Multi-Task Test-time Adaptation via Gradient Consensus and Plasticity Constraint

Zhong Ye[1], **Yu Hu**[1*], **Zhenguo Yang**[1]

[1]School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China

Abstract

Multi-task test-time adaptation (MT-TTA) aims to adapt pre-trained models to dynamic environments during multi-task inference by leveraging unlabeled test data. This task is particularly challenging as different tasks respond divergently to distribution shifts, and mixed input streams containing both in-distribution (ID) and out-of-distribution (OOD) samples make the models after test-time adaptation prone to catastrophic forgetting of ID knowledge. Although the existing methods like M-TENT extend the classic test entropy minimization (TENT) by minimizing multi-task entropies and employing task-average gradient to adapt a model, it suffers from two key limitations: 1) the average gradient strategy proposed by M-TENT may exacerbate multi-task test-time optimization conflicts, harming individual tasks when gradients are directionally non-consensual; 2) aggressive updates on mixed ID/OOD data cause severe forgetting of ID knowledge. In this paper, we theoretically establish a formal connection between multi-task loss differences and test-time performance under the first-order Taylor analysis, demonstrating that consensual multi-task entropy reductions are likely to increase the performance, while non-consensual ones might decrease the performance. To this end, we propose Consensus-driven Constrained Multi-Task Test-Time Adaptation (CoCo-MT-TTA), consisting of 1) multi-task gradient consensus adaptation, which aligns cross-task gradient directions to seek a consensus gradient; 2) multi-task plasticity-constraint adaptation, which constrains parameter updates using second-moment statistics to preserve ID knowledge. Extensive experiments on benchmark datasets, including CelebA and PlantData, demonstrate that our method achieves an absolute improvement of up to 16.02% in mean ID/OOD F1-score (Mean I&O) under domain shifts over non-adapted models, outperforming the recent baselines.

Code — https://github.com/leafheavy/MT-TTA

1 Introduction

Multi-task pre-trained models have been widely deployed on edge computing devices, such as enabling simultaneous execution of various tasks, including face recognition, scenery classification, background segmentation, and other tasks simultaneously (Ranjan, Patel, and Chellappa 2017; Chen et al. 2025). However, in inference stages, the performance of these models can significantly degrade due to distributional shifts, such as illumination changes or sensor noise. This presents a critical obstacle to the reliability of multi-task models in real-world applications.

Test-time adaptation (TTA) methods aim to adjust models in response to distributional shifts under unlabeled and resource-limited environments for a single task. For instance, test entropy minimization (TENT) (Wang et al. 2021)

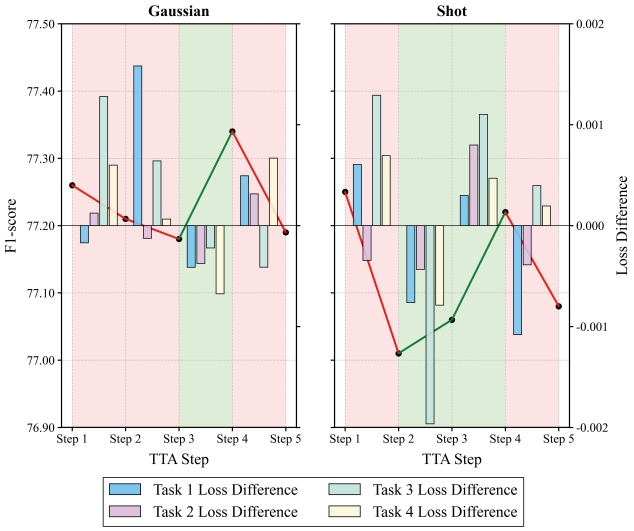

Figure 1: Experiments on CelebA with domain shifts of Gaussian and Shot noise. The X-axis denotes the test-time adaptation steps of performing M-TENT (Chatterjee et al. 2024) on tested samples. The figure shows a relationship between the relative change of task-specific loss (entropy) and F1 score on distribution shifts, i.e., consensual multi-task entropy reductions are likely to increase the performance while non-consensual ones might decrease the performance. The broken lines represent the model performance, with red indicating a decline in performance and green indicating an increase. The red in the background indicates that there is a directional conflict between the multi-task loss differences as shown in the bar. (best viewed with color)

*Yu Hu is the corresponding author

adjusts a pre-trained model by optimizing the entropy of test samples during adaptation for image classification. In contrast, multi-task test-time adaptation (MT-TTA) can leverage the shared feature space across tasks to extract more generalizable knowledge representations. This shared inductive bias improves the model's robustness under previously unseen shifts. Moreover, overfitting during adaptation can cause models to forget robust features learned during pre-training, leading to unpredictable behavior. For example, in autonomous driving, adapting a model to road scenes under thunderstorm conditions might cause it to forget how to recognize road conditions in sunny weather. Therefore, it is essential to retain knowledge from the source (in-distribution, ID) domain to ensure stable and controllable behavior under continuous, open-world shifts.

