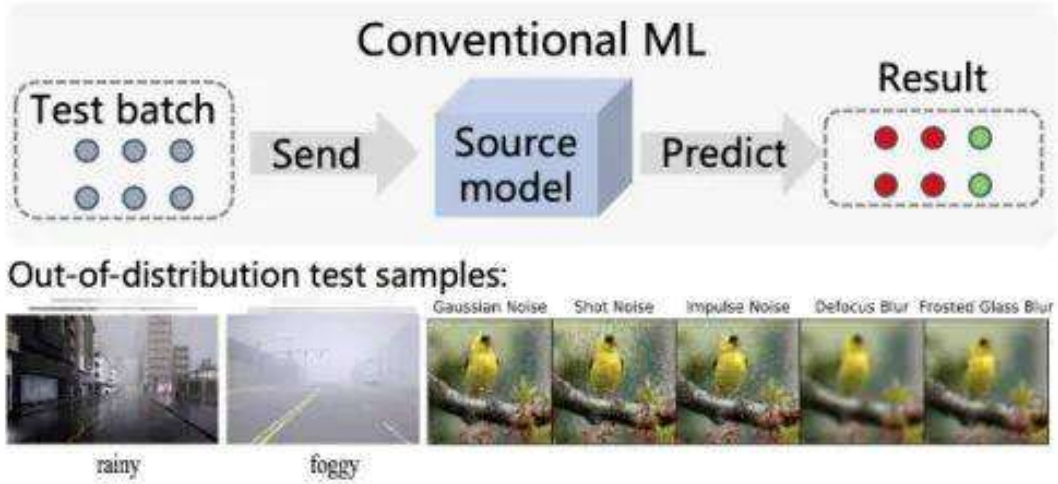
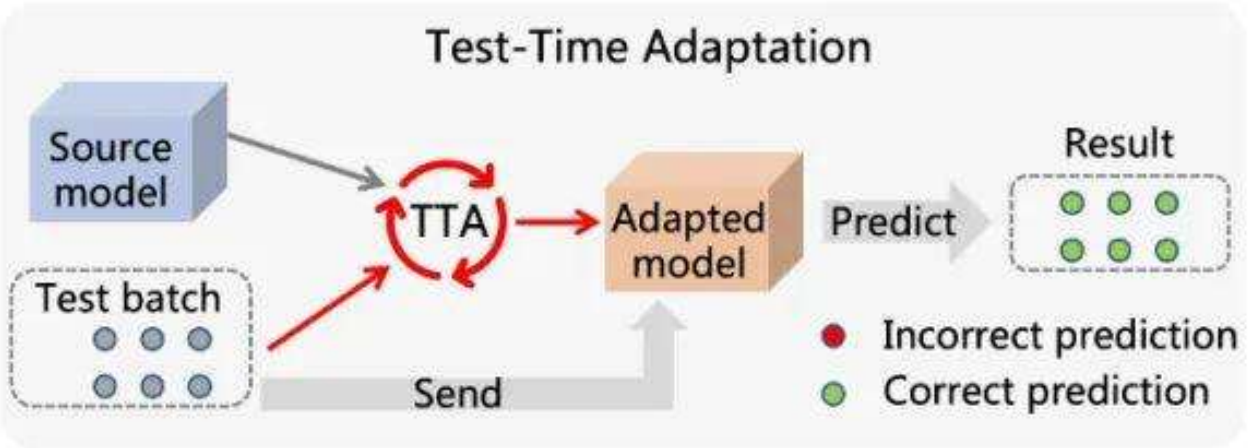
Adaptive Test-Time Personalization for Federated Learning

Wenxuan Bao[1*], **Tianxin Wei**[1*], **Haohan Wang**[1], **Jingrui He**[1],
[1]University of Illinois Urbana-Champaign
{wbao4,twei10,haohanw,jingrui}@illinois.edu

