ODS: Test-Time Adaptation in the Presence of Open-World Data Shift

Zhi Zhou[1], Lan-Zhe Guo[1], Lin-Han Jia[1], Ding-Chu Zhang[1], Yu-Feng Li[1†]

1 Nanjing University, Nanjing, China

LAMDA

Learning And Mining from DatA

http://www.lamda.nju.edu.cn

† Corresponding author

What is this work about

Test-time adaptation adapts the model to distribution shifts without source data.

However, current test-time adaptation account for relatively simple distribution shift, such as covariate shift, which challenges in the following two aspects:

- **TTA degenerates when label and covariate distribution shifts are mixed**
- **TTA cannot adapt to changed label distribution shift**

These two points are very crucial for deploying test-time adaptation in the real world.

✓ In our work, we study an **Open-World Data Shift** setting for test-time adaptation and where the model needs to adapt to **both covariate and label distribution shifts**.

✓ We propose **a test-time adaptation framework ODS** to solve the above open-world data shift setting, which can apply to many existing test-time algorithms.

✓ Our proposal is **clearly better than** one baseline and six test-time adaptation methods evaluated on two benchmark datasets.

Outline

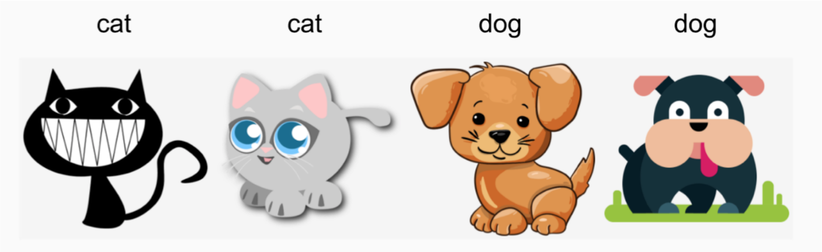
- **Background**
- ODS Framework
- Experiments
- Conclusions

Distribution Shift

Covariate Shift

Training stage

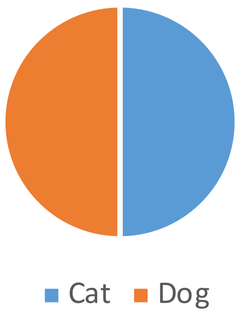
Testing stage

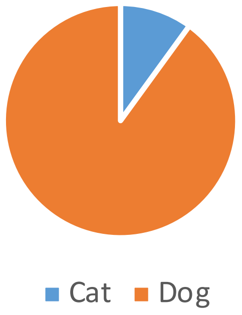


Label Shift

Training stage

Testing stage

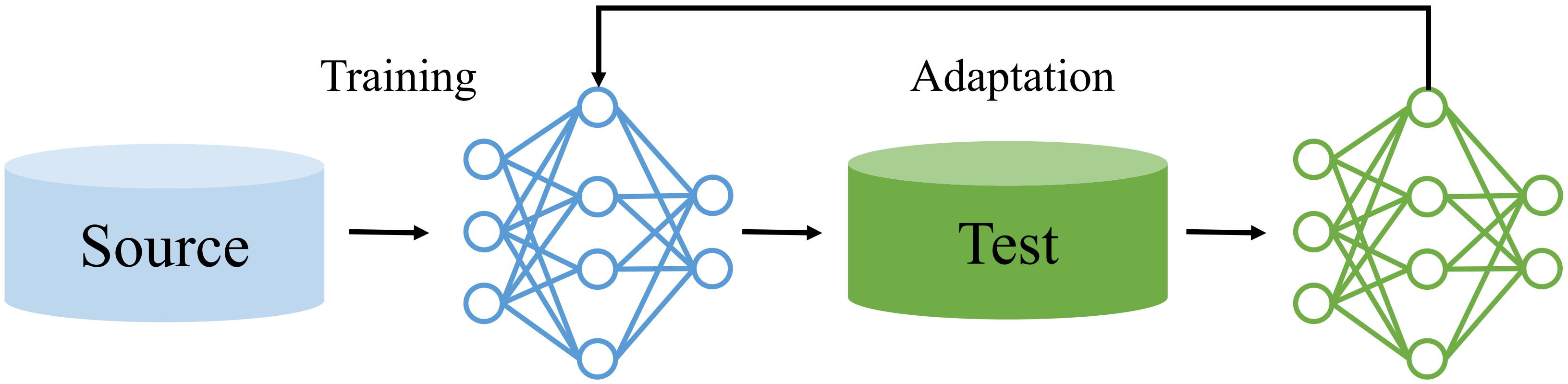




Machine learning models suffer from performance degradation when encountering distribution shifts.

* Graphs come from Dive into deep learning.