To this end, we aim to tackle the problem of MT-TTA, which involves concurrently adapting to multiple out-of-distribution (OOD) target domains while preserving knowledge from the source domain, all under the constraints of unlabeled data and limited computational resources. Recently, M-TENT (Chatterjee et al. 2024) extends TENT (Wang et al. 2021) to the multi-task setting and minimizes the average of overall task entropies, which faces two major challenges in MT-TTA: (1) Different tasks exhibit varying sensitivities to domain shifts, which are known as task heterogeneity. The average strategy proposed by M-TENT might cause non-consensual update directions and thereby result in performance degradation, as theoretically analyzed in Section 3.2. (2) Mixed distributional shifts in the input stream under the open-world environment and the high sensitivity of pre-trained models to batch-specific noise. When deployed, the input often consists of a mixture of ID and OOD data, and aggressive parameter updates during adaptation can lead to severe forgetting of ID-domain knowledge (Niu et al. 2022).

Our motivation stems from empirical observations that the information entropy difference for m-th task across consecutive time steps, denoted as $\Delta L_{(m)}$, strongly correlates with the performance of the adapted model, as shown in Figure 1. We find that non-consensual $\Delta L_{(m)}$ directions across tasks may harm overall performance, while a consensus of these directions tends to improve it. Furthermore, we observe that catastrophic forgetting frequently occurs during MT-TTA as shown in Figure 3.

To address these issues, we propose Consensus-driven Constrained Multi-Task Test-Time Adaptation (CoCo-MT-TTA) to coordinate the parameter update directions across M tasks based on the consensus of their loss difference $\Delta L_{(m)}$, $\forall m \in [1, M]$, and simultaneously constrain the magnitude of parameters updates to retain knowledge of the source domain. Specifically, we propose CoCo-MT-TTA, which incorporates two strategies: 1) Multi-task Gradient Consensus Adaptation. It searches for a consensual update direction, which alleviates the optimization conflict caused by non-consensual multi-task update directions. 2) Multi-task Plasticity-Constrained Adaptation. It restrains the update magnitude during adaptation based on the parameter importance scores, which maintains the memory of ID domain knowledge.

The main contributions can be summarized as follows:

- To our knowledge, we are the first to identify and empirically validate that orientation-consensual gradients across tasks are beneficial for multi-task test-time domain adaptation. (Section 3.2)
- Based on this insight, we propose a multi-task gradient consensus method to enforce a consensual parameter updating orientation during test-time. (Section 4.1)
- We further introduce a multi-task plasticity-constrained method to alleviate catastrophic forgetting and stabilize adaptation. (Section 4.2)
- Extensive experiments on two diverse benchmarks, CelebA and PlantData, demonstrate the effectiveness and generalizability of our method. (Section 5)

2 Related Work

Multi-Task Learning. Multi-task learning (MTL) aims to learn various related tasks (Ruder 2017). MTL encompasses several classic research directions: model architecture design and optimization. The survey (Ruder 2017) typically categorizes the model architecture into soft parameter sharing and hard parameter sharing. Soft parameter sharing assigns separate backbones to tasks (Misra et al. 2016). In contrast, hard parameter sharing, dating back to the 1990s (Bromley et al. 1993), uses a common backbone with task-specific output heads. From the optimization standpoint, recent works tackle task conflicts through adversarial training (Liu, Qiu, and Huang 2017), scalarization (Kendall, Gal, and Cipolla 2018), and multi-objective optimization (Sener and Koltun 2018). To the best of our knowledge, M-TENT (Chatterjee et al. 2024) is the only recent work that addresses multi-task test-time adaptation (MT-TTA). It extends test entropy minimization (TENT) (Wang et al. 2021) to the multi-task setting by minimizing the mean entropy of all task outputs. However, M-TENT (Chatterjee et al. 2024) overlooks the heterogeneity across tasks, which hinders adaptation performance.

Test-Time Adaptation. Test-time adaptation (TTA) aims to improve model performance on out-of-distribution (OOD) data by adapting models using label-free test samples. TENT (Wang et al. 2021) adapts models by minimizing inference prediction entropy via updating the batch normalization statistics. Efficient Anti-forgetting Test-time Adaptation (EATA) (Niu et al. 2022) introduces a sample entropy selection mechanism to enhance the efficiency of adaptation. Adversarial training on Penultimate Activations (APA) (Sun, Lu, and Ling 2023) perturbs the penultimate layer features and minimizes the divergence between the original and perturbed predictions. Activation Matching (ActMAD) (Mirza et al. 2023) aligns inference local statistics with reference statistics from the source domain. Sharpness-aware and reliable entropy minimization (SAR) (Niu et al. 2023) filters out high-entropy samples and employs Sharpness-Aware Minimization (SAM) (Foret et al. 2021) to enhance generalization.

Continuous Learning and Anti-Forgetting. McCloskey and Cohen first elaborates the problem of catastrophic forgetting. Subsequent research has focused on mitigating for-

getting, which has also been extended to the TTA area. EATA (Niu et al. 2022) uses Fisher information estimated from source data to regularize the parameter updates during test-time adaptation, thereby preserving performance on the source domain. However, its effectiveness is highly sensitive to hyperparameter tuning. Continual test-time adaptation (CoTTA) (Wang et al. 2022) mitigates forgetting by periodically restoring a subset of model neurons to the initial values with a fixed probability, but this stochastic restoration can interfere with the acquisition of new knowledge. Collaborative Lifelong Adaptation (CoLA) (Chen et al. 2024) stores and dynamically aggregates parameter deltas and domain features during adaptation.