NeurIPS 2023

Test-Time Adaptation





Setting	Source data	Target data	Training loss	Testing loss	Offline	Online
Fine-tuning	×	x^t, y^t	$\mathcal{L}(x^t, y^t)$	--	√	×
Unsupervised domain adaptation	x^s, y^s	x^t	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^s, x^t)$	--	√	×
Test-time training [1]	x^s, y^s	x^t	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^s)$	$\mathcal{L}(x^t)$	×	√
Fully test-time adaptation [2]	×	x^t	×	$\mathcal{L}(x^t)$	×	√

Background

- 1) Federated Learning (FL) Core Challenge
- Clients exhibit distinct **distribution shifts**
- **Need:** Models must adapt to each client's unique data distribution for good performance

- 2) Limitations of Existing Methods

Personalized FL

- Relies on **labeled data** from test clients for personalization
- Infeasible for real-world scenarios

Test-Time Adaptation (TTA)

- Assumes training data from a **single source domain**
- Designed for **specific distribution shifts**
- Predefines adaptive modules — **ineffective for mismatched shifts**

Background

- 3) **Novel Setting: Test-Time Personalized FL (TTPFL)**
- **Training Phase:** Server trains a global model using source clients.
- **Test Phase:** Target clients adapt the global model locally with unlabeled data.
- **Goal:** Address complex distribution shifts for cross-device FL with unlabeled target clients.

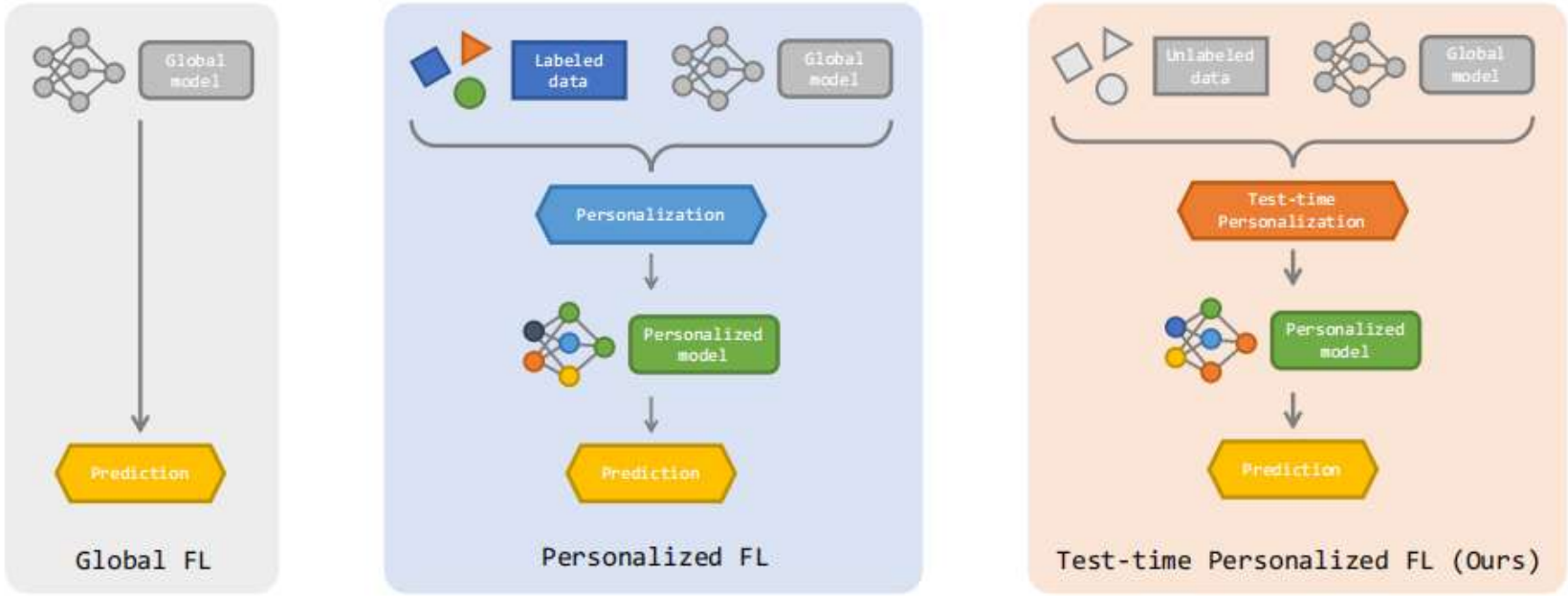


Figure 1: Comparison between the testing phase of GFL, PFL, and TTPFL. TTPFL enables model personalization without requiring labeled data.