Test-Time Adaptation



- ➢ Do not require source data during adaptation
- ➢ Lightweight and efficient adaptation for trained models
- ➢ Continuously adapt to distribution shifts

Test-time adaptation methods fail to adapt to both label and covariate distribution shifts at the same time.

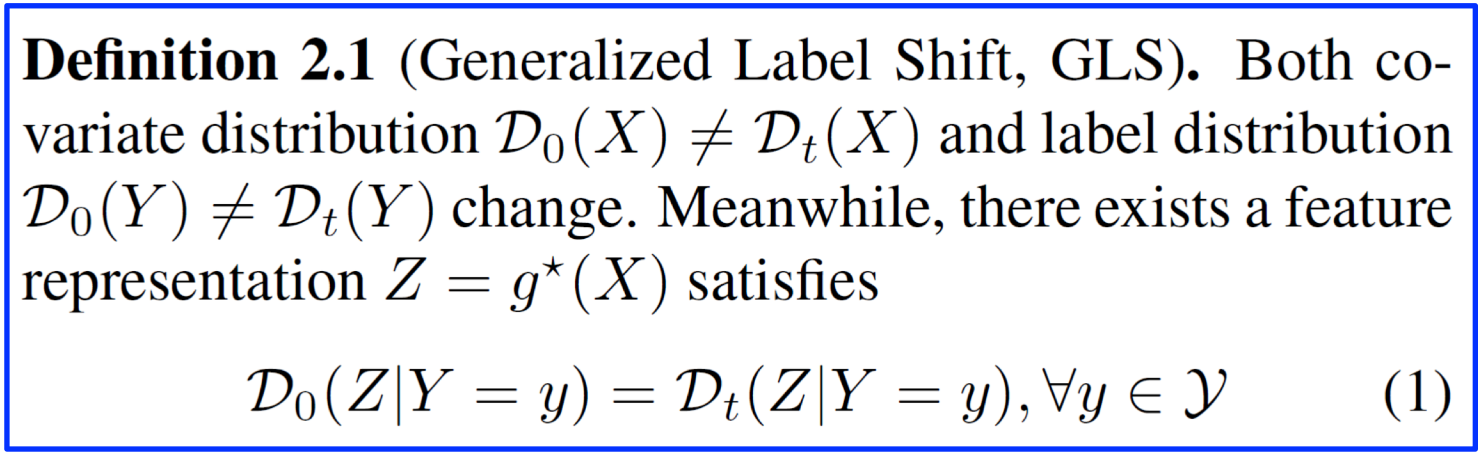
Outline

- Background
- **ODS Framework**
- Experiments
- Conclusions

Motivation

The assumption of shift in label and covariate distribution

- To ensure task feasibility, we adopt the generalized label shift assumption [1].
- X and Y represent random variables of the sample and label.
- Z represents the random variable of optimal feature representation.
- The generalized label shift assumption ensures there is an optimal feature representation, making $\mathcal{D}_t(Z|Y)$ remain fixed over time t.



Definition 2.1 (Generalized Label Shift, GLS)**.** Both covariate distribution $\mathcal{D}_0(X) \neq \mathcal{D}_t(X)$ and label distribution $\mathcal{D}_0(Y) \neq \mathcal{D}_t(Y)$ change. Meanwhile, there exists a feature representation $Z = g^\star(X)$ satisfies

$$\mathcal{D}_0(Z|Y = y) = \mathcal{D}_t(Z|Y = y), \forall y \in \mathcal{Y} \qquad (1)$$

[1] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, Geoffrey J. Gordon: Domain Adaptation with Conditional Distribution Matching and Generalized Label Shift. NeurIPS 2020

Motivation

Optimizing the label distribution of prediction

- We adopt a simple yet effective strategy, i.e., logit adjustment [1], to optimize the label distribution of prediction.
- $f_{\theta_t}(Y = y|X)$ represents the logit predicted by the model for class y at time t.
- $w_{t,y}$ represents the ground-truth label distribution for class y at time t.
- Therefore, the prediction is the class with the maximum calibrated logit value.
- We find that the classification error can be reduced using appropriate w_t.

