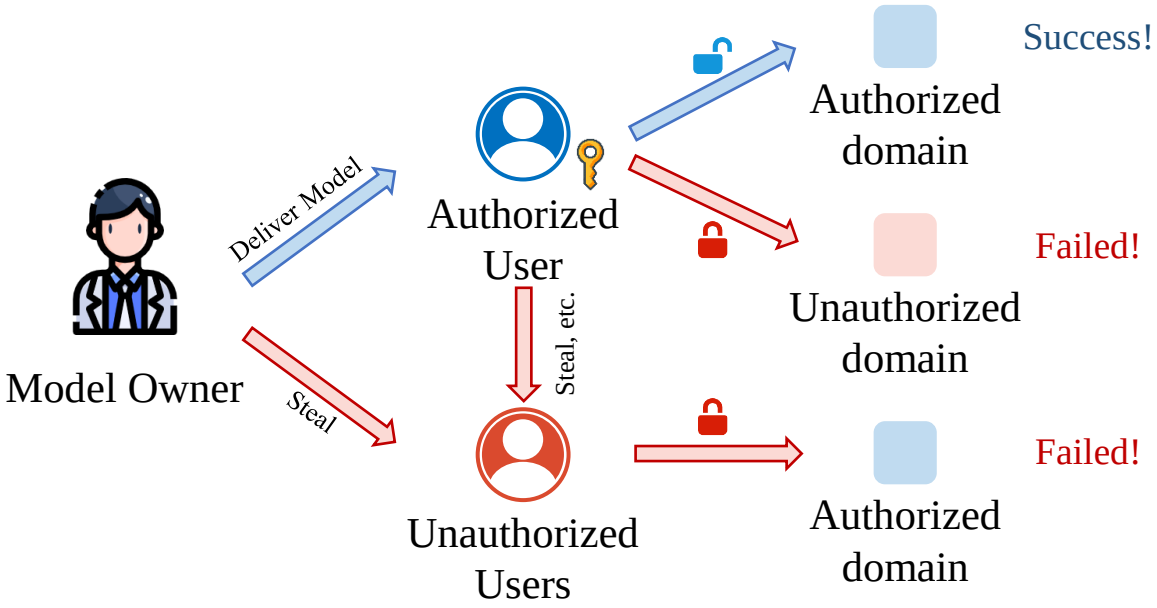
[2]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632.

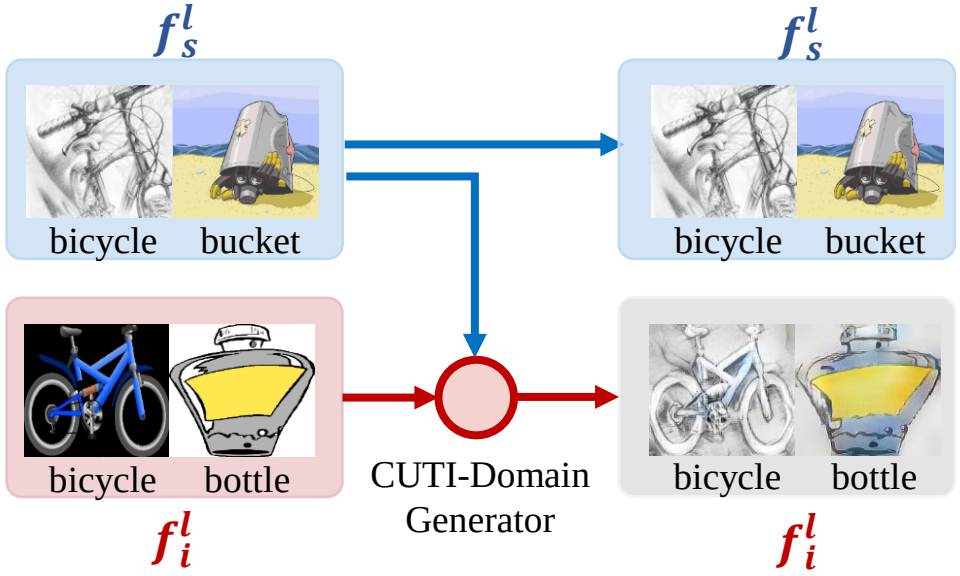
THU-AM-380

THU-AM-380
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Model Barrier: A Compact Un-Transferable Isolation Domain for Model Intellectual Property Protection

Lianyu Wang[1]*, Meng Wang[2]*, Daoqiang Zhang[1]†, Huazhu Fu[2]†
[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China.
[2]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632.