3 Problem Formulation and Motivation

3.1 Problem Formulation.

Following the recent study on multi-task test-time adaptation (MT-TTA), we consider the multi-task models typically consisting of a shared encoder followed by the task-specific classifiers (Chatterjee et al. 2024). Without loss of generality, let us denote the source domain data as $\{(\mathbf{x}_{n_s}, y_{n_s})\}_{n_s=1}^{N_S} \sim P(\mathbf{x})$ and the model as $f(\cdot)$, where $y_{n_s} = [y_{(1),n_s}, y_{(2),n_s}, ..., y_{(M),n_s}]$ corresponds to the labels for M tasks. The model parameters on the source domain are denoted by Θ^0, which aggregates the parameters of size N_p. The pre-trained model $f(\Theta^0)$ performs well on the in-distribution (ID) data. However, once deployed in open-world edge scenarios (e.g., mobile or wearable devices), the model may encounter distribution shifts denoted as $Q(\mathbf{x})$, where $Q(\mathbf{x}) \neq P(\mathbf{x})$, leading to significant performance degradation.

Multi-task test-time adaptation (MT-TTA) refers to the setting where, under target domain shifts, the model adapts to multiple out-of-distribution (OOD) domains concurrently without access to labeled data, while retaining knowledge of the source domain. Specifically, given test samples $\{\mathbf{x}_{n_t}\}_{n_t=1}^{N_T}$ with $\mathbf{x}_{n_t} \sim Q(\mathbf{x})$, we aim to enhance the model's performance via adaptation by minimizing average unsupervised losses over all tasks:

$$\min_{\Theta_l} \mathcal{L}_{\text{ave}}(\Theta) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_{(m)}(\Theta), \tag{1}$$

where a typical choice of $\mathcal{L}_{(m)}$, $\forall m$, is the Shannon entropy (Chatterjee et al. 2024), and $\Theta_l \subseteq \Theta$ represents the subset of parameters that are learnable during test time.

3.2 Motivation.

Denoting the parameters after the t-th adaptation step by Θ^t, standard stochastic gradient descent (Rumelhart, Hinton, and Williams 1986) gives:

$$\Theta^t = \Theta^{t-1} - \eta \cdot d, \tag{2}$$

where η is the learning rate and $d = \nabla\mathcal{L}(\Theta^{t-1})$ is **the update direction**. For brevity, we write $\mathcal{L}_{(m)}(\Theta^t)$ as $\mathcal{L}^t_{(m)}$. Considering the tiny value of learning rate η, performing first-order Taylor expansion for task-specific loss $\mathcal{L}^{t-1}_{(m)}$:

$$\mathcal{L}^t_{(m)} \approx \mathcal{L}^{t-1}_{(m)} + \nabla\mathcal{L}^{t-1\top}_{(m)} \cdot (\Theta^t - \Theta^{t-1}). \tag{3}$$

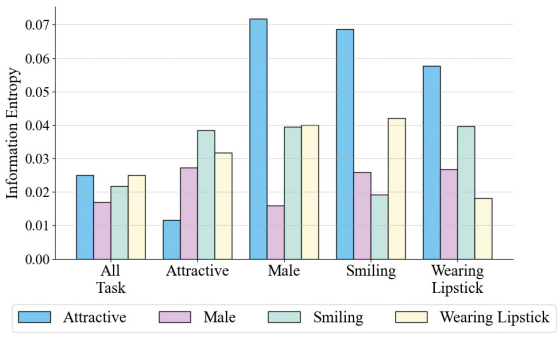

Figure 2: The x-axis represents the task-specific update gradient during adaptation. A single-task gradient reduces the loss of the corresponding task but increases that of others, indicating optimization conflicts in multi-task adaptation.

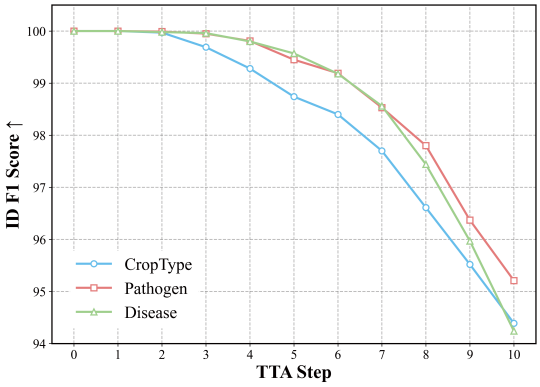

Figure 3: The dashed line represents the model's performance trajectory, which exhibits a consensual decline during adaptation across all three evaluated tasks on PlantData.

Substituting Eq. (2) into Eq. (3), we arrive at:

$$\mathcal{L}^t_{(m)} - \mathcal{L}^{t-1}_{(m)} \overset{\text{def}}{=} \Delta\mathcal{L}^t_{(m)} \approx -\eta \cdot \nabla\mathcal{L}^{t-1\top}_{(m)} \cdot d, \tag{4}$$

where $\Delta\mathcal{L}^t_{(m)}$ represents the loss difference between two successive adaptation steps.

