Test-Time Classifier Adjustment Module for Model-Agnostic Domain Generalization

Yusuke Iwasawa, and Yutaka Matsuo

THE UNIVERSITY OF TOKYO

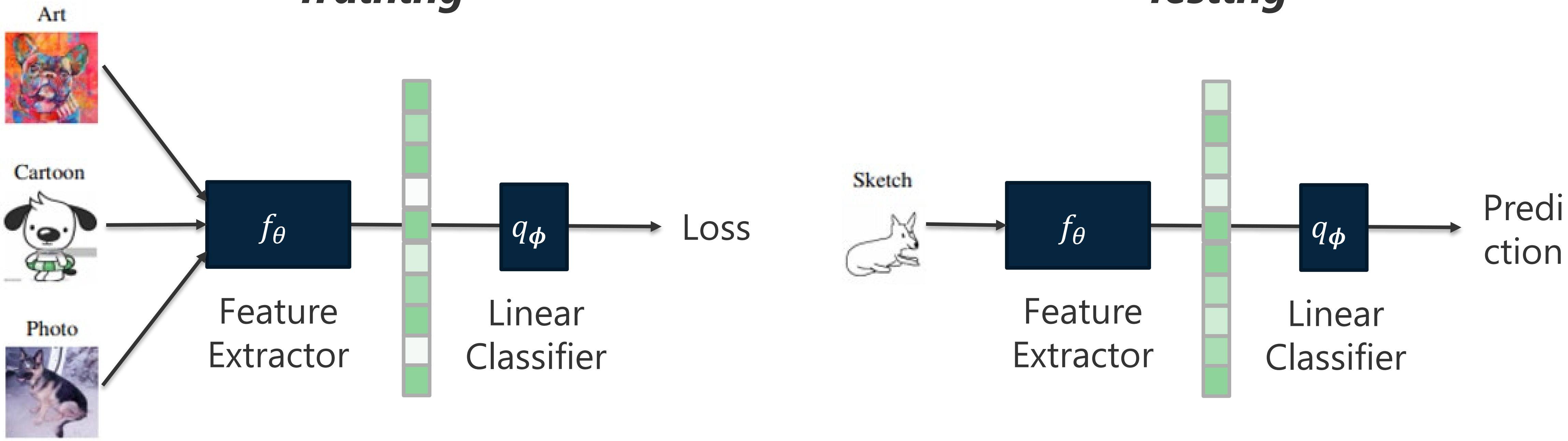
NeurIPS2021 Spotlight presentation

Robustness of DNNs and Cybersecurity

- DNNs become an important component of intelligent system.
  - Preventing catastrophic failure of DNNs become important topic.
- Its behavior under distribution shift might cause security issue.
  - Adversarial attack
  - Weather change in autonomous driving
  - etc

Domain Generalization (DG)

Training ***Testing***



Domain shift

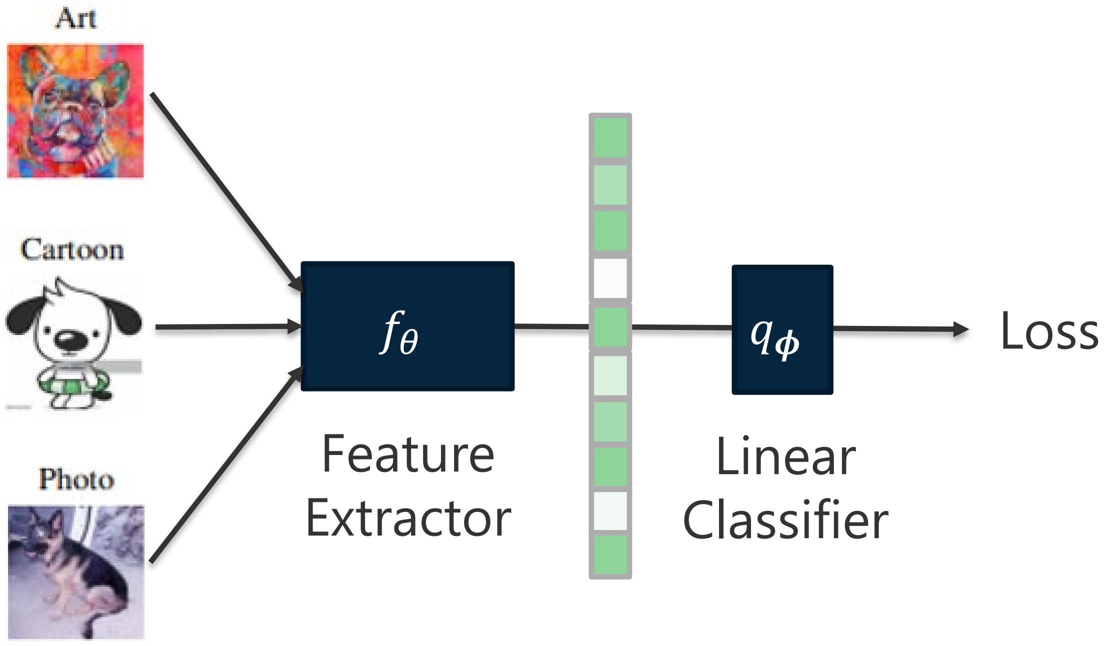
Training on several source domains (Art, Cartoon, and Photo).