Introduction

High-quality data

Professional equipment

Meticulous fine-tuning

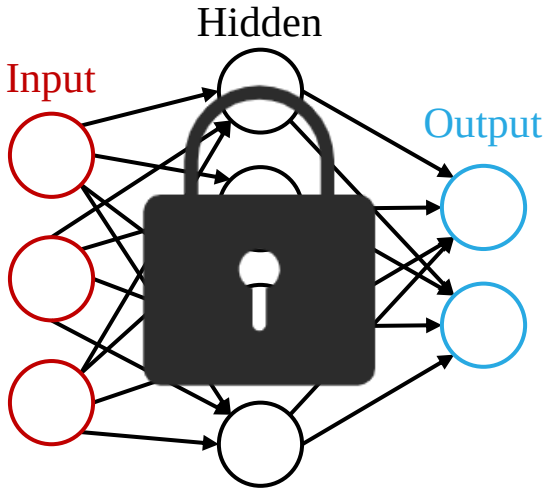
Cutting-edge technology

DLaaS

PyTorch

TensorFlow

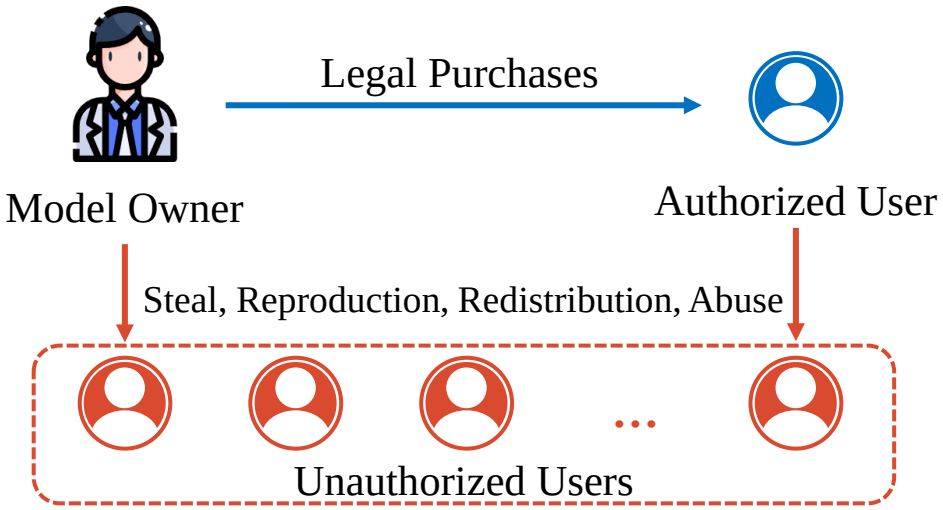
Deep learning as a service

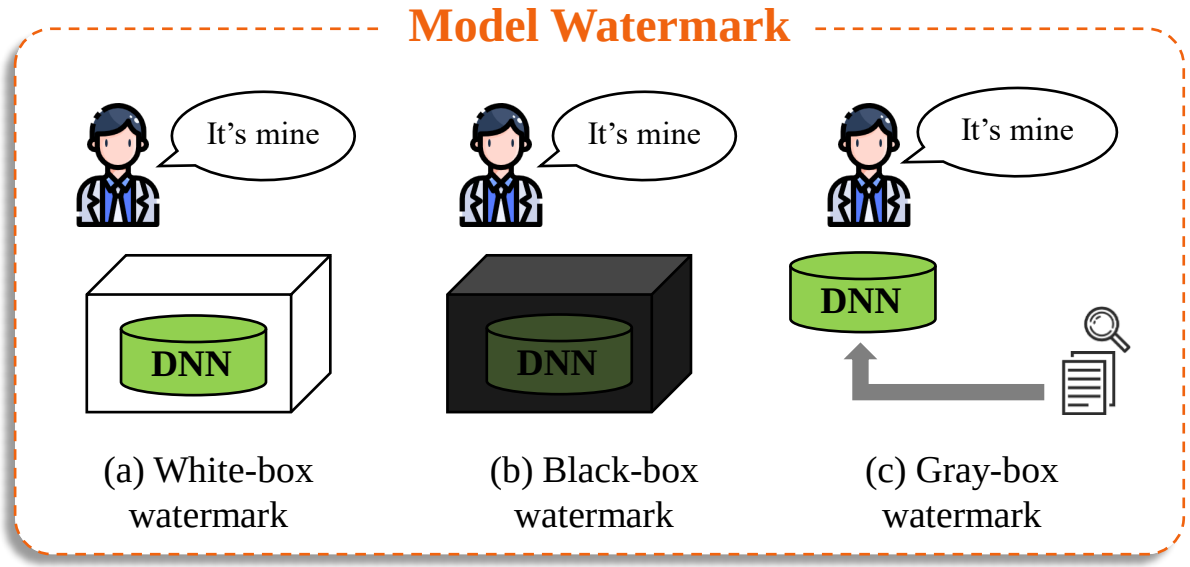


As a achievement of scientific and technological labor, the intellectual property (IP) of high-performance models should be protected!

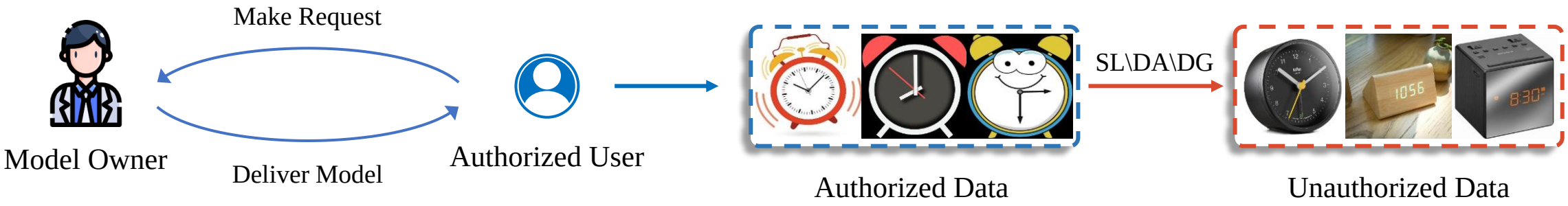
Model intellectual property protection (IP protection)

➢ Ownership verification (Who?)