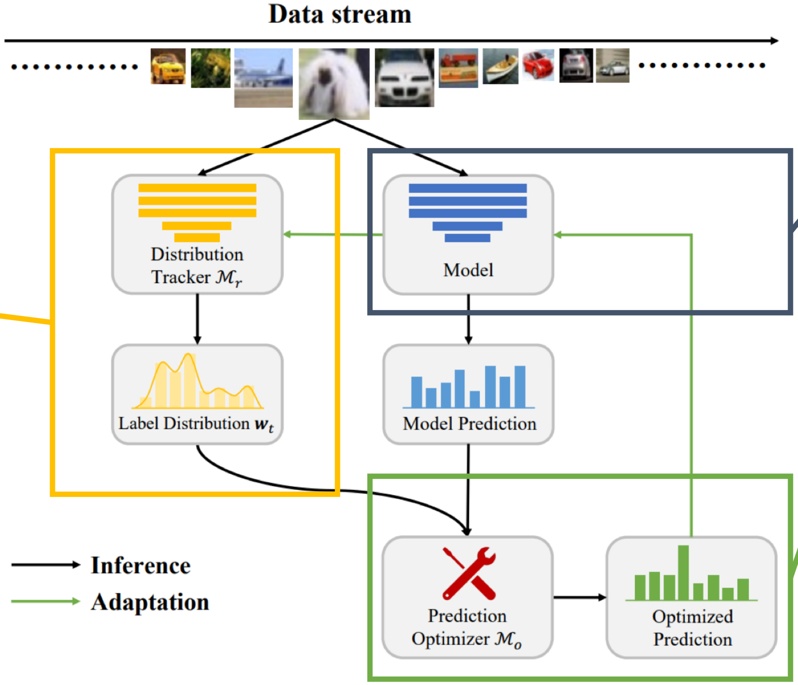
$$\boxed{\hat{Y}_o = \arg\max_{y \in \mathcal{Y}} f_{\theta_t}(Y = y|X) + \ln w_{t,y}}$$

[1] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, Sanjiv Kumar: Long-tail learning via logit adjustment. ICLR 2021

ODS Framework

Overall illustration

- The ODS framework contains two modules:
  - Distribution Tracker $\mathcal{M}_T$: Estimating label distribution $\boldsymbol{w}_t$ for subsequent adaptation and predictive optimization.
  - Prediction Optimizer $\mathcal{M}_O$: Improving the prediction using $\boldsymbol{w}_t$.


Data stream
Distribution Tracker $\mathcal{M}_r$
Model
Label Distribution $\boldsymbol{w}_t$
Model Prediction
Prediction Optimizer $\mathcal{M}_o$
Optimized Prediction
Inference
Adaptation

Estimating shifted label distribution w_t at each timestamp t

Help the model perform class-balanced adaptation

Optimizing the prediction to accommodate label distribution shift

ODS Framework

Optimizing objective

- The objective of ODS framework contains two parts:
  - A weighting term: Applying the estimated label distribution $\boldsymbol{w}_t$ to weight the entropy minimization loss.
  - An entropy minimization term: Adapting the model in an unsupervised fashion.

Weighted updating for test-time adaptation to maintain an internal class-balanced model

Entropy minimization loss to adapt the model in an unsupervised fashion

$$\min_{\theta_t} \frac{1}{N_t} \sum_{i=1}^{N_t} \sum_{k=1}^{K} \boxed{S(\boldsymbol{w}_t)_k} \boxed{f_{\theta_t}(Y = k|\boldsymbol{x}_i) \log f_{\theta_t}(Y = k|\boldsymbol{x}_i)}$$

s.t. $\boldsymbol{w}_t$ is estimated by $\mathcal{M}_T$

ODS Framework

Distribution Tracker $\mathcal{M}_T$

- Distribution tracker estimates the label distribution $\boldsymbol{w}_t$
  - Black box shift estimation (BBSE) [1] is a powerful tool to estimate the test label distribution shift $\mathcal{D}_t(Y)/\mathcal{D}_0(Y)$.
  - Using covariance matrix $\hat{C}$ model and current label distribution γ_t estimated by source model f_{θ_0}.

$$\frac{\mathcal{D}_t(Y)}{\mathcal{D}_0(Y)} = \hat{C}_{\hat{Y},Y}^{-1}\gamma_t$$

  - **However, this does not work!**
  - **BBSE assume $\mathcal{D}_0(X|Y) = \mathcal{D}_t(X|Y)$ which not holds for test-time adaptation settings.**

[1] Zachary C. Lipton, Yu-Xiang Wang, Alexander J. Smola: Detecting and Correcting for Label Shift with Black Box Predictors. ICML 2018: 3128-3136

ODS Framework

Distribution Tracker $\mathcal{M}_T$

- Distribution tracker estimates the label distribution $\boldsymbol{w}_t$
  - Recall the generalized label shift assumption, we have $\mathcal{D}_0(Z|Y) = \mathcal{D}_t(Z|Y)$.
  - **The adapted feature representation Z can help!**
  - We adopt the following objective to optimize the pseudo labels for estimating the label distribution $\boldsymbol{w}_t$.

$$\min_{\boldsymbol{w}_t} \sum_{i=1}^{N_t} \left[\boldsymbol{z}_i^\top \log f_{\theta_0}(Y|\boldsymbol{x}_i) + \boldsymbol{z}_i^\top \log \boldsymbol{z}_i - \sum_{j=1}^{N_t} \boxed{s_{ij}} \boldsymbol{z}_i^\top \boldsymbol{z}_j \right]$$

$$\text{s.t.} \quad \boldsymbol{w}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \boldsymbol{z}_i$$

Instance-wise similarity is calculated by Z.