Testing on unseen domain (Sketch).

Domain Generalization is a common benchmark setup to the robustness of a predictor to distribution shift (such as variation in light, weather, or object backgrounds)

Existing Domain Generalization Algorithm

Training



Common question:
How to regularize the predictor?

■ Domain Invariant feature learning

- Reduce domain gaps on a space of latent representations.
- **DANN**, **CORAL**, **MMD**, etc.

■ Meta learning

- Learn how to regularize the model to improve the robustness.
- **MLDG** etc.

■ Many others

- **IRM** regularize gradient norm penalty.
- **Domain Mixup** implicitly enhance domain invariance using data augmentation.

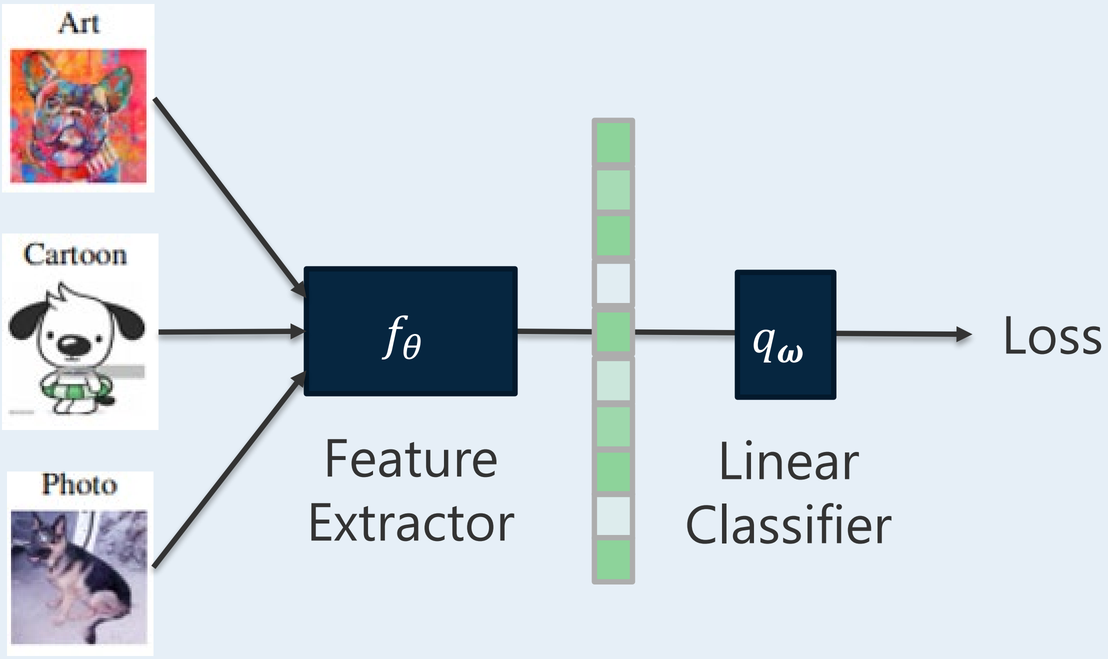
… But ERM is Often Better than DG methods [Gulrajani+ICLR2011]

Table 1: Our ERM baseline outperforms the state-of-the-art in terms of average domain generalizatic performance, even when picking the best competitor per dataset.