As illustrated in Figure 1, our experiments support the correlation between the loss difference and the adaptation performance. In brief, **a non-consensual direction of the task-specific loss-difference**, $\exists m$, $\Delta\mathcal{L}_{(m)} > 0$, across consecutive adaptation steps **precipitates a degradation in model performance** (e.g., F1-score); conversely, **a consensual decrease yields performance gains**, i.e., $\Delta\mathcal{L}_{(m)} < 0$, $\forall m$.

For M-TENT (Chatterjee et al. 2024), which averages the losses over all tasks, we have update direction $d = d_{\text{M-TENT}} = \nabla\mathcal{L}^{t-1} = \frac{1}{M}\sum_{j=1}^{M} \nabla\mathcal{L}^{t-1}_{(j)}$. Thus, the entropy value change of m-th task is:

$$\Delta\mathcal{L}^t_{(m)} \approx -\eta \cdot \nabla\mathcal{L}^{t-1\top}_{(m)} \left(\frac{1}{M} \sum_{j=1}^{M} \nabla\mathcal{L}^{t-1}_{(j)} \right). \tag{5}$$

If $\exists j, \nabla\mathcal{L}_{(m)}^{t-1} \cdot \nabla\mathcal{L}_{(j)}^{t-1} < 0$, there is a possibility that $\Delta\mathcal{L}_{(m)}^{t} > 0$. In practice, the non-consensual gradient phenomenon does exist, as demonstrated in Figure 2. When parameters are updated solely with the gradient from one task, rather than the aggregated gradient over all tasks, the loss for the selected task decreases while the losses of all remaining tasks simultaneously increase, which also known as optimization conflict. Therefore, **reducing inter-task gradient conflict is essential for effective multi-task test-time adaptation (MT-TTA)**.

Ideally, if all task gradients have consensual directions, i.e., $\forall j \in [1, M], \nabla\mathcal{L}_{(m)}^{t-1} \cdot \nabla\mathcal{L}_{(j)}^{t-1} > 0$, it can be derived that the loss difference of m-th task $\Delta\mathcal{L}_{(m)}^{t} < 0, \forall t$, which motivates us to develop a gradient consensus method in Section 4.1.

Furthermore, in practice, test-time data probably consists of both ID and OOD samples. Adapting the model solely based on OOD inputs may lead to catastrophic forgetting of source domain, i.e., ID, knowledge, as empirically shown in Figure 3, which shows the updated multi-task model has a consistently lower F1-score on ID test samples than the pre-trained and un-adapted model.

4 Methodology

We propose Consensus-driven Constrained Multi-Task Test-Time Adaptation (CoCo-MT-TTA) that addresses the dual challenge of adapting to out-of distribution (OOD) domains while preserving performance on the source (in-distribution, ID) domain. As shown in Figure 4, our approach incorporates two strategies: (1) **Multi-task Gradient Consensus Test-time Adaptation** search the appropriate update gradient d to improve the overall adaptation performance, and (2) **Multi-task Plasticity-Constrained Adaptation** to mitigate catastrophic forgetting by constraining the change magnitude of model parameters Θ. The pseudo-code of CoCo-MT-TTA is summarized in Algorithm 1.

4.1 Multi-task Gradient Consensus Test-time Adaptation

Given the heterogeneity of task-specific gradients and our insight about the correlation of loss difference $\Delta\mathcal{L}$, update direction d, and adapted model performance, we **aim to find a d that leads to $\Delta\mathcal{L}_{(m)}^{t} < 0$ for all $m \in [1, M]$**. First, based on Eq. (4), the following mathematical formula can be obtained:

$$-\frac{1}{\eta}\Delta\mathcal{L}_{(m)}^{t} \approx \nabla\mathcal{L}_{(m)}^{t-1\top} \cdot d = \left\langle \nabla\mathcal{L}_{(m)}^{t-1}, d \right\rangle. \tag{6}$$

For brevity, we denote $\nabla\mathcal{L}_{(m)}^{t}$ as $g_{(m)}^{t}$. Inspired by conflict-averse gradient descent (CAGrad) (Liu et al. 2021), we formulate the following constrained optimization problem. Note that the $\left\langle g_{(m)}^{t-1}, d \right\rangle$ is the inner product between the gradient of m-th task and the update direction and measures their conflict, in which the lower values of the inner product mean the higher conflict. We first identify which task gradient among $m \in [1, M]$ exhibits the greatest conflict with the current update direction d as $\min_{m\in[1,M]} \left\langle g_{(m)}^{t-1}, d \right\rangle$. Subsequently, within a neighborhood centered at $g_{(0)}^{t-1} = \frac{1}{M}\sum_{m=1}^{M} g_{(m)}^{t-1}$, we search for a refined direction d that maximally aligns the conflicting gradients across tasks.