  - Luckily, this objective can be effectively optimized [1].

[1] Malik Boudiaf, Romain Müller, Ismail Ben Ayed, Luca Bertinetto: Parameter-free Online Test-time Adaptation. CVPR 2022: 8334-8343

ODS Framework

Prediction Optimizer $\mathcal{M}_O$

- Prediction Optimizer optimizes the model prediction
  - **Statistics Optimization**

$$\hat{Y}_o = \arg\max_{y \in \mathcal{Y}} f_{\theta_t}(Y = y | X) + \ln w_{t,y}$$

  - **Distribution Optimization**

$$\hat{Y}_o = \arg\max_{k \in \mathcal{Y}} \frac{\sqrt{\boldsymbol{z}_{i,k} f_{\theta_t}(Y = k \mid \boldsymbol{x}_i)}}{\sum_{k' \in \mathcal{Y}} \sqrt{\boldsymbol{z}_{i,k'} f_{\theta_t}(Y = k' \mid \boldsymbol{x}_i)}}$$

Outline

- Background
- ODS Framework
- **Experiments**
- Conclusions

Experiments

- In our paper, we mainly answer the following three research questions:

  - **RQ1:** Whether ODS can outperform prior TTA methods when encountering open-world data shift?
  - **RQ2:** Whether ODS is generic to integrate with different TTA methods and boost their performance?
  - **RQ3:** Does ODS accurately estimate label distribution and effectively optimize the prediction?

Experiments

- **RQ1:** Whether ODS can outperform prior TTA methods when encountering open-world data shift?

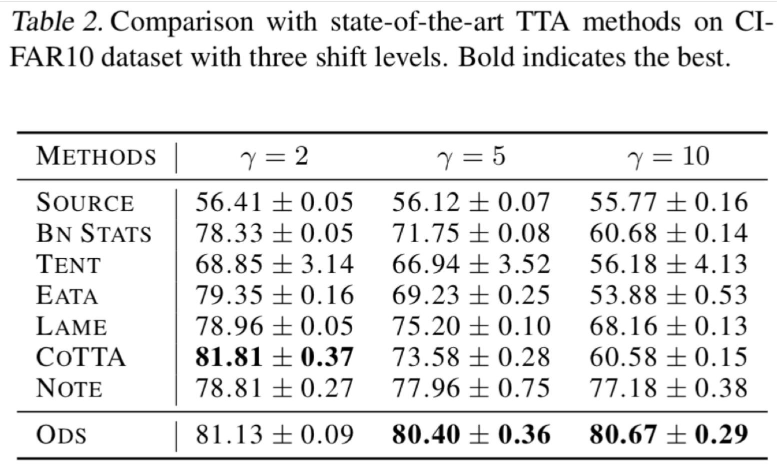


Table 2. Comparison with state-of-the-art TTA methods on CIFAR10 dataset with three shift levels. Bold indicates the best.

METHODS	$\gamma = 2$	$\gamma = 5$	$\gamma = 10$
SOURCE	56.41 ± 0.05	56.12 ± 0.07	55.77 ± 0.16
BN STATS	78.33 ± 0.05	71.75 ± 0.08	60.68 ± 0.14
TENT	68.85 ± 3.14	66.94 ± 3.52	56.18 ± 4.13
EATA	79.35 ± 0.16	69.23 ± 0.25	53.88 ± 0.53
LAME	78.96 ± 0.05	75.20 ± 0.10	68.16 ± 0.13
COTTA	**81.81 ± 0.37**	73.58 ± 0.28	60.58 ± 0.15
NOTE	78.81 ± 0.27	77.96 ± 0.75	77.18 ± 0.38
ODS	81.13 ± 0.09	**80.40 ± 0.36**	**80.67 ± 0.29**