Dataset / algorithm	DG accuracy per test domain						Average
Rotated MNIST (full)	0°	15°	30°	45°	60°	75°	
DIVA (Ilse et al., 2019)	95.3	98.7	98.7	98.4	97.7	94.5	97.2
Our ERM	95.9	98.9	98.8	98.9	98.9	96.4	**98.0**
VLCS	C	L	S	V			
G2DM (Albuquerque et al., 2019)	95.5	67.6	69.4	71.1			75.9
Our ERM	97.7	64.3	73.4	74.6			**77.5**
PACS	A	C	P	S			
RSC (Huang et al., 2020)	87.9	82.1	97.9	83.4			**87.8**
Our ERM	84.7	80.8	97.2	79.3			85.5
OfficeHome	A	C	P	R			
DDAIG (Zhou et al., 2020)	59.2	52.3	74.6	76.0			65.5
Our ERM	61.3	52.4	75.8	76.6			**66.5**
All datasets							
Best SOTA competitor							81.6
Our ERM							**81.9**

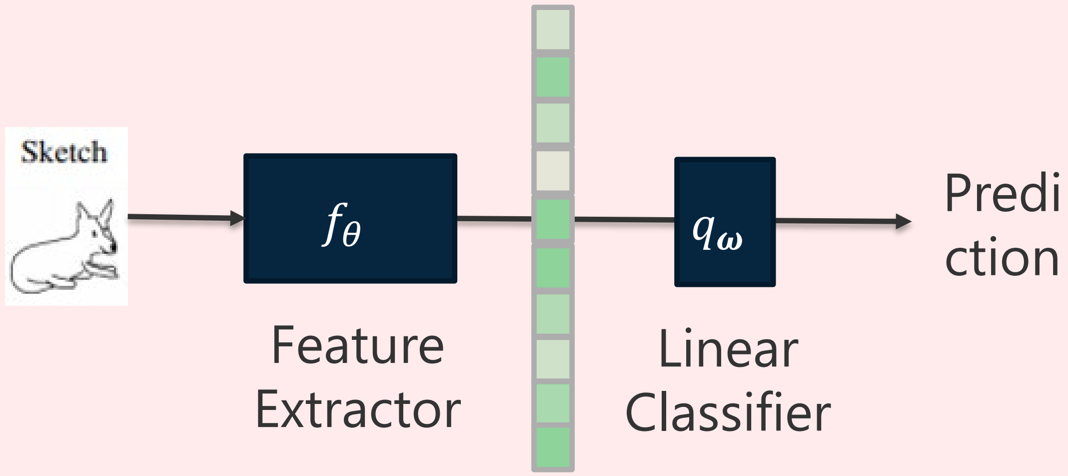
Proposal: Test-Time Adaptation for DG

Training (Prior works)



***Labeled** data from **source domains** are available*

Testing (Our work)



***Unlabeled** and **online** data from **target** domain are available*

Research Question: How can we use off-the-shelf data available at test-time to correlate its prediction by itself?