$$\begin{aligned} &\max_{d\in\mathbb{R}^N} \min_{m\in[1,M]} \left\langle g_{(m)}^{t-1}, d \right\rangle \\ &\text{s.t.} \quad ||d - g_{(0)}^{t-1}|| \leq c \cdot ||g_{(0)}^{t-1}||, \end{aligned} \tag{7}$$

where c controls the search space of the update gradient. Solving for d at each adaptation step requires high computational resources because the dimension of d, which equals the total number of parameters, could be millions. Instead, the dual problem of Eq. (7) reduces to a tractable optimization problem over an M-dimensional space. To get the dual problem, we introduce a variable w, matching the number of tasks M. Formally, $w = [w_{(1)}, w_{(2)}, ..., w_{(M)}] \in \mathcal{W}$, where $\mathcal{W} = \{w : \forall m, w_{(m)} \geq 0 \text{ and } \sum_{m=1}^{M} w_{(m)} = 1\}$, is the probability simplex. We arrive at the new version of the minimization problem: $\min_m \left\langle g_{(m)}^{t-1}, d \right\rangle = \min_w \left\langle \sum_m w_{(m)} g_{(m)}^{t-1}, d \right\rangle = \min_w \left\langle g_{w_{(m)}}^{t-1}, d \right\rangle$. Consequently, we obtain the following optimization problem:

$$\begin{aligned} &\max_{d\in\mathbb{R}^N} \min_{w\in W} \left\langle g_{w_{(m)}}^{t-1}, d \right\rangle \\ &\text{s.t.} \quad ||d - g_{(0)}^{t-1}|| \leq c \cdot ||g_{(0)}^{t-1}||. \end{aligned} \tag{8}$$

Applying Eq. (8), we derive the corresponding Lagrangian's equation as follows:

$$\max_{d\in\mathbb{R}^N} \min_{w\in W, \lambda\geq 0} g_{w_{(m)}}^{t-1\top} \cdot d - \frac{\lambda}{2}(||d - g_{(0)}^{t-1}||^2 - \phi), \tag{9}$$

where $\phi = c^2||g_{(0)}^{t-1}||^2$ and λ means Lagrange multiplier. Since $\phi > 0$, there exists a strictly feasible d that satisfies $||d - g_{(0)}^{t-1}|| \leq \phi$, which induces the convex feasible set. Consequently, one can check that Slater's condition (Shapiro and Scheinberg 2000) holds and strong duality applies, which allows switching the max and the min without changing the solution. Yielding the dual of the primal problem as follows:

$$\min_{w\in W, \lambda\geq 0} \max_{d\in\mathbb{R}^N} g_{w_{(m)}}^{t-1\top} \cdot d - \frac{\lambda}{2}(||d - g_{(0)}^{t-1}||^2 - \phi). \tag{10}$$

Using the Lagrange multiplier method repeatedly, i.e., setting the derivative w.r.t. d and λ alternately to zero, we obtain an optimal update direction d^*.

$$d^* = g_{(0)}^{t-1} + \frac{\sqrt{\phi}}{\|g_{(w)}^{t-1}\|} g_{(w)}^{t-1}, \tag{11}$$

where $g_{(w)}^{t-1} = \frac{1}{M}\sum_{m=1}^{M} w_{(m)} g_{(m)}^{t-1}$.

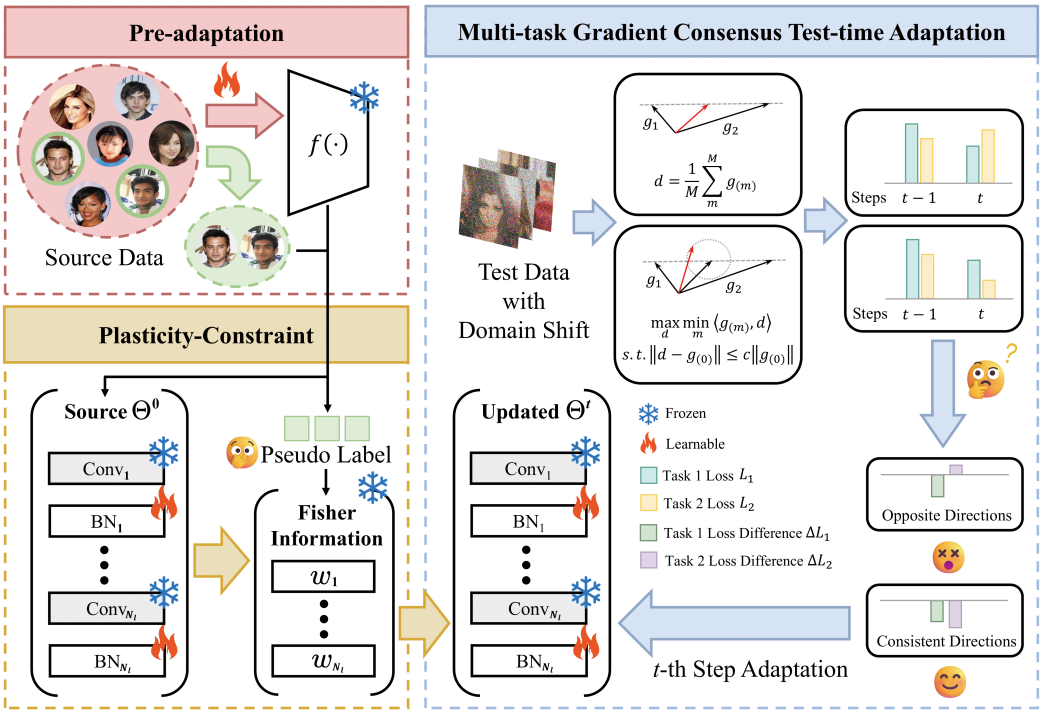

Figure 4: Our framework for MT-TTA. In the pre-adaptation phase, the model f is trained on the source domain to obtain the pre-trained parameters Θ^0. The pre-trained model generates pseudo labels corresponding to a random subset of source samples. The Fisher information matrix is computed for the plasticity constraint. During adaptation, the model explores the updated gradient space to find gradients that achieve a consensual direction on the loss difference $\Delta\mathcal{L}$, which are combined with the Fisher information to update Θ^t jointly.