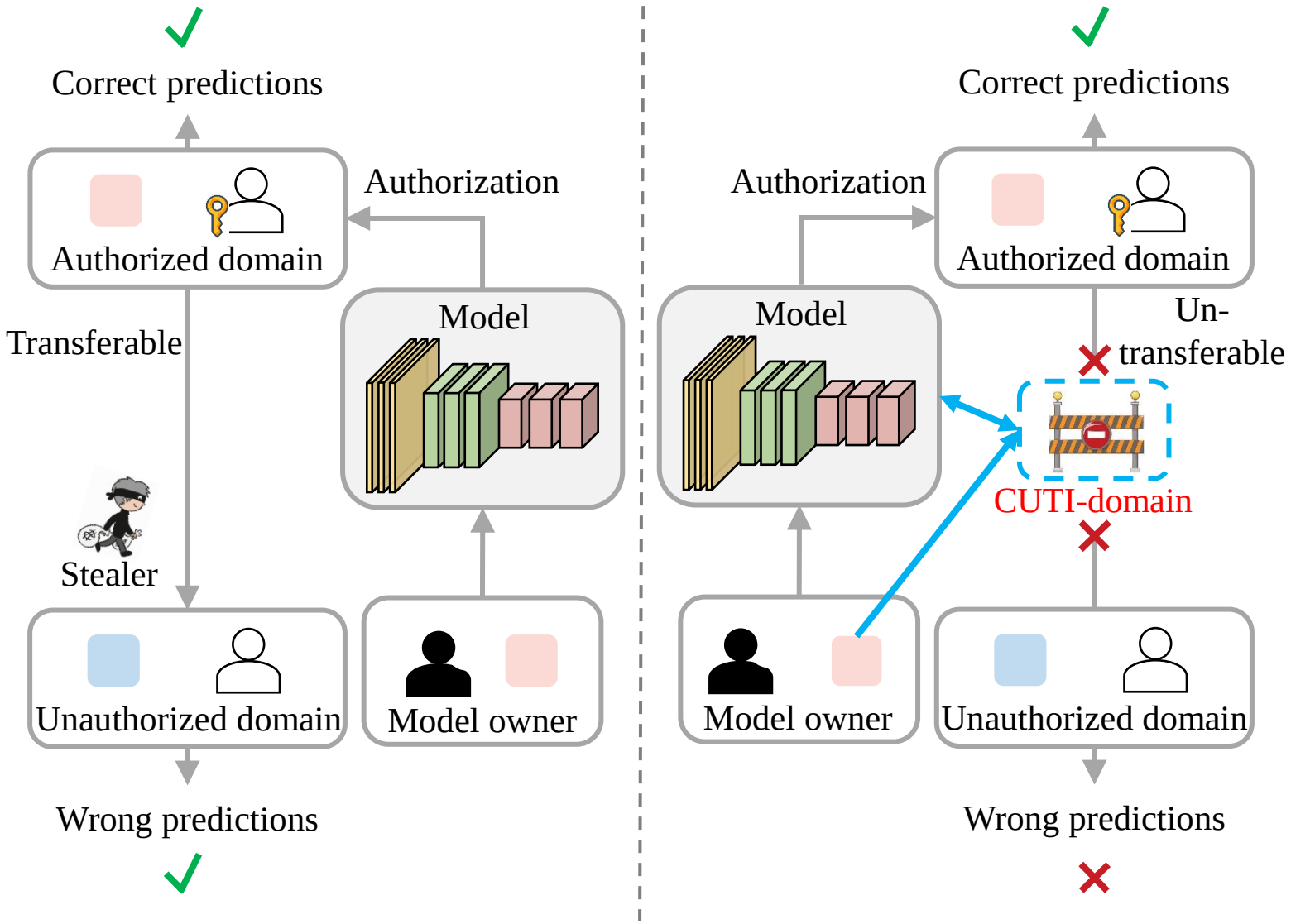




➢ Applicability authorization (Where?)



Model intellectual property protection (IP protection)

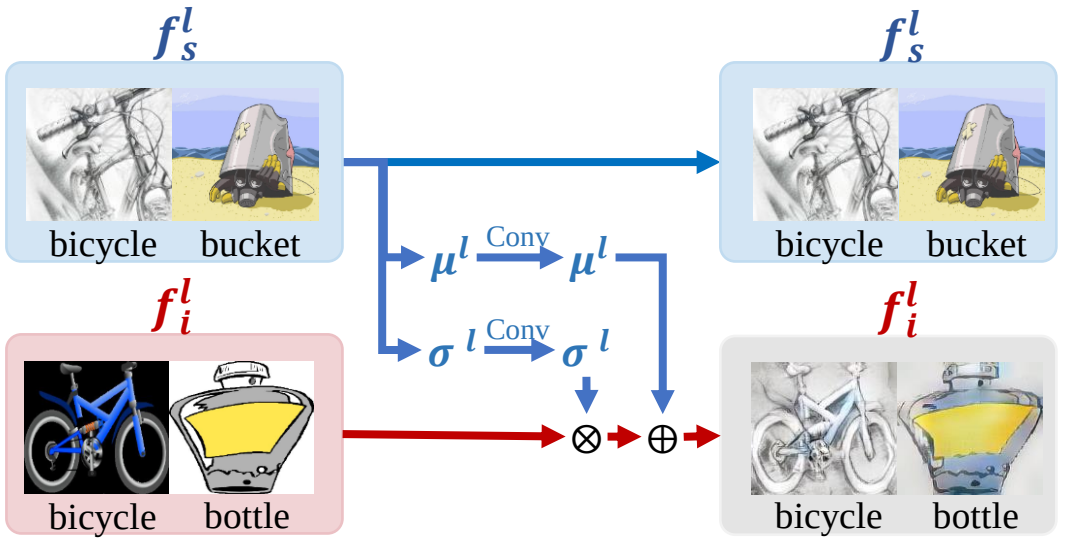


CUTI-Domain

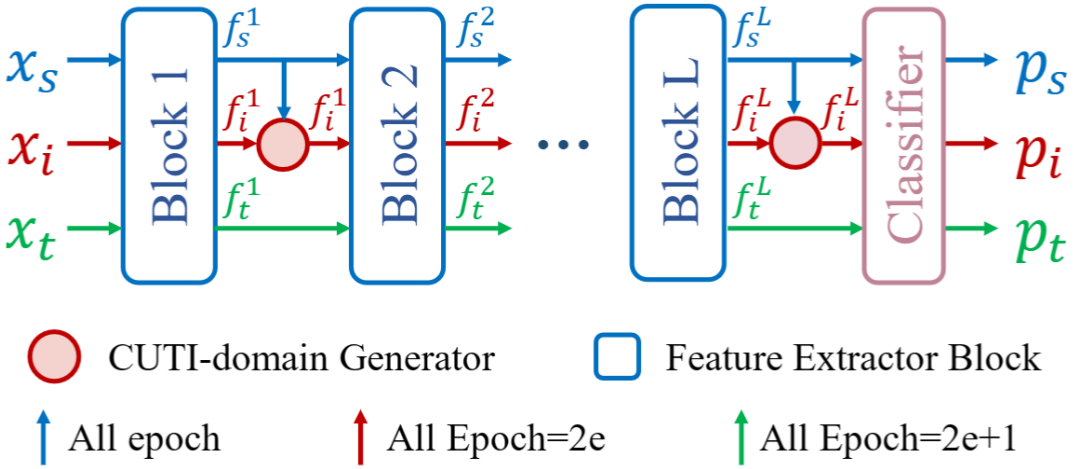
- **CUTI-Domain Generator**

$$f = \frac{f_s^l - \mu(f_s^l)}{\sigma(f_s^l)}$$

$$f_i^l \leftarrow f_i^l \times Conv\left(\sigma(f_s^l)\right) + Conv\left(\mu(f_s^l)\right)$$



- **Model IP Protection with CUTI-domain**

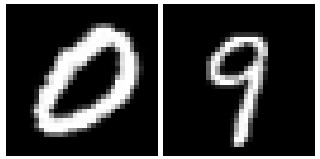


$$\mathcal{L} = \begin{cases} KL(p_s||y_s) - KL(p_i||y_i), & \text{if } epoch = 2e, \\ KL(p_s||y_s) - KL(p_t||y_t), & \text{if } epoch = 2e+1, \end{cases}$$

Experiment: Implementation Details

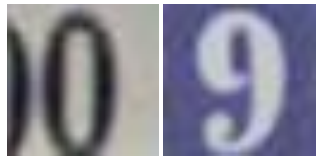
- Digit datasets

10-classification task

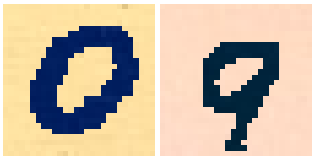


MNIST (MT)