Motivation

- **Core Conflict in TTPFL**
- Existing methods fail to handle diverse distribution shifts in FL with unlabeled test clients.

- **Gap 1:** TTA Ignores FL's Multi-Source Nature
- In FL, data comes from **multiple source clients** with inherent distribution shifts.
- Overlooking inter-source domain relationships degrades generalization to target clients.

- **Gap 2:** TTA **Lacks Flexibility** for Diverse Shifts
- TTA predefines adaptive modules
- **Performance Trade-off:** A module effective for one shift harms performance on another

- **Motivation for ATP**
- Adapts to **multiple distribution shifts**
- Learns optimal **adaptation rates** from FL's multi-source clients
- Enables effective unlabeled personalization in TTPFL

Motivation

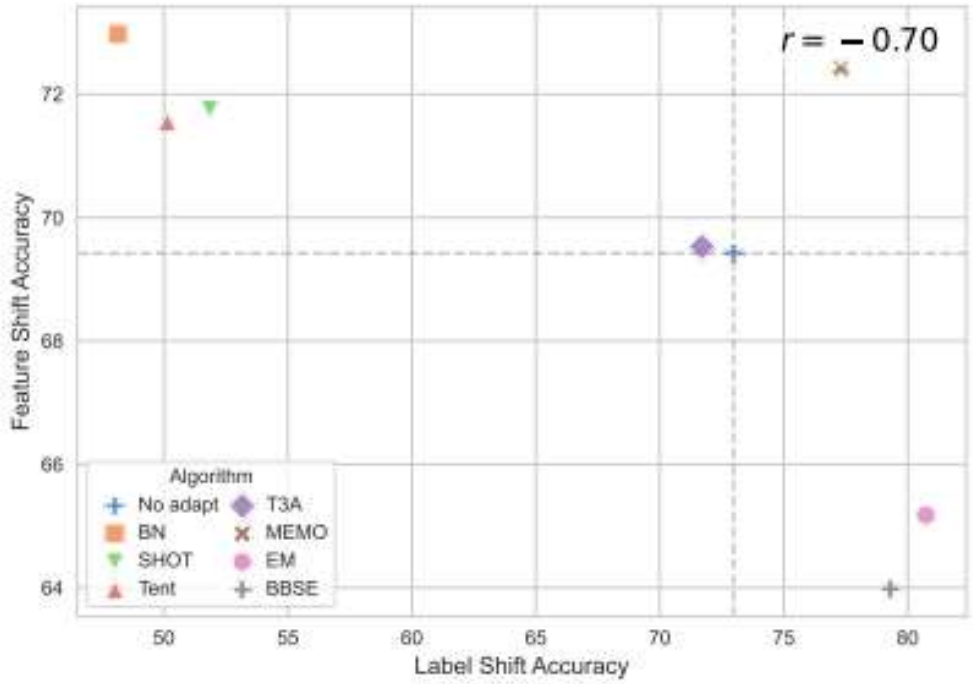


Figure 2: Performance trade-off of existing TTA methods under two distribution shifts.