Remark. Although Eq. (10) involves three variables, d, λ and w, the Lagrange multipliers d, λ can be analytically eliminated, leaving w as the sole variable in the dual problem. Specifically, denoting the objective in Eq. (10) as O. Setting $\nabla_d O = 0$ yields $d^* = g_{w_{(m)}}^{t-1}/\lambda + g_{(0)}^{t-1}$. Substituting d^* into O gives a reduced objective O'. Then setting $\nabla_\lambda O' = 0$ gives $\lambda^* = g_{w_{(m)}}^{t-1}/\sqrt{\phi}$. Substituting λ^* into O' yields the final dual problem: $\min_{w\in W} \sqrt{\phi}||g_{w_{(m)}}^{t-1}|| + g_{w_{(m)}}^{t-1\top} g_{(0)}^{t-1}$. This convex problem is solved efficiently via gradient descent to obtain w^*. The optimal gradient direction d^* is then recovered by plugging w^* into the closed-form expressions for d^* and λ^*.

Assumption 1. *Assume the m-th loss functions at the t-th step $\mathcal{L}_{(m)}^t, \forall m \in [1, M]$ are differentiable for all M tasks and its gradients $g_{(m)}$ are all H-Lipschitz, i.e., $||g_{(m)}^t - g_{(m)}^{t-1}|| \le H \cdot ||\Theta^t - \Theta^{t-1}||$, where $H \in (0, \infty)$. There exists a sufficiently small learning rate $\eta \in (0, \frac{1}{H}]$.*

Theorem 1 (Effectiveness of Gradient Consensus)**.** *When Assumption 1 holds, using the Gradient Consensus strategy to search update direction as Eq.* (11) *can yield:*

$$\Delta\mathcal{L}_{(m)}^{t-1} \le 0, \forall m \in \{1, 2, ..., M\} \tag{12}$$

Proof. The proof is provided in the Supplementary. □

Algorithm 1: The pipeline of proposed CoCo-MT-TTA.

Input: Training dataset $D_{\text{tr}} = \{(\mathbf{x}_{n_s}, y_{n_s})\}_{n_s=1}^{N_S}$; Domain-shifted dataset $D_{\text{te}} = \{(\mathbf{x}_{n_t}, y_{n_t})\}_{n_t=1}^{N_T}$; The randomly sampled dataset $D_{\text{rand}} \subset D_{\text{tr}}$; Test time adaption steps: T; Pre-trained parameters: Θ^0; Learnable parameters: $\Theta_l \subseteq \Theta$;
Output: Adapted parameters: Θ^T;

Initialize: Step $t = 1$;
Calculating Fisher information matrix F by Eq. (14);
while $t < T$ **do**
  for data batch $\mathbf{x}$ in D_{te} **do**
    Forward propagation through Θ^t;
    Calculate the final loss $\mathcal{L}$ as Eq. (16);
    Update learnable parameters Θ_l^t based on Eq. (11);
  end for
  $t \leftarrow t + 1$;
end while

With the fixed hyper-parameter $c \in [0, 1]$, we theoretically prove that $\Delta\mathcal{L}_{(m)} \le 0, \forall m$, which can give rise to better performance as evidenced in Figure 1.

4.2 Multi-task Plasticity-Constrained Adaptation

To mitigate catastrophic forgetting induced by overfitting to out-of-distribution (OOD) samples, we restrict parame-

Dataset	Method	Gaussian	Shot	Impulse	Defocus Blur	Brightness	Contrast	Average
PlantData	Source	73.03	75.55	73.14	80.88	74.85	75.84	$75.55_{\pm 2.62}$
	M-TENT	<u>92.17</u>	<u>95.84</u>	82.55	**97.05**	**98.62**	76.88	$90.52_{\pm 8.05}$
	EATA	92.63	93.88	**88.99**	95.83	97.08	**77.85**	<u>91.04</u>$_{\pm 6.43}$
	ActMAD	39.96	42.41	34.75	41.88	30.49	17.16	$34.44_{\pm 8.79}$
	SAR	89.09	91.73	83.83	92.92	92.95	77.44	$87.99_{\pm 5.67}$
	Ours	**93.12**	**95.98**	<u>87.25</u>	<u>96.95</u>	<u>98.56</u>	<u>77.52</u>	**91.57**$_{\pm 7.25}$
CelebA	Source	84.29	84.37	84.77	83.81	96.29	81.88	$85.90_{\pm 4.42}$
	M-TENT	<u>93.09</u>	**93.04**	<u>93.33</u>	<u>91.72</u>	<u>95.64</u>	<u>93.29</u>	<u>93.35</u>$_{\pm 1.16}$
	EATA	92.96	92.93	93.24	91.58	95.54	93.17	$93.24_{\pm 1.17}$
	ActMAD	79.16	76.37	81.69	76.15	79.45	48.25	$73.51_{\pm 11.46}$
	SAR	92.99	92.93	93.25	91.61	95.59	93.21	$93.26_{\pm 1.18}$
	Ours	**93.11**	<u>93.03</u>	**93.37**	**91.76**	**95.66**	**93.31**	**93.37**$_{\pm 1.15}$