USPS (US)



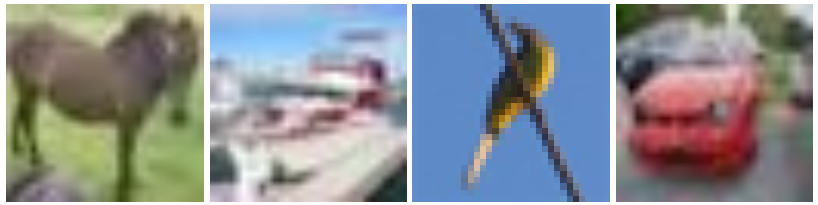
SVHN (SV)



MNIST-M (MM)

- CIFAR & STL

10-classification task



CIFAR10

STL10

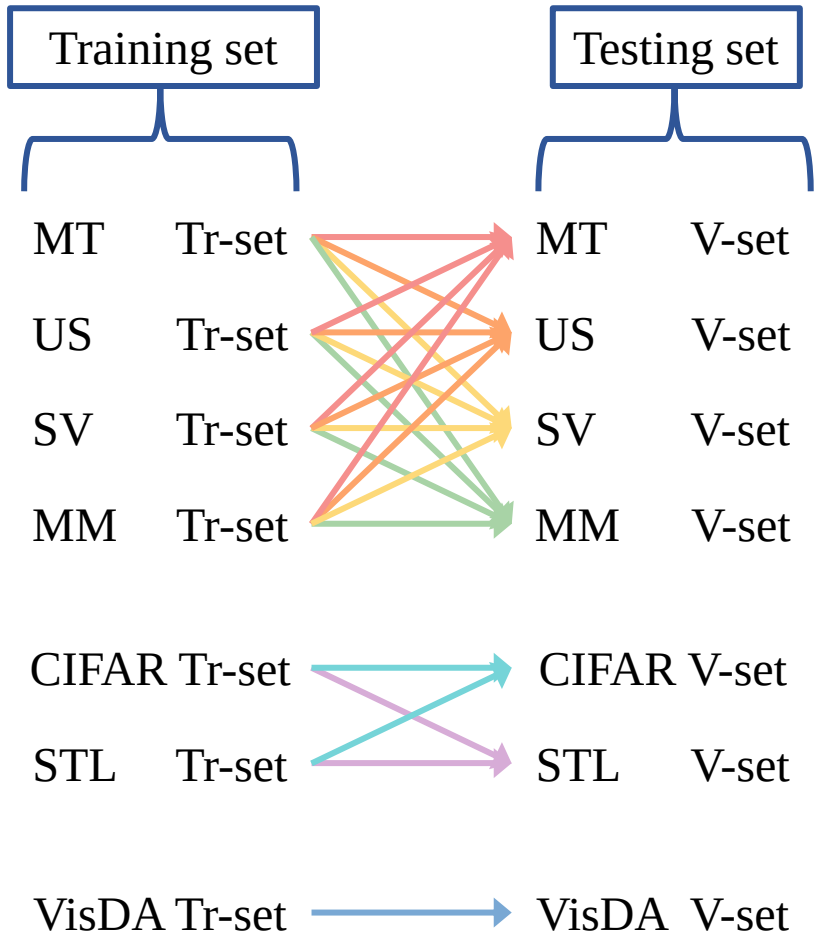
- VisDA-2017

12-classification task

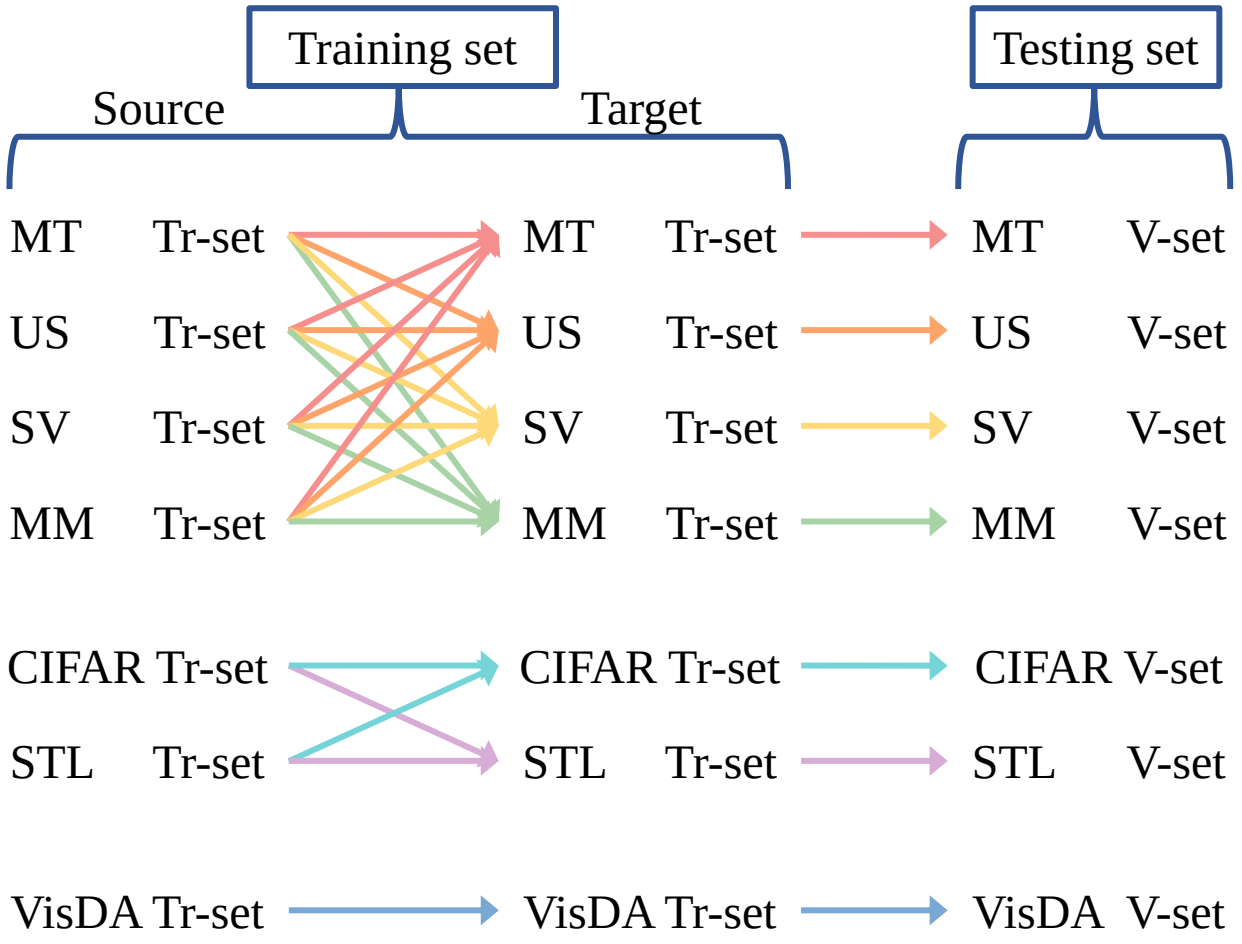
VisDA-2017

Experiment: Implementation Details

Baseline task construction (SL):

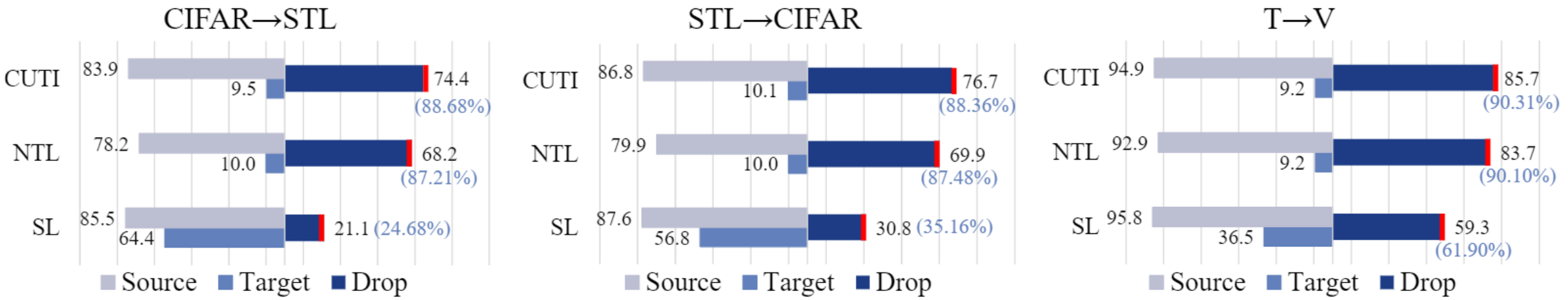


CUTI-Domain task construction:

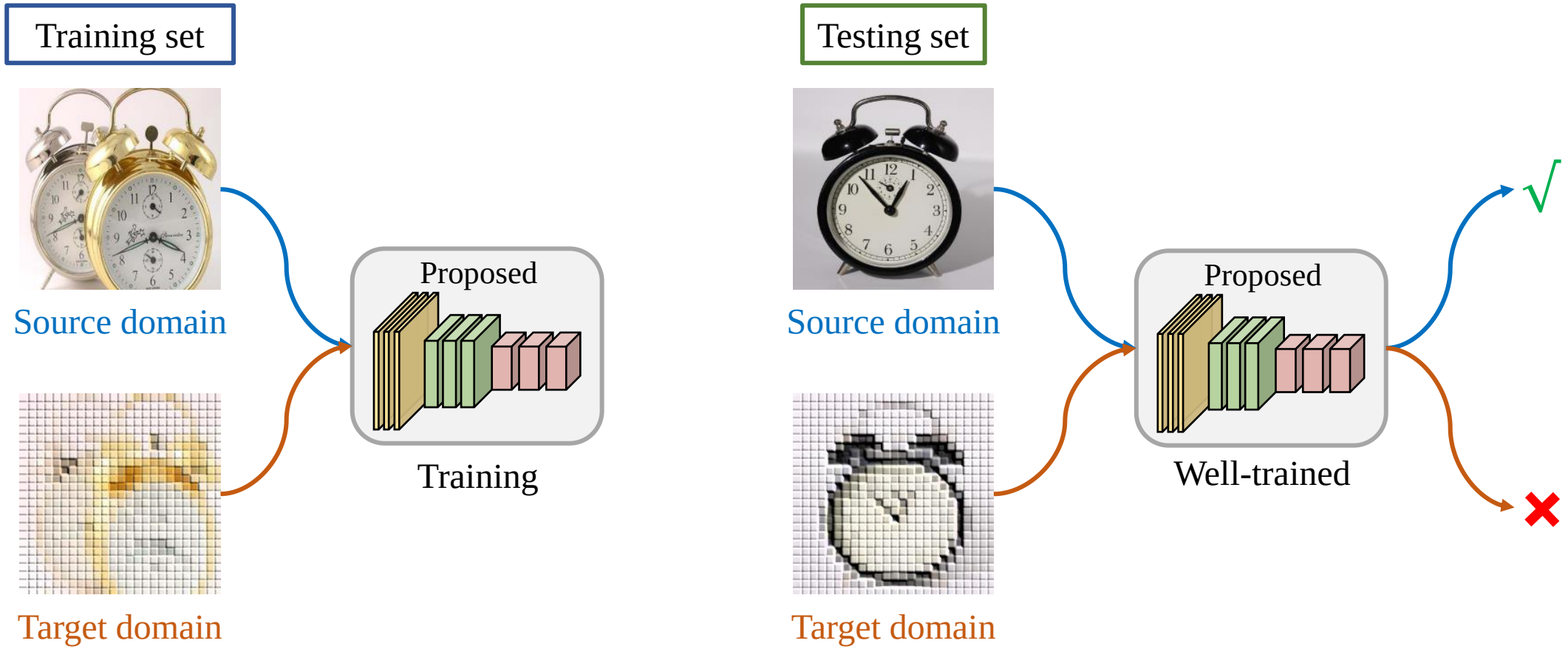


Experiment: Result of Target-Specified CUTI-Domain

Source/Target	MT	US	SN	MM	CUTI Source Drop↓	CUTI Target Drop↑	NTL Source Drop↓	NTL Target Drop↑
MT	99.2 ⇒ 99.1	98.0 ⇒ 6.7	38.2 ⇒ 5.6	67.8 ⇒ 8.7	**0.10 (0.10%)**	**61.00 (88.56%)**	1.00 (1.01%)	46.57 (75.60%)
US	92.6 ⇒ 10.0	99.7 ⇒ 99.6	25.5 ⇒ 6.8	41.2 ⇒ 8.4	**0.10 (0.10%)**	**44.70 (80.72%)**	1.00(1.00%)	38.67 (75.55%)
SN	66.7 ⇒ 9.2	70.5 ⇒ 6.7	91.2 ⇒ 90.9	34.6 ⇒ 10.9	**0.30 (0.33%)**	**48.33 (81.73%)**	1.10(1.23%)	40.60 (77.25%)
MM	98.4 ⇒ 9.5	88.4 ⇒ 6.8	46.3 ⇒ 7.6	95.4 ⇒ 95.4	**0.00 (0.00%)**	**69.73 (88.75%)**	2.10(2.30%)	60.10 (76.95%)
Mean	/	/	/	/	**0.13 (0.13%)**	**55.94 (84.94%)**	1.30 (1.39%)	46.48 (76.34%)