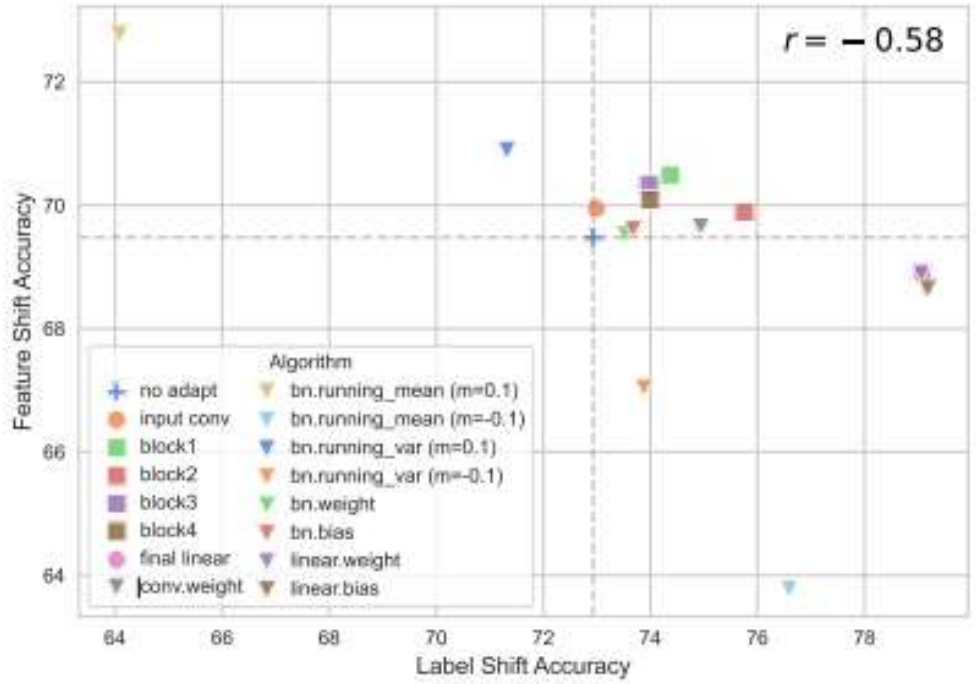


Figure 3: Performance trade-off of entropy minimization when adapting different modules.

Method——ATP: adaptive test-time personalization

- **Modules: Fine-grained model components — each has an adaptation rate $\alpha^{[l]}$**

- BN layers split into running mean/variance/weight/bias

- **Update Direction:** Unsupervised shift-aware direction for adaptation

Update trainable parameters

$$\boldsymbol{h}_k^{[l]} = -\nabla_{\boldsymbol{w}^{[l]}} \ell_H(f(\boldsymbol{X}_k; \boldsymbol{w}_G))$$

$$\ell_H(\hat{\boldsymbol{Y}}) = \frac{1}{B}\sum_{b=1}^{B}(-\sum_c \hat{y}_{b,c}\log \hat{y}_{b,c})$$

Update running statistics

$$\boldsymbol{h}_k^{[l]} = \hat{\boldsymbol{w}}_k^{[l]} - \boldsymbol{w}_G^{[l]}$$

$$\boldsymbol{w}_k^{[l]} \leftarrow \boldsymbol{w}_G^{[l]} + \alpha^{[l]}\boldsymbol{h}_k^{[l]}$$

- **Module Update:** $\boldsymbol{w}_k^{[l]} \leftarrow \boldsymbol{w}_G^{[l]} + \alpha^{[l]}\boldsymbol{h}_k^{[l]} \qquad \boldsymbol{w}_k \leftarrow \boldsymbol{w}_G + (\boldsymbol{A}\boldsymbol{\alpha}) \odot \boldsymbol{h}_k$

Method——ATP: adaptive test-time personalization

- **Training Phase**

Algorithm 1 ATP Training

`ServerTrain`($\boldsymbol{w}_G$, $\boldsymbol{\alpha}_G^0 = \mathbf{0}$)

Broadcast $\boldsymbol{w}_G$ to all source clients
for communication round $t = 1$ to T **do**
  $\mathbb{S}^t \leftarrow$ (random set of C source clients)
  for source client $\mathcal{S}_i \in \mathbb{S}^t$ **in parallel do**
    $\boldsymbol{\alpha}_i^t \leftarrow$ `ClientTrain`$(\mathcal{S}_i, \boldsymbol{\alpha}_G^{t-1})$
  $\boldsymbol{\alpha}_G^t = \frac{1}{C}\sum_{\mathcal{S}_i \in \mathbb{S}^t} \boldsymbol{\alpha}_i^t$
return $\boldsymbol{\alpha}_G^T$