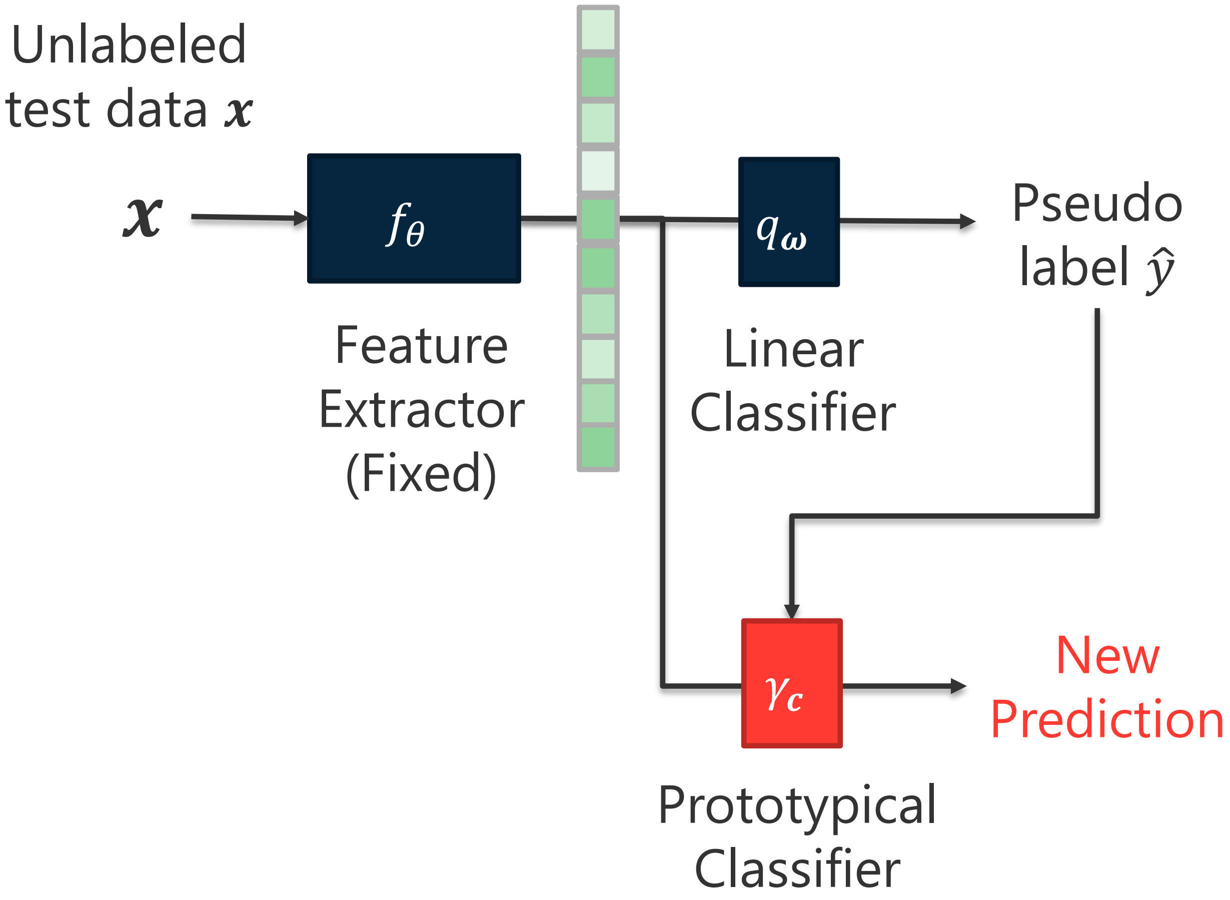
SGD during Test-Time is not Desirable

- Natural way to achieve the goal is to use SGD at test-time.
  - SHOT [Liang+2020] and Tent [Wang+2021] updates parameters to minimize prediction entropy.

- Using SGD during test-time is not desirable.
  - (1) It harm inference throughput.
  - (2) It can lead catastrophic failure.

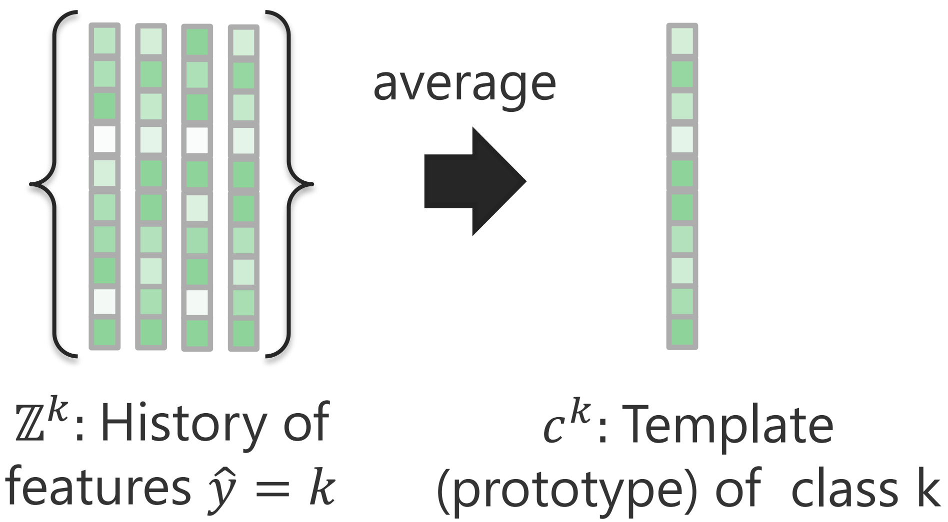
- Tent [Wang+2021] avoid the second issue by only updating small portion of parameters (BN layer).
  - But many recent architecture (BiT, ViT, and MLP-Mixer) does not employ BN.

Proposal: Test-Time Template Adjuster (T3A)



(1) Pseudo Prototype

Update templates using pseudo label and intermediate features.