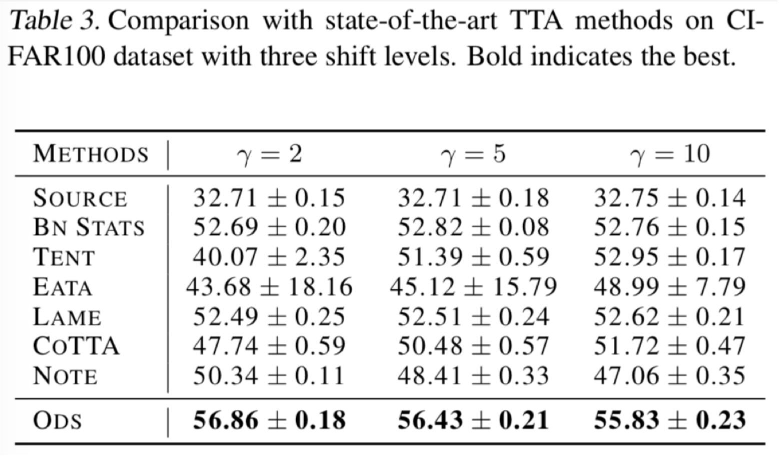
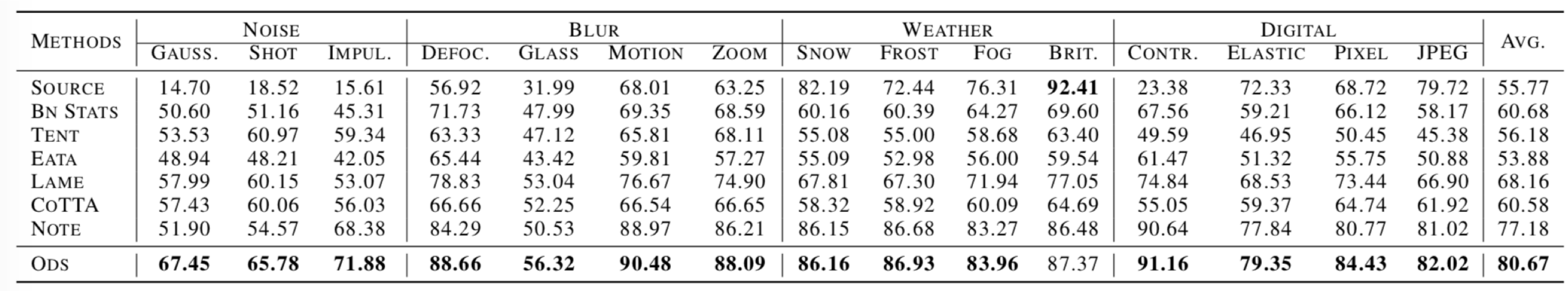


Table 3. Comparison with state-of-the-art TTA methods on CIFAR100 dataset with three shift levels. Bold indicates the best.

METHODS	$\gamma = 2$	$\gamma = 5$	$\gamma = 10$
SOURCE	32.71 ± 0.15	32.71 ± 0.18	32.75 ± 0.14
BN STATS	52.69 ± 0.20	52.82 ± 0.08	52.76 ± 0.15
TENT	40.07 ± 2.35	51.39 ± 0.59	52.95 ± 0.17
EATA	43.68 ± 18.16	45.12 ± 15.79	48.99 ± 7.79
LAME	52.49 ± 0.25	52.51 ± 0.24	52.62 ± 0.21
COTTA	47.74 ± 0.59	50.48 ± 0.57	51.72 ± 0.47
NOTE	50.34 ± 0.11	48.41 ± 0.33	47.06 ± 0.35
ODS	**56.86 ± 0.18**	**56.43 ± 0.21**	**55.83 ± 0.23**



METHODS	NOISE			BLUR				WEATHER				DIGITAL				AVG.
	GAUSS.	SHOT	IMPUL.	DEFOC.	GLASS	MOTION	ZOOM	SNOW	FROST	FOG	BRIT.	CONTR.	ELASTIC	PIXEL	JPEG	
SOURCE	14.70	18.52	15.61	56.92	31.99	68.01	63.25	82.19	72.44	76.31	**92.41**	23.38	72.33	68.72	79.72	55.77
BN STATS	50.60	51.16	45.31	71.73	47.99	69.35	68.59	60.16	60.39	64.27	69.60	67.56	59.21	66.12	58.17	60.68
TENT	53.53	60.97	59.34	63.33	47.12	65.81	68.11	55.08	55.00	58.68	63.40	49.59	46.95	50.45	45.38	56.18
EATA	48.94	48.21	42.05	65.44	43.42	59.81	57.27	55.09	52.98	56.00	59.54	61.47	51.32	55.75	50.88	53.88
LAME	57.99	60.15	53.07	78.83	53.04	76.67	74.90	67.81	67.30	71.94	77.05	74.84	68.53	73.44	66.90	68.16
COTTA	57.43	60.06	56.03	66.66	52.25	66.54	66.65	58.32	58.92	60.09	64.69	55.05	59.37	64.74	61.92	60.58
NOTE	51.90	54.57	68.38	84.29	50.53	88.97	86.21	86.15	86.68	83.27	86.48	90.64	77.84	80.77	81.02	77.18
ODS	**67.45**	**65.78**	**71.88**	**88.66**	**56.32**	**90.48**	**88.09**	**86.16**	**86.93**	**83.96**	87.37	**91.16**	**79.35**	**84.43**	**82.02**	**80.67**

ODS gives **the best results** on benchmark datasets in most cases. The detailed performance of different corruptions is also the same.