Table 1: Mean In-Distribution and Out-of-Distribution (I&O) F1-score under various domain shift types (severity level 5) on the CelebA and PlantData dataset. Higher values indicate better performance. Our method outperforms all baseline approaches across all domain shifts. The **bold** number indicates the best result and the <u>underlined</u> number indicates the second best result.

ter updates based on the Fisher information matrix. Inspired by recent studies on anti-forgetting (Kirkpatrick et al. 2017), we enforce plasticity-constraint on the parameters during the adaptation stage by restricting their update range:

$$\mathcal{R}(\Theta_l, \Theta_l^0) = \sum_{\theta_i \in \Theta_l} w_f(\theta_i)(\theta_i - \theta_i^0)^2, \tag{13}$$

where $w_f(\theta_i)$ reflects the importance of θ_i, which represents the i-th parameter.

Inspired by Efficient Anti-forgetting Test-Time Adaptation (EATA) (Niu et al. 2022), we propose to multi-task aware parameter importance weight estimation based on the Fisher information matrix as:

$$w_f(\theta_i) = \frac{\sum_{m=1}^{M}\sum_{n_{sc}=1}^{N_{SC}} \left(\nabla_{\theta_i}\mathcal{L}_{\mathrm{CE},(m)}(f(\mathbf{x}_{n_{sc}};\Theta^0), \hat{y}_{n_{sc}})\right)^2}{MN_{SC}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{CE},(m)}$ is the m-th task's cross-entropy loss. N_{SC} means the number of the chosen samples from ID data. $f(\mathbf{x}_{n_{sc}};\Theta^0)$ and $\hat{y}_{n_{sc}}$ indicate soft and hard pseudo-labels from the pre-trained model $f(\Theta^0)$, respectively.

4.3 Overall Objective

For brevity, we denote $\left(\nabla_{\theta_i}\mathcal{L}_{\mathrm{CE},(m)}(f(\mathbf{x}_{n_{sc}};\Theta^0), \hat{y}_{n_{sc}})\right)$ as $g^f_{\theta_i,n_{sc}}$. Consequently, the final loss function integrates the plasticity-constrained term as follows:

$$\mathcal{L} = \frac{\sum_{m=1}^{M}}{M}\left[\mathcal{L}_{(m)} + \frac{\beta}{N_{SC}}\sum_{\theta_i\in\Theta_l}\sum_{n_{sc}=1}^{N_{SC}}(g^f_{\theta_i,n_{sc}})^2\right]. \tag{15}$$

Inspired by the entropy-based unsupervised loss for M-TENT (Chatterjee et al. 2024), we set $\mathcal{L}_{(m)}$ as the Shannon entropy, yielding the overall objective.

$$\mathcal{L} = \frac{1}{M}\sum_{m=1}^{M}\Bigg[-\sum_{c=1}^{C_{(m)}} p(\hat{y}_c)\log p(\hat{y}_c) \quad + \quad \frac{\beta}{N_{SC}}\sum_{\theta_i\in\Theta_l}\sum_{n_{sc}=1}^{N_{SC}}\left(g^f_{\theta_i,n_{sc}}\right)^2\Bigg], \tag{16}$$

where $C_{(m)}$ is the number of classes for m-th task, and β controls the regularization strength.

5 Experiments

5.1 Setup

Benchmarks. We evaluated our method on two datasets: CelebA (Liu et al. 2015) and PlantData, as shown in Table 2. We use a classic image corruptions algorithm (Hendrycks and Dietterich 2019) to create out-of-distribution (OOD) domain data, and source (in-distribution, ID) domain data has no corruption. We adopt the corruption types from M-TENT: Gaussian noise, shot noise, impulse noise, defocus blur, brightness, and contrast, and evaluate only on the highest corruption severity level (level 5) to assess model performance under challenging conditions.

Dataset	Total	Train	Test	Tasks
CelebA	202,599	141,819	60,780	4
PlantData	19,219	13,453	5,766	3

Table 2: Dataset statistics. Train/Test denotes a 70%/30% split. CelebA includes 4 binary attribute tasks; PlantData covers 3 multiclass plant health tasks.

Pre-Training Configuration. Our backbone follows the architecture of M-TENT (Chatterjee et al. 2024). During the training phase, we used a batch size of 256 and an initial learning rate of 1×10^{-3}. The model was trained for 100/50 epochs on the CelebA/PlantData dataset. We used the Adam optimizer (Kingma and Ba 2014) for both training and test-time adaptation phases, and all input images were resized to 224×224 before being fed into the network. We set the random seed to 42 for all experiments to ensure reproducibility.