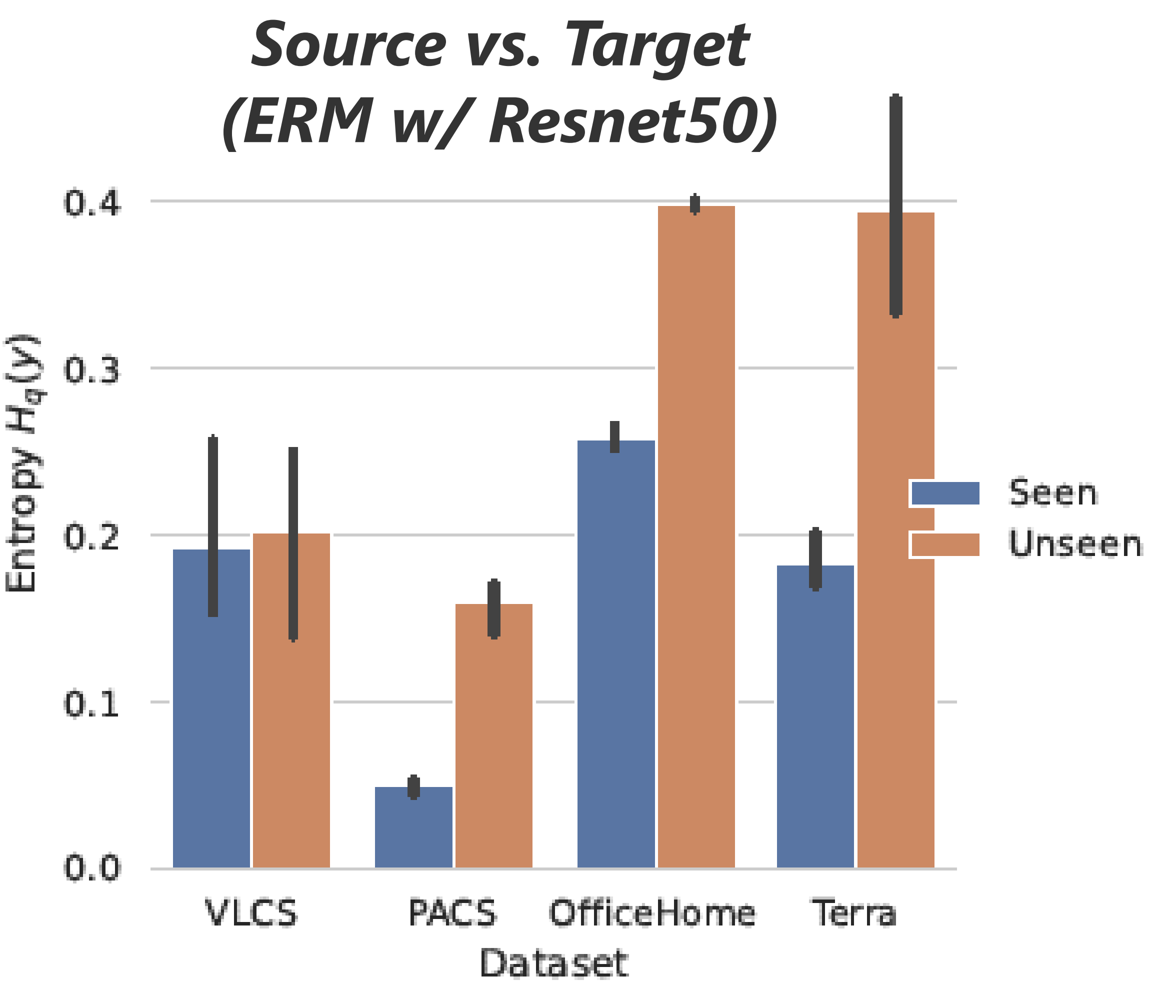


(2) Prototypical Classification

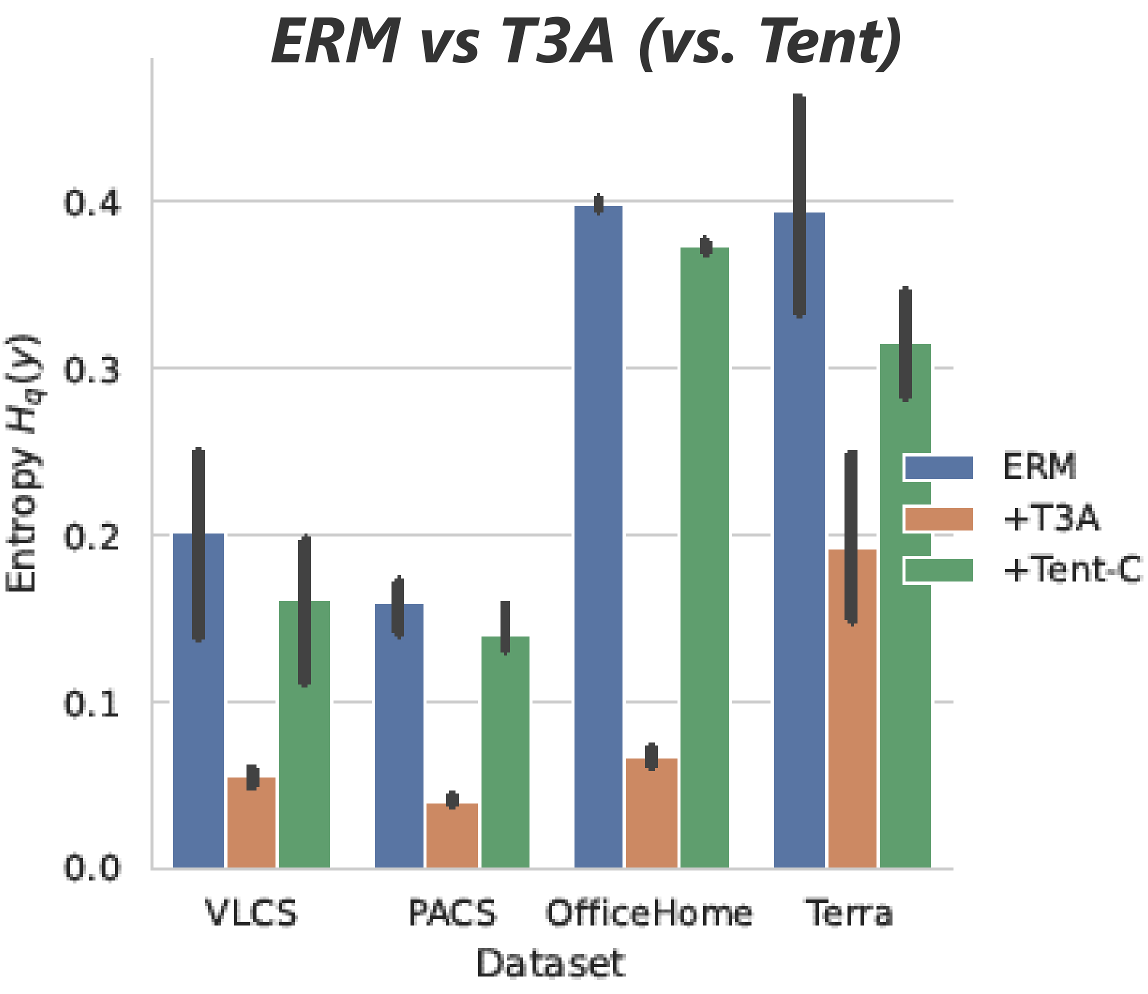
Classify each sample based on its distance to the pseudo-prototype

This procedure will be repeated every time the model encounter new examples

T3A Implicitly Reduce Prediction Entropy



Source <<< Target



ERM >> T3A

Experimental Setup

- Dataset
  - VLCS, PACS, OfficeHome, and TerraIncognita

- Experimental procedure strictly follows DomainBed [Gulrajani+ICLR2011]
  - Training-domain validation for selecting hyperparameters
  - All experiments repeat 3 times with different seeds

Results: Comparison to DG and Tent

Table 1: Domain generalization accuracy for all datasets and algorithms. Bold type indicates performance improvement from the base model, and * indicates statistical significance in one-sided paired t-test (** indicates $p \leq 0.01$, * indicates $p \leq 0.05$).

Algorithm	VLCS	PACS	OfficeHome	Terra	Avg
ERM	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	69.0
IRM	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	47.6 ± 0.8	68.5
GroupDRO	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	43.2 ± 1.1	67.6
Mixup	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	69.5