`ClientTrain`($\mathcal{S}_i, \boldsymbol{\alpha}$) *# Run on source client $\mathcal{S}_i$*
for local epoch $e = 1$ to E **do**
  $\mathbb{B}^{\mathcal{S}_i} \leftarrow$ (split $\mathbb{D}^{\mathcal{S}_i}$ into $K^{\mathcal{S}_i}$ batches of size B)
  for batch $k = 1$ to $K^{\mathcal{S}_i}$ **do**
    $(\boldsymbol{X}_k^{\mathcal{S}_i}, \boldsymbol{Y}_k^{\mathcal{S}_i}) \leftarrow$ (k-th labeled batch in $\mathbb{B}^{\mathcal{S}_i}$)
    Estimate update direction $\boldsymbol{h}_k^{\mathcal{S}_i}$ with *unlabeled* $\boldsymbol{X}_k^{\mathcal{S}_i}$ according to Eq. (4) and (5)
    $\boldsymbol{w}_k^{\mathcal{S}_i} \leftarrow \boldsymbol{w}_G + (\boldsymbol{A}\boldsymbol{\alpha}) \odot \boldsymbol{h}_k^{\mathcal{S}_i}$
    $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \nabla_{\boldsymbol{\alpha}} \ell_{CE}(f(\boldsymbol{X}_j^{\mathcal{S}_i}; \boldsymbol{w}_k^{\mathcal{S}_i}), \boldsymbol{Y}_k^{\mathcal{S}_i})$
return $\boldsymbol{\alpha}$

- **Testing phase**

ATP-batch for test-time batch adaptation/
ATP-online for online test-time adaptation

Algorithm 2 ATP Testing

`ClientTest`($\mathcal{T}_j, \boldsymbol{w}_G, \boldsymbol{\alpha}$) *# Run on target client $\mathcal{T}_j$*
$\mathbb{B}^{\mathcal{T}_j} \leftarrow$ (split $\mathbb{X}^{\mathcal{T}_j}$ into $K^{\mathcal{T}_j}$ batches of size B)
$\boldsymbol{h}_{\text{history}} \leftarrow \mathbf{0}$ *# Cumulative moving average*
for batch $k = 1$ to $K^{\mathcal{T}_j}$ **do**
  Estimate update direction $\boldsymbol{h}_k^{\mathcal{T}_j}$ with *unlabeled* $\boldsymbol{X}_k^{\mathcal{T}_j}$ according to Eq. (4) and (5)
  if *TTBA* **then**
    $\boldsymbol{w}_k^{\mathcal{T}_j} \leftarrow \boldsymbol{w}_G + (\boldsymbol{A}\boldsymbol{\alpha}) \odot \boldsymbol{h}_k^{\mathcal{T}_j}$
  else if *OTTA* **then**
    $\boldsymbol{h}_{\text{history}} \leftarrow \frac{k-1}{k}\boldsymbol{h}_{\text{history}} + \frac{1}{k}\boldsymbol{h}_k^{\mathcal{T}_j}$
    $\boldsymbol{w}_k^{\mathcal{T}_j} \leftarrow \boldsymbol{w}_G + (\boldsymbol{A}\boldsymbol{\alpha}) \odot \boldsymbol{h}_{\text{history}}$
  Make prediction: $\hat{\boldsymbol{Y}}_k^{\mathcal{T}_j} = f(\boldsymbol{X}_k^{\mathcal{T}_j}; \boldsymbol{w}_k^{\mathcal{T}_j})$

Communication Cost: $2TD \rightarrow D + 2Td$

Experiments

RQ1: Can ATP handle different distribution shift?