Experiments

- **RQ2:** Whether ODS is generic to integrate with different TTA methods and boost their performance?

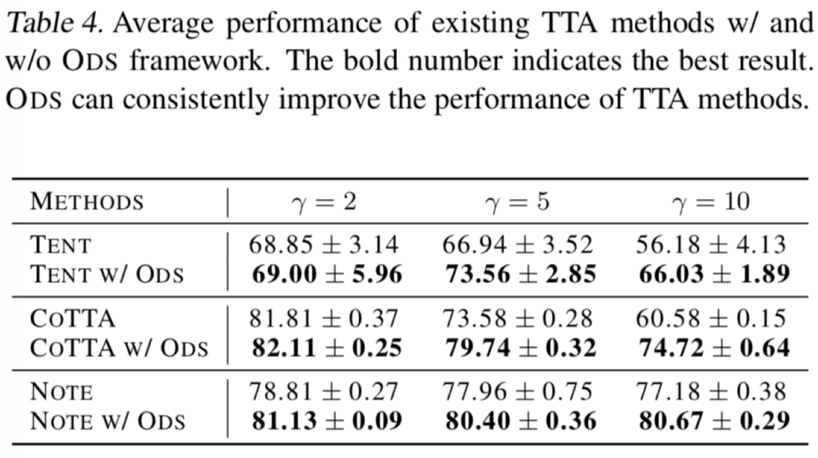
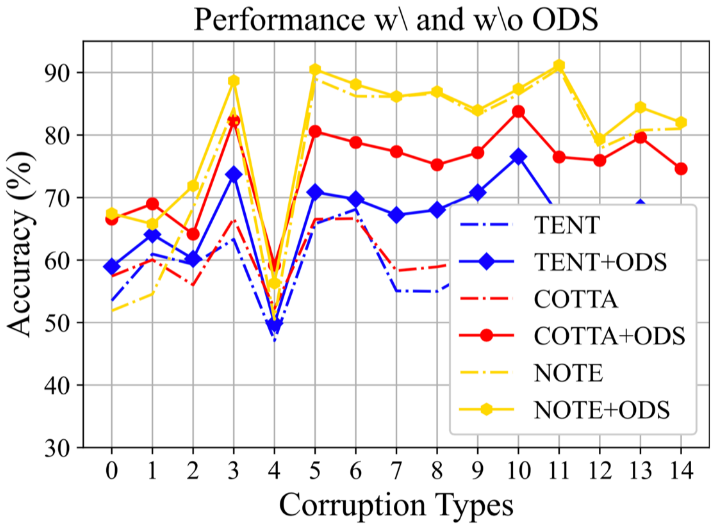


Table 4. Average performance of existing TTA methods w/ and w/o ODS framework. The bold number indicates the best result. ODS can consistently improve the performance of TTA methods.

METHODS	$\gamma = 2$	$\gamma = 5$	$\gamma = 10$
TENT	68.85 ± 3.14	66.94 ± 3.52	56.18 ± 4.13
TENT W/ ODS	$\mathbf{69.00 \pm 5.96}$	$\mathbf{73.56 \pm 2.85}$	$\mathbf{66.03 \pm 1.89}$
COTTA	81.81 ± 0.37	73.58 ± 0.28	60.58 ± 0.15
COTTA W/ ODS	$\mathbf{82.11 \pm 0.25}$	$\mathbf{79.74 \pm 0.32}$	$\mathbf{74.72 \pm 0.64}$
NOTE	78.81 ± 0.27	77.96 ± 0.75	77.18 ± 0.38
NOTE W/ ODS	$\mathbf{81.13 \pm 0.09}$	$\mathbf{80.40 \pm 0.36}$	$\mathbf{80.67 \pm 0.29}$