MLDG	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	47.7 ± 0.9	69.2
CORAL	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	47.6 ± 1.0	70.3
MMD	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	42.2 ± 1.6	67.7
DANN	78.6 ± 0.4	83.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	68.7
CDANN	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	45.8 ± 1.6	67.9
MTL	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	45.6 ± 1.2	68.5
SagNet	77.8 ± 0.5	86.3 ± 0.2	68.1 ± 0.1	48.6 ± 1.0	70.2
ARM	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	45.5 ± 0.3	68.3
VREx	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	46.4 ± 0.6	69.0
RSC	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	46.6 ± 1.0	68.6
ERM†	77.7 ± 0.1	83.6 ± 0.9	66.4 ± 0.3	46.5 ± 0.3	68.6
+T3C (Ours)	**80.0** ± 0.2	**85.3** ± 0.6	**68.3** ± 0.1	**47.0** ± 0.6	**70.1****
+Tent-BN	68.2 ± 0.2	**84.8** ± 0.5	**67.0** ± 0.4	44.7 ± 0.3	66.2
+Tent-C	77.0 ± 0.4	82.3 ± 1.2	65.7 ± 0.2	45.5 ± 0.4	67.6
CORAL†	78.6 ± 0.5	84.2 ± 0.3	68.3 ± 0.1	48.1 ± 1.3	69.8
+T3C (Ours)	**79.5** ± 0.5	**85.6** ± 0.2	**69.2** ± 0.2	47.3 ± 0.7	**70.4***
+Tent-BN	71.4 ± 0.7	**85.6** ± 0.2	**69.2** ± 0.2	46.5 ± 0.5	68.2
+Tent-C	78.1 ± 0.5	83.7 ± 0.4	68.2 ± 0.1	47.8 ± 1.1	69.5