Table 1: Accuracy (mean ± s.d. %) on target clients under various distribution shifts on CIFAR-10

Method	Feature shift	Label shift	Hybrid shift	Avg. Rank
No adaptation	69.42 ± 0.13	72.98 ± 0.24	63.68 ± 0.24	7.7
BN-Adapt	73.52 ± 0.22	54.54 ± 0.10	50.42 ± 0.39	7.0
SHOT	71.76 ± 0.17	48.13 ± 0.18	44.68 ± 0.32	9.3
Tent	71.76 ± 0.09	50.13 ± 0.21	46.05 ± 0.26	8.3
T3A	69.53 ± 0.08	71.70 ± 0.32	62.17 ± 0.17	8.0
MEMO	72.43 ± 0.22	77.30 ± 0.15	68.07 ± 0.28	4.3
EM	65.18 ± 0.12	80.73 ± 0.18	69.85 ± 0.43	5.0
BBSE	63.98 ± 0.17	79.30 ± 0.17	67.96 ± 0.43	6.7
Surgical	69.85 ± 0.22	76.00 ± 0.17	66.94 ± 0.43	6.3
ATP-batch	73.68 ± 0.10	79.90 ± 0.22	73.05 ± 0.35	2.3
ATP-online	**74.06 ± 0.18**	**81.96 ± 0.14**	**75.37 ± 0.22**	**1.0**

Table 2: Accuracy (mean ± s.d. %) on target clients under hybrid shift on Digits-5 and PACS

Method	Digits-5					PACS			
	MNIST	SVHN	USPS	SynthDigits	MNIST-M	Art	Cartoon	Photo	Sketch
No adaptation	95.47 ± 0.22	52.28 ± 1.45	89.62 ± 0.44	79.75 ± 0.69	55.62 ± 0.80	71.57 ± 1.16	74.71 ± 0.70	90.25 ± 0.75	74.20 ± 0.72
BN-Adapt	94.90 ± 0.29	57.57 ± 0.53	89.51 ± 0.39	75.34 ± 0.48	59.68 ± 0.44	73.55 ± 0.51	71.54 ± 0.55	92.07 ± 0.26	70.92 ± 0.53
SHOT	94.69 ± 0.31	57.91 ± 0.23	89.55 ± 0.69	76.43 ± 0.34	60.19 ± 0.69	69.32 ± 0.67	67.77 ± 0.40	86.97 ± 0.60	59.40 ± 0.91
Tent	95.48 ± 0.29	60.67 ± 0.49	91.65 ± 0.61	78.56 ± 0.45	62.49 ± 0.73	71.59 ± 0.71	71.03 ± 0.97	88.06 ± 0.24	63.15 ± 1.10
T3A	94.63 ± 0.61	49.90 ± 1.10	88.46 ± 0.75	75.47 ± 1.14	51.25 ± 1.55	72.15 ± 0.72	75.02 ± 0.78	91.51 ± 0.62	70.14 ± 1.21
MEMO	95.92 ± 0.19	52.85 ± 1.09	89.84 ± 0.44	80.12 ± 0.90	55.48 ± 1.13	71.47 ± 1.29	75.57 ± 0.98	90.65 ± 0.90	76.30 ± 0.65
EM	96.64 ± 0.31	57.21 ± 1.65	92.29 ± 0.32	85.69 ± 0.46	62.08 ± 0.60	73.96 ± 1.85	78.91 ± 0.92	92.30 ± 0.92	80.82 ± 1.52
BBSE	94.47 ± 0.58	57.26 ± 1.47	91.34 ± 0.39	85.54 ± 0.46	61.59 ± 0.91	74.33 ± 1.78	78.69 ± 1.00	91.82 ± 0.68	80.15 ± 1.42
Surgical	97.35 ± 0.13	59.93 ± 2.01	94.19 ± 0.40	86.06 ± 0.44	65.87 ± 0.78	74.59 ± 2.69	77.48 ± 0.64	92.34 ± 0.78	80.90 ± 3.42
ATP-batch	97.81 ± 0.27	62.18 ± 1.71	95.41 ± 0.26	87.91 ± 0.45	69.98 ± 1.96	82.92 ± 0.96	**79.64 ± 0.75**	95.40 ± 0.41	82.28 ± 1.57
ATP-online	**97.81 ± 0.23**	**62.64 ± 1.92**	**95.56 ± 0.23**	**88.33 ± 0.47**	**70.78 ± 2.36**	**83.51 ± 0.84**	79.46 ± 0.77	**95.52 ± 0.40**	**82.80 ± 1.69**

Experiments

Table 5: ATP with different model architectures, accuracy (mean ± s.d. %) on target clients