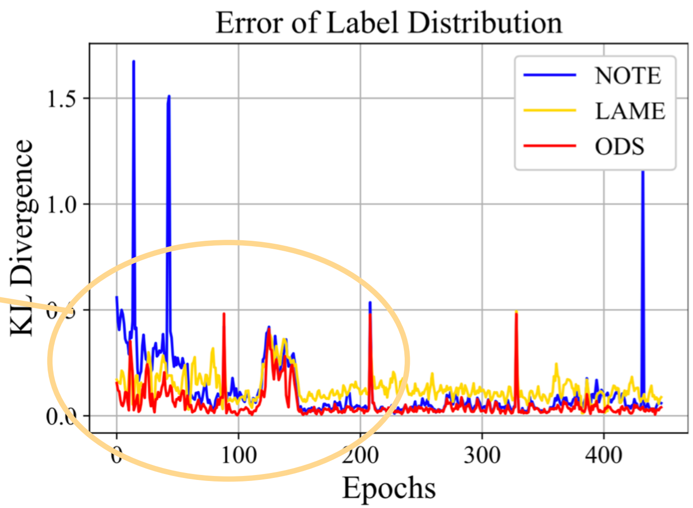


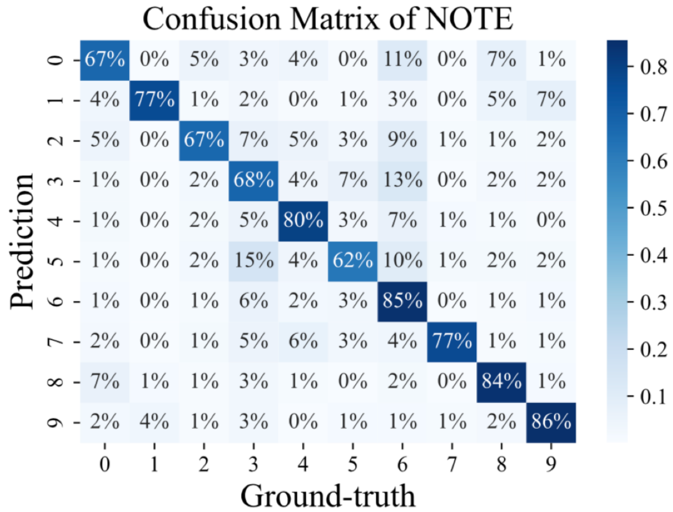
ODS gives **better overall results** on benchmark datasets with different levels of label distribution shifts. Detailed results on each corruption are similar.

Experiments

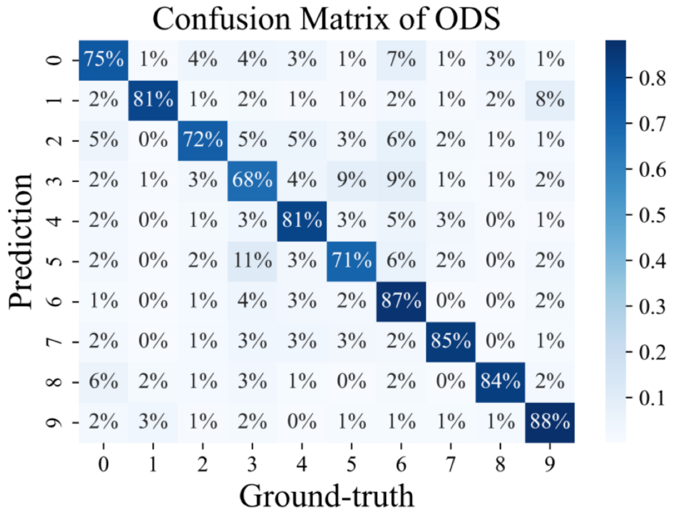
- **RQ3:** Does ODS accurately estimate label distribution and effectively optimize the prediction?

The KL Divergence of ODS is smaller than LAME and NOTE methods.


Error of Label Distribution


Confusion Matrix of NOTE

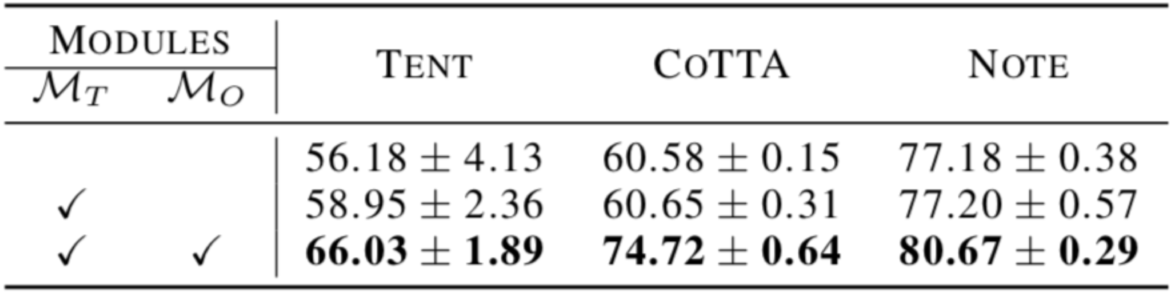
	0	1	2	3	4	5	6	7	8	9
0	67%	0%	5%	3%	4%	0%	11%	0%	7%	1%
1	4%	77%	1%	2%	0%	1%	3%	0%	5%	7%
2	5%	0%	67%	7%	5%	3%	9%	1%	1%	2%
3	1%	0%	2%	68%	4%	7%	13%	0%	2%	2%
4	1%	0%	2%	5%	80%	3%	7%	1%	1%	0%
5	1%	0%	2%	15%	4%	62%	10%	1%	2%	2%
6	1%	0%	1%	6%	2%	3%	85%	0%	1%	1%
7	2%	0%	1%	5%	6%	3%	4%	77%	1%	1%
8	7%	1%	1%	3%	1%	0%	2%	0%	84%	1%
9	2%	4%	1%	3%	0%	1%	1%	1%	2%	86%