DG [Gulrajani+2021]

T3A >= DG

ERM + T3A (and Tent)

T3A > ERM

T3A > TTA

CORAL + T3A (and Tent)

T3A > CORAL

Results: Performance on Various Backbone Networks

Table 2: Domain generalization accuracy with different backbone networks. T3A increases the performance agnostic to backbone networks. Note that, this experiments is conducted only on the default hyperparameters of ERM. Bold type indicates performance improvement, and * indicates statistical significance in paired t-test (** indicates $p \leq 0.01$, * indicates $p \leq 0.05$).

	Models	VLCS	PACS	OfficeHome	Terra	Avg
Convolution	resnet18	73.2 ± 0.9	80.3 ± 0.4	55.7 ± 0.2	40.7 ± 0.3	62.5
	+T3A	**76.5 ± 0.9**	**81.7 ± 0.1**	**57.0 ± 0.4**	**41.6 ± 0.5**	64.2*
	resnet50	75.5 ± 0.1	83.9 ± 0.2	64.4 ± 0.2	45.4 ± 1.2	67.3
	+T3A	**78.3 ± 0.7**	**84.5 ± 0.3**	**66.5 ± 0.2**	**45.9 ± 0.5**	68.8*
	BiT-M-R50x3	76.7 ± 0.1	84.4 ± 1.2	69.2 ± 0.6	52.5 ± 0.3	70.7
	+T3A	**79.7 ± 0.3**	**85.4 ± 0.9**	**71.7 ± 0.6**	52.2 ± 0.6	72.3*
	BiT-M-R101x3	75.0 ± 0.6	84.0 ± 0.7	67.7 ± 0.5	47.8 ± 0.8	68.6
	+T3A	**78.6 ± 0.4**	**85.4 ± 0.5**	**69.9 ± 0.4**	**48.1 ± 0.8**	70.5*
	BiT-M-R152x2	76.7 ± 0.3	85.2 ± 0.1	71.3 ± 0.6	51.4 ± 0.6	71.1
	+T3A	**79.1 ± 0.4**	**86.4 ± 0.1**	**73.2 ± 0.5**	50.9 ± 0.7	72.4*
ViT	ViT-B16	79.2 ± 0.3	85.7 ± 0.1	78.4 ± 0.3	41.8 ± 0.6	71.3
	+T3A	**80.2 ± 0.4**	**86.0 ± 0.1**	**78.9 ± 0.3**	**42.5 ± 0.7**	71.9*
	ViT-L16	78.2 ± 0.5	84.6 ± 0.5	78.0 ± 0.1	42.7 ± 1.9	70.9
	+T3A	**79.0 ± 0.6**	**85.5 ± 0.6**	**78.7 ± 0.2**	**45.3 ± 0.4**	72.1**
Hybrid	DeiT	79.3 ± 0.4	87.8 ± 0.5	76.6 ± 0.3	50.0 ± 0.2	73.4
	+T3A	**81.3 ± 0.4**	**89.5 ± 0.4**	**78.3 ± 0.2**	**50.1 ± 0.2**	74.8*
	HViT	79.2 ± 0.5	89.7 ± 0.4	80.0 ± 0.2	51.4 ± 0.9	75.1
	+T3A	**81.0 ± 0.1**	**90.4 ± 0.5**	**80.5 ± 0.2**	**52.3 ± 1.0**	76.1*
MLP-Mixer	Mixer-L16	76.4 ± 0.2	81.3 ± 1.0	69.4 ± 1.6	37.1 ± 0.4	66.1
	+T3A	**80.3 ± 0.3**	**83.0 ± 0.8**	**72.3 ± 1.8**	**37.5 ± 0.8**	68.3*