Compared Methods. We compared our approach with representative test-time adaptation (TTA) methods. M-TENT (Chatterjee et al. 2024) adapts the model by minimiz-

Gradient Consensus	Plasticity Constraint	Average Mean I&O ↑		Average Mean I&O across Datasets ↑
		PlantData	CelebA	
✗	✗	90.52	93.35	91.94
✗	✓	<u>90.66</u>	93.35	<u>92.00</u>
✓	✗	90.57	<u>93.37</u>	91.97
✓	✓	**91.57**	**93.37**	**92.47**

Table 3: Performance comparison on the CelebA and PlantData under different domain shift types and settings. The **bold** number indicates the best result, and the <u>underlined</u> number indicates the second best result.

ing the entropy of multi-labels and updating the batch normalization statistics during inference. Since there are limited MT-TTA methods as a baseline, we extended the following methods from single to multi-task: Efficient Anti-forgetting Test-time Adaptation (EATA) (Niu et al. 2022) introduces a confidence-based sample selection strategy and employs Fisher regularization for model stability. Activation Matching (ActMAD) (Mirza et al. 2023) aligns the activation statistics of test samples with reference statistics from the source domain. Sharpness-aware and reliable entropy minimization (SAR) (Niu et al. 2023) filters out high-entropy samples and applies Sharpness-Aware Minimization (SAM) (Foret et al. 2021) during adaptation.

Test-Time Adaptation Configuration. At the test-time adaptation stage, we adapted the model using 10 steps for PlantData and CelebA (Liu et al. 2015). The learning rate during adaptation was set to $2.5 \times 10^{-3}/2.5 \times 10^{-4}$ by using SGD (Niu et al. 2022). We employed two adaptation-specific hyperparameters: β in Eq. (16) was fixed at 2000.0/1000.0, while c in Eq. (7) was set to 0.01/0.9 for PlantData/CelebA.

Evaluation Metrics. To evaluate the performance of test-time adaptation across both the source (in-distribution, ID) and out-of-distribution (OOD) domains, we adopt the $\text{I\&O}_{(m)}$ metric, defined as:

$$\text{I\&O}_{(m)} = 0.7 \cdot \text{F1}_{(m),(\text{ID})} + 0.3 \cdot \text{F1}_{(m),(\text{OOD})}, \tag{17}$$

where $\text{F1}_{(m),(\text{ID})}$ and $\text{F1}_{(m),(\text{OOD})}$ represent the F1-scores for the m-th task on the ID and OOD domains, respectively. The weights are set to 0.7/0.3 in line with the ratio of ID/OOD samples. To aggregate performance across all tasks, we compute the Mean I&O as:

$$\text{Mean I\&O} = \frac{1}{M} \sum_{m=1}^{M} \text{I\&O}_{(m)}, \tag{18}$$

where M is the total number of tasks in the dataset.

5.2 Performance comparison of the approaches

We benchmark our proposed method against existing classic TTA baselines under various domain shifts as shown in Table 1. **Task-generalization**. Across the 12 scenarios of CelebA and PlantData (6 shifts per benchmark), our method attains 7 state-of-the-art records and 5 the second-best results, evidencing superior multi-task generalization over all baselines. For example, under the Gaussian corruption of PlantData, we report 93.12 mean I&O, which means our adaptation method performs best across several tasks in the current scenario. **Shift-generalization**. Our method achieves the best performance across domains. As on CelebA, our method achieves the highest average mean I&O across diverse domain shifts, while simultaneously attaining the smallest standard deviation among all baselines, which demonstrates exceptional consensus in shift generalization. **Benchmark-generalization**. It is noteworthy that despite the significant semantic differences between the two datasets (face vs. plant images), our method demonstrates robust generalization capability.

5.3 Ablation Study

We perform an ablation study to quantify the contributions of two core modules in our method: Gradient Consensus (GC) and Plasticity Constraint (PC). As shown in Table 3, by selectively disabling these components during adaptation and evaluating on both ID and OOD data, we observe that the complete method (with both modules) achieves the best performance on mean I&O metrics. Adapted models with only one component perform worse, and without either module, perform the worst, affirming the necessity and complementary nature of both design choices.

We further evaluate modules robustness under six types of synthetic domain shifts with varying datasets. Our method maintains superior performance across all shift types and task types, achieving the highest mean I&O scores (mI&O). This indicates that our approach generalizes well to diverse domain shift patterns and offers a reliable solution for real-world MT-TTA deployments.

6 Conclusion.

In this paper, by theoretically analyzing Loss-performance with Taylor expansion, our findings on directional consensus as a performance indicator provide new insights for stable multi-task test-time adaptation (MT-TTA). Based on this insight, we present COnsensus-driven COnstrainted Multi-Task Test-Time Adaptation (CoCo-MT-TTA) for MT-TTA that explicitly addresses gradient conflicts and catastrophic forgetting. Evaluated on CelebA and PlantData, our approach achieves average improvements of $+11.75\%$ and $+10.30\%$ in the mean in-distribution and out-of-distribution F1-score (Mean I&