Experiment: Result of Ownership Verification



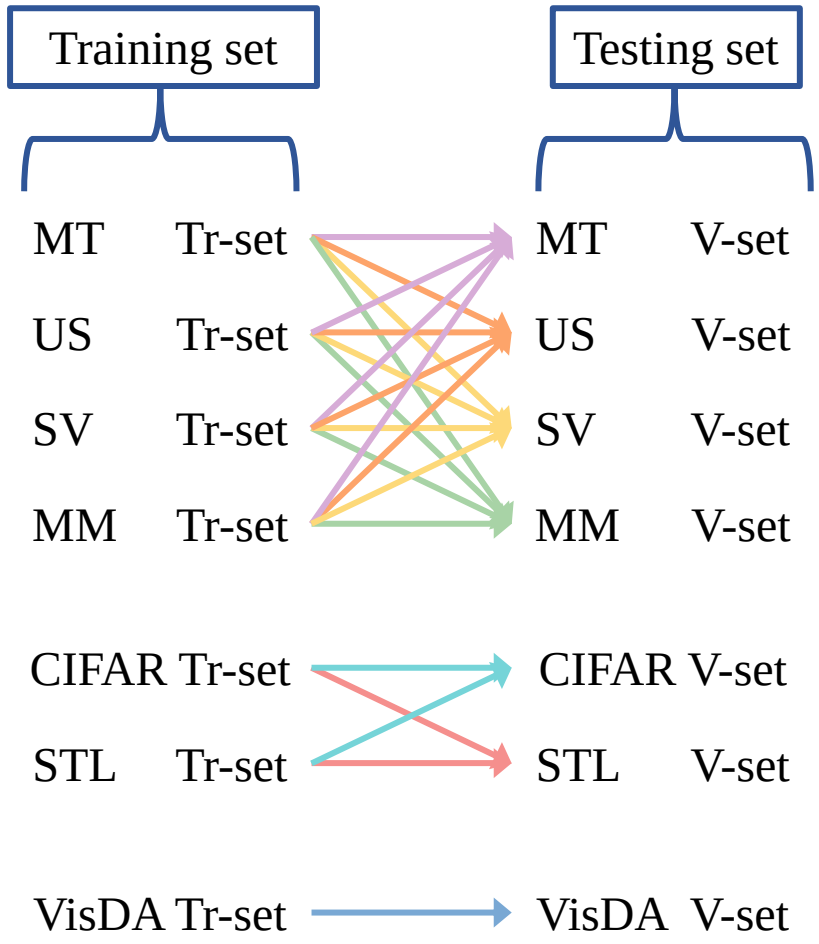
Experiment: Result of Ownership Verification

Source without Patch	Training Methods		Watermark Removal Approaches on CUTI					Avg Drop↑	
	SL	CUTI	FTAL [1]	RTAL [1]	EWC [5]	AU [5]	Overwriting	CUTI	NTL
	[Test w/o Watermark (%)]		[Test w/o Watermark (%)]						
MT	99.0 / 99.3	11.3 / 99.1	9.0 / 100.0	9.4 / 100.0	9.7 / 100.0	9.0 / 100.0	9.4 / 96.2	**89.9**	88.4
US	99.8 / 99.8	7.7 / 99.8	8.0 / 100.0	8.7 / 100.0	9.4 / 100.0	9.4 / 100.0	8.7 / 98.6	**90.9**	85.7
SN	91.3 / 92.3	9.9 / 92.1	9.4 / 98.3	13.9 / 97.6	10.8 / 100.0	8.7 / 100.0	10.4 / 95.8	**87.7**	79.0
MM	96.6 / 96.0	16.8 / 96.0	14.3 / 95.4	24.0 / 98.6	14.6 / 100.0	14.6 / 100.0	14.9 / 95.8	**81.5**	77.3
CIFAR	83.3 / 75.1	10.7 / 86.8	14.9 / 97.9	14.9 / 93.8	14.9 / 100.0	9.4 / 97.2	16.7 / 90.3	**81.7**	74.6
STL	87.9 / 93.2	22.0 / 88.2	20.0 / 96.9	26.4 / 93.8	13.9 / 100.0	22.9 / 94.1	21.2 / 89.6	**74.0**	**74.0**
VisDA	93.6 / 92.2	13.1 / 94.1	15.0 / 95.5	20.5 / 95.1	15.3 / 100.0	21.9 / 95.1	19.4 / 96.2	**78.0**	76.8
Mean	/	/	/	/	/	/	/	**83.4**	79.4