Results: Performance on Various Backbone Networks

Models	VLCS	PACS	OfficeHome	Terra	Avg
resnet18	73.2 ± 0.9	80.3 ± 0.4	55.7 ± 0.2	40.7 ± 0.3	62.5
+T3A	$\mathbf{76.5 \pm 0.9}$	$\mathbf{81.7 \pm 0.1}$	$\mathbf{57.0 \pm 0.4}$	$\mathbf{41.6 \pm 0.5}$	64.2*
resnet50	75.5 ± 0.1	83.9 ± 0.2	64.4 ± 0.2	45.4 ± 1.2	67.3
+T3A	$\mathbf{78.3 \pm 0.7}$	$\mathbf{84.5 \pm 0.3}$	$\mathbf{66.5 \pm 0.2}$	$\mathbf{45.9 \pm 0.5}$	68.8*
BiT-M-R50x3	76.7 ± 0.1	84.4 ± 1.2	69.2 ± 0.6	52.5 ± 0.3	70.7
+T3A	$\mathbf{79.7 \pm 0.3}$	$\mathbf{85.4 \pm 0.9}$	$\mathbf{71.7 \pm 0.6}$	52.2 ± 0.6	72.3*
BiT-M-R101x3	75.0 ± 0.6	84.0 ± 0.7	67.7 ± 0.5	47.8 ± 0.8	68.6
+T3A	$\mathbf{78.6 \pm 0.4}$	$\mathbf{85.4 \pm 0.5}$	$\mathbf{69.9 \pm 0.4}$	$\mathbf{48.1 \pm 0.8}$	70.5*
BiT-M-R152x2	76.7 ± 0.3	85.2 ± 0.1	71.3 ± 0.6	51.4 ± 0.6	71.1
+T3A	$\mathbf{79.1 \pm 0.4}$	$\mathbf{86.4 \pm 0.1}$	$\mathbf{73.2 \pm 0.5}$	50.9 ± 0.7	72.4*
ViT-B16	79.2 ± 0.3	85.7 ± 0.1	78.4 ± 0.3	41.8 ± 0.6	71.3
+T3A	$\mathbf{80.2 \pm 0.4}$	$\mathbf{86.0 \pm 0.1}$	$\mathbf{78.9 \pm 0.3}$	$\mathbf{42.5 \pm 0.7}$	71.9*
ViT-L16	78.2 ± 0.5	84.6 ± 0.5	78.0 ± 0.1	42.7 ± 1.9	70.9
+T3A	$\mathbf{79.0 \pm 0.6}$	$\mathbf{85.5 \pm 0.6}$	$\mathbf{78.7 \pm 0.2}$	$\mathbf{45.3 \pm 0.4}$	72.1**
DeiT	79.3 ± 0.4	87.8 ± 0.5	76.6 ± 0.3	50.0 ± 0.2	73.4
+T3A	$\mathbf{81.3 \pm 0.4}$	$\mathbf{89.5 \pm 0.4}$	$\mathbf{78.3 \pm 0.2}$	$\mathbf{50.1 \pm 0.2}$	74.8*
HViT	79.2 ± 0.5	89.7 ± 0.4	80.0 ± 0.2	51.4 ± 0.9	75.1
+T3A	$\mathbf{81.0 \pm 0.1}$	$\mathbf{90.4 \pm 0.5}$	$\mathbf{80.5 \pm 0.2}$	$\mathbf{52.3 \pm 1.0}$	76.1*
Mixer-L16	76.4 ± 0.2	81.3 ± 1.0	69.4 ± 1.6	37.1 ± 0.4	66.1
+T3A	$\mathbf{80.3 \pm 0.3}$	$\mathbf{83.0 \pm 0.8}$	$\mathbf{72.3 \pm 1.8}$	$\mathbf{37.5 \pm 0.8}$	68.3*

Conv

ViT

ViT

Hybrid

Hybrid

MLP-Mixer

MLP

Concluding Remarks

- We present T3A, optimization-free test-time adaptation method for improves robustness against domain shift.
  - vs. DG: T3A focus on test-phase
  - vs. Test time adaptation: T3A is optimization-free

- Our method stably improves robustness against domain shift on various backbone networks and various datasets.

- Further results will be presented on paper and poster.
  - Full results for each datasets and backbone networks.
  - Hyperparameter sensitivity.
  - Comparison with various test-time adaptation methods.

Comparison with Existing Test-Time Adaptation

Method	Description	Optimization-free	Model-agnostic
Pseudo Label	Update parameters to minimize cross entropy with pseudo label.		✓
SHOT	Update parameters w/ PL loss, entropy,		✓
TENT	Update BN transformation parameters to minimize entropy.		
BN Norm	Updates BN statistics during test time.	✓	
T3A (Ours)	**Replace classifier templates w/ pseudo labeling**	✓	✓