Method	Shallow CNN on CIFAR-10				ResNet-50 on CIFAR-100			
	Feature shift	Label shift	Hybrid shift	Avg. Rank	Feature shift	Label shift	Hybrid shift	Avg. Rank
No adaptation	64.39 ± 0.18	69.33 ± 0.37	61.99 ± 0.47	7.3	45.31 ± 0.30	51.63 ± 0.15	40.01 ± 0.17	7.3
BN-Adapt	66.46 ± 0.22	54.99 ± 0.38	50.40 ± 0.43	7.0	47.75 ± 0.29	34.85 ± 0.26	30.31 ± 0.09	7.3
SHOT	65.60 ± 0.18	49.98 ± 0.29	45.95 ± 0.47	9.0	45.42 ± 0.30	31.06 ± 0.32	27.44 ± 0.14	9.3
Tent	65.61 ± 0.24	50.12 ± 0.25	45.91 ± 0.49	8.7	45.91 ± 0.46	31.34 ± 0.11	27.93 ± 0.31	8.3
T3A	64.31 ± 0.27	66.96 ± 0.43	59.65 ± 0.58	8.3	45.31 ± 0.30	51.42 ± 0.15	39.89 ± 0.20	7.7
MEMO	65.89 ± 0.31	71.95 ± 0.25	64.17 ± 0.47	5.3	48.42 ± 0.14	55.19 ± 0.28	42.53 ± 0.20	3.7
EM	61.74 ± 0.25	76.28 ± 0.29	67.54 ± 0.41	5.0	43.00 ± 0.31	59.34 ± 0.15	44.82 ± 0.27	5.0
BBSE	56.92 ± 0.53	75.99 ± 0.44	66.64 ± 0.53	6.3	37.26 ± 0.64	56.97 ± 0.20	40.09 ± 0.51	7.0
Surgical	64.45 ± 0.12	73.75 ± 0.42	65.67 ± 0.44	5.7	45.18 ± 0.38	54.83 ± 0.26	42.50 ± 0.33	6.7
ATP-batch	66.90 ± 0.05	76.23 ± 0.32	68.88 ± 0.35	2.3	48.35 ± 0.45	58.06 ± 0.53	46.82 ± 0.32	2.7
ATP-online	**67.13 ± 0.17**	**78.56 ± 0.32**	**71.52 ± 0.51**	**1.0**	**49.08 ± 0.26**	**61.86 ± 0.25**	**49.51 ± 0.23**	**1.0**

Experiments

RQ2: Does ATP learn adaptation rates specific to distribution shift?