Confusion Matrix of ODS

	0	1	2	3	4	5	6	7	8	9
0	75%	1%	4%	4%	3%	1%	7%	1%	3%	1%
1	2%	81%	1%	2%	1%	1%	2%	1%	2%	8%
2	5%	0%	72%	5%	5%	3%	6%	2%	1%	1%
3	2%	1%	3%	68%	4%	9%	9%	1%	1%	2%
4	2%	0%	1%	3%	81%	3%	5%	3%	0%	1%
5	2%	0%	2%	11%	3%	71%	6%	2%	0%	2%
6	1%	0%	1%	4%	3%	2%	87%	0%	0%	2%
7	2%	0%	1%	3%	3%	3%	2%	85%	0%	1%
8	6%	2%	1%	3%	1%	0%	2%	0%	84%	2%
9	2%	3%	1%	2%	0%	1%	1%	1%	1%	88%

Other Results

Ablation Study

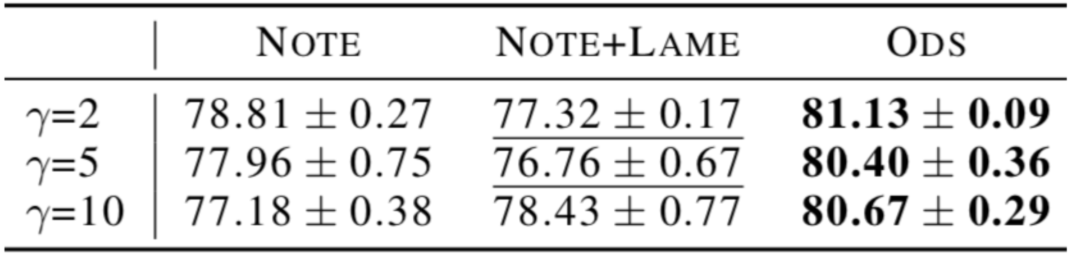


Modules $\mathcal{M}_T$	Modules $\mathcal{M}_O$	Tent	CoTTA	Note
		56.18 ± 4.13	60.58 ± 0.15	77.18 ± 0.38
✓		58.95 ± 2.36	60.65 ± 0.31	77.20 ± 0.57
✓	✓	**66.03 ± 1.89**	**74.72 ± 0.64**	**80.67 ± 0.29**

The two components proposed in ODS can only get the best results if they are integrated.

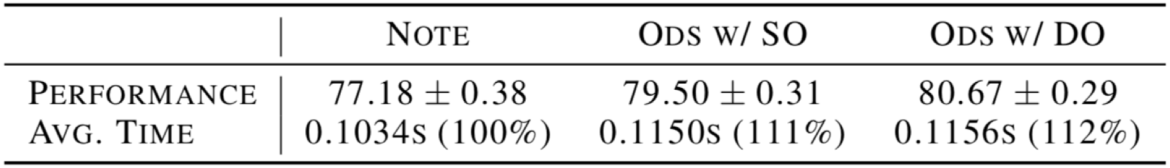
In-depth Comparison with LAME

ODS performs better than directly combining NOTE and LAME methods together.



	Note	Note+Lame	Ods
γ=2	78.81 ± 0.27	77.32 ± 0.17	**81.13 ± 0.09**
γ=5	77.96 ± 0.75	76.76 ± 0.67	**80.40 ± 0.36**
γ=10	77.18 ± 0.38	78.43 ± 0.77	**80.67 ± 0.29**

Time Consumption



	Note	Ods w/ SO	Ods w/ DO
Performance	77.18 ± 0.38	79.50 ± 0.31	80.67 ± 0.29
Avg. Time	0.1034s (100%)	0.1150s (111%)	0.1156s (112%)

ODS does not bring a large calculation burden to the existing TTA algorithm.

Outline

- Background
- ODS Framework
- Experiments
- **Conclusions**

Conclusions

In this paper, we consider a realistic setting, i.e.,

Open-World Data Shift setting for test-time adaptation

✓ A simple yet effective ODS test-time adaptation framework

✓ Extensive experiments demonstrate the effectiveness of ODS

Future work

➢ Ensure the safety of adaptation

➢ Test-time adaptation for large vision-language models

Code:

https://www.lamda.nju.edu.cn/code_ODS.ashx

Thank you!

If you are interested in, feel free to contact me:
Zhi Zhou (zhouz@lamda.nju.edu.cn)