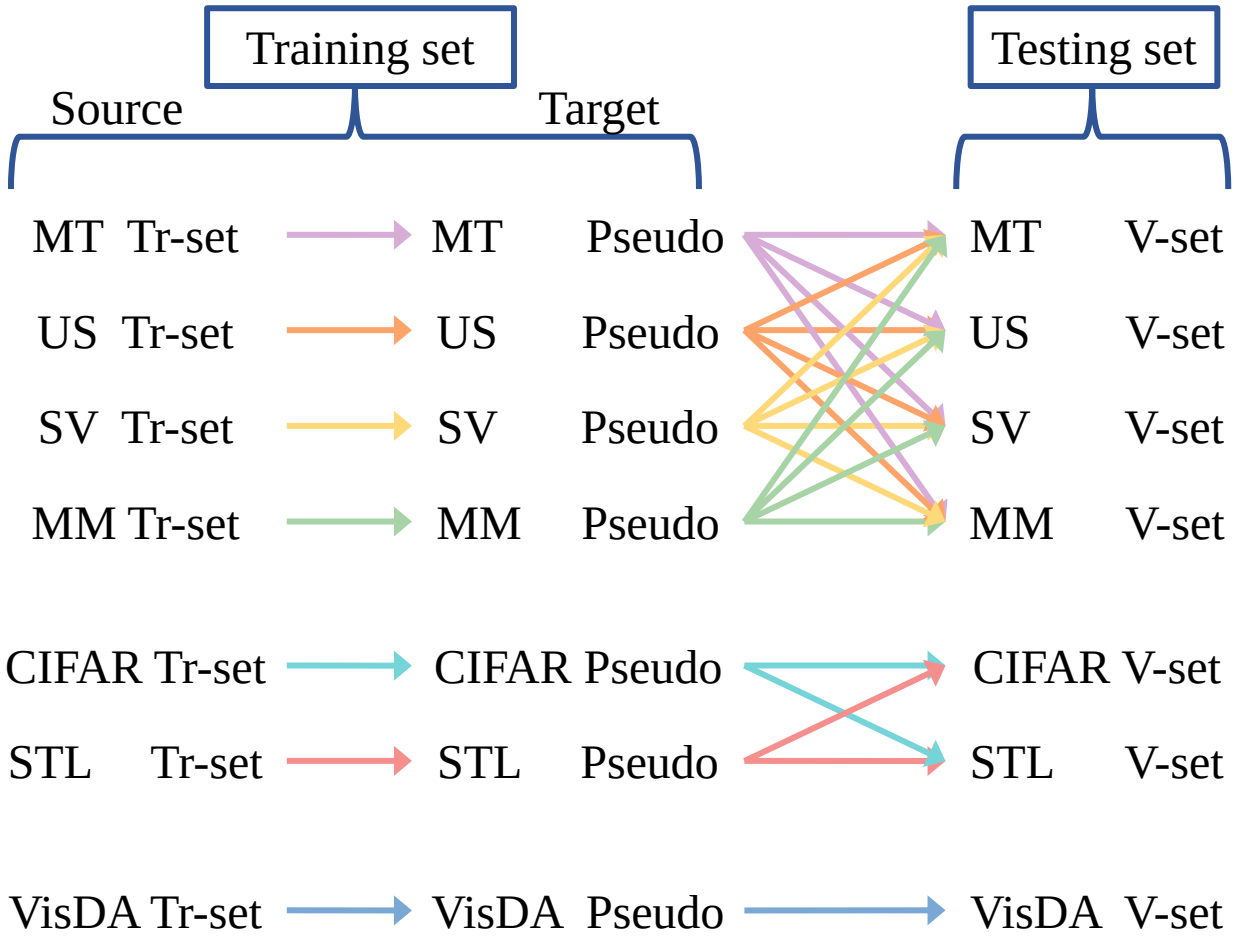
Experiment: Result of Target-free CUTI-Domain

iBRAIN
Intelligent BRain Analysis through Images and Networks

Baseline task construction (SL):

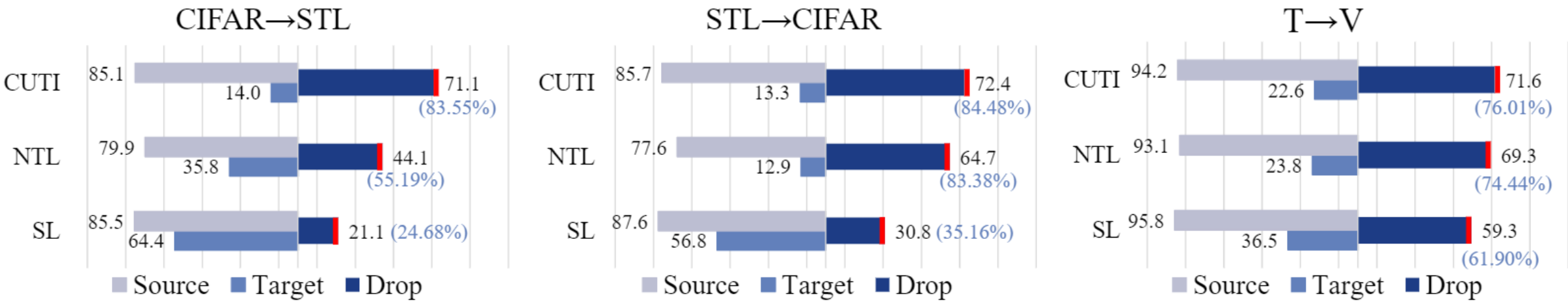


CUTI-Domain task construction:

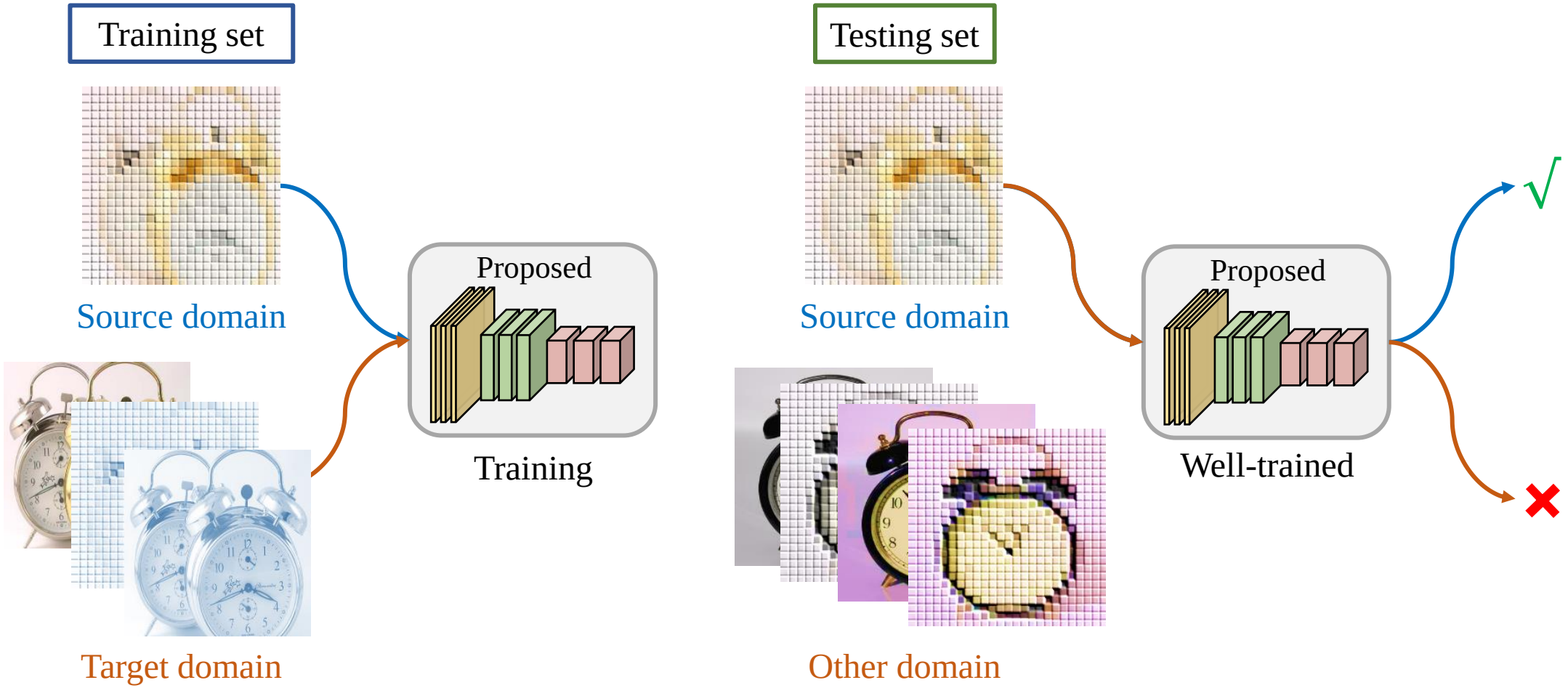


Experiment: Result of Target-free CUTI-Domain

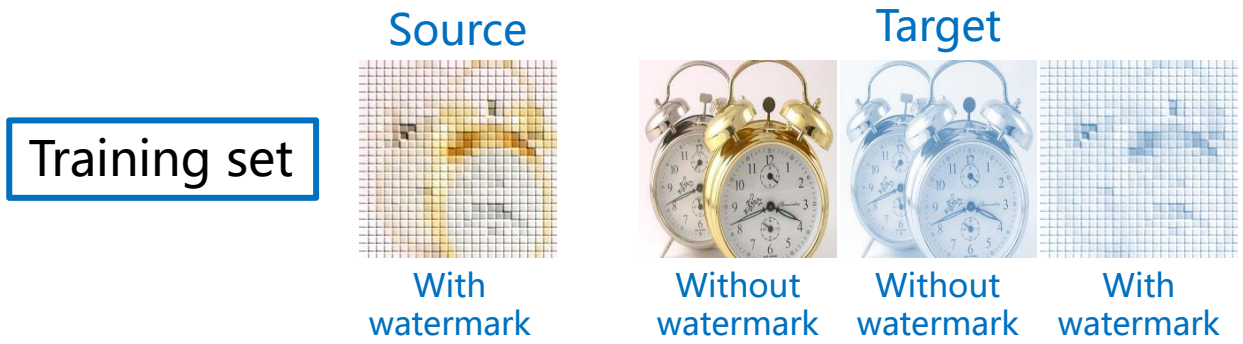
Source/Target	MT	US	SN	MM	CUTI Source Drop↓	CUTI Target Drop↑	NTL Source Drop↓	NTL Target Drop↑
MT	99.2 ⇒ 98.8	98.0 ⇒ 6.7	38.2 ⇒ 6.7	67.8 ⇒ 13.1	**0.40 (0.40%)**	**59.17 (85.43%)**	0.70 (0.71%)	57.30 (84.06%)
US	92.6 ⇒ 9.1	99.7 ⇒ 99.7	99.1 ⇒ 25.5	6.8 ⇒ 8.5	**0.60 (0.60%)**	**44.97 (80.96%)**	**0.60 (0.60%)**	42.90 (74.71%)
SN	66.7 ⇒ 11.9	70.5 ⇒ 14.3	91.2 ⇒ 88.7	34.6 ⇒ 13.6	**2.50 (2.74%)**	**44.00 (74.19%)**	3.20 (3.51%)	41.53 (64.09%)
MM	98.4 ⇒ 19.6	88.4 ⇒ 6.8	46.3 ⇒ 9.5	95.4 ⇒ 95.1	**0.30 (0.31%)**	**65.73 (83.96%)**	2.00 (2.10%)	63.03 (77.62%)
Mean	/	/	/	/	**0.95 (1.02%)**	**53.47 (81.13%)**	1.63 (1.73%)	51.19 (75.12%)



Experiment: Result of Applicability Authorization



Experiment: Result of Applicability Authorization



Source with Path	Test with Path(%)				Test without Path(%)				CUTI Authorized Domain↑	CUTI Other Domain↓	CUTI Drop ↑	NTL Authorized Domain↑	NTL Other Domain↓	NTL Drop ↑
	MT	US	SN	MM	MT	US	SN	MM						
MT	100.0	14.3	17.6	12.9	10.3	8.6	18.3	14.1	**100.0**	**13.7**	**86.27(86.27%)**	99.8	14.5	85.31(85.49%)
US	9.6	99.2	14.9	10.7	10.2	6.7	8.6	10.6	**99.2**	**10.2**	**89.01(89.73%)**	98.5	13.3	85.20(86.50%)
SN	10.8	13.5	99.1	23.0	9.6	9.2	17.7	8.9	99.1	**13.2**	**85.86(86.64%)**	**99.3**	15.8	83.51(84.10%)
MM	8.9	9.1	15.0	99.5	10.0	9.1	11.9	25.9	**99.5**	**12.8**	**86.66(87.09%)**	**99.5**	14.0	85.49(85.92%)
	CIFAR		STL		CIFAR		STL		/					
CIFAR	97.9		42.5		13.1		11.6		**97.9**	**22.4**	**75.50(77.12%)**	97.5	24.6	72.90(74.77%)
STL	29.0		99.9		13.3		15.2		**99.9**	**19.2**	**80.73(80.81%)**	98.6	20.5	78.10(79.21%)
	T		V		T		V		/					
T	100.0		22.9		20.3		10.1		**100.0**	**17.8**	**82.23(82.23%)**	**100.0**	23.3	76.70(76.70%)
Mean	/								**99.4**	**15.5**	**83.75(84.27%)**	99.0	18.0	81.03(81.81%)

Model Barrier: A Compact Un-Transferable Isolation Domain for Model Intellectual Property Protection

Thanks!

http://ibrain.nuaa.edu.cn/

https://www.a-star.edu.sg/

https://www.a-star.edu.sg/ihpc