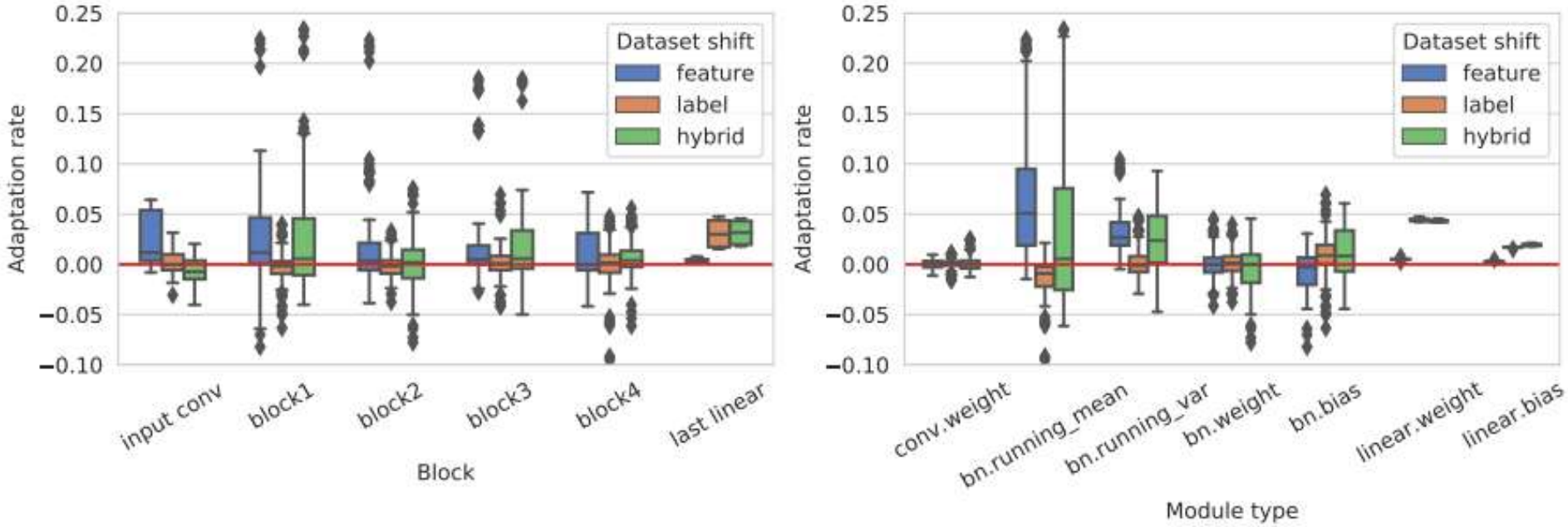


Figure 4: Adaptation rates learned by ATP with different distribution shifts on CIFAR-10

Experiments

RQ2: Does ATP learn adaptation rates specific to distribution shift?

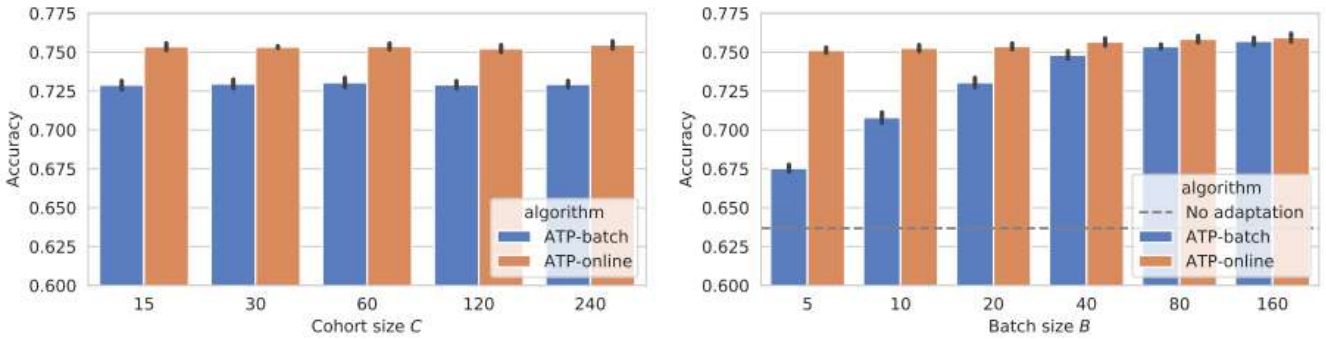


Figure 5: Effect of cohort size and batch size

Table 3: Train and test adaptation rates with different distribution shifts, accuracy (mean ± s.d. %)

Train	Test		
	Feature shift	Label shift	Hybrid shift
No adaptation	69.42 ± 0.13	72.98 ± 0.24	63.68 ± 0.24
Feature shift	**73.68 ± 0.10**	65.05 ± 1.82	60.64 ± 1.43
Label shift	67.99 ± 0.28	**79.90 ± 0.22**	69.50 ± 0.52
Hybrid shift	72.69 ± 0.14	78.92 ± 0.34	**73.05 ± 0.35**

Table 7: Accuracy (%), ATP enhances different global models under hybrid shift on CIFAR-10

Method	FedAvg	FedProx ($\mu = 0.01$)	q-FFL ($q = 1$)
No adaptation	63.68 ± 0.24	63.77 ± 0.25	63.87 ± 0.23
ATP-batch	73.05 ± 0.35	72.95 ± 0.33	73.15 ± 0.21
ATP-online	75.37 ± 0.22	75.51 ± 0.19	75.79 ± 0.15

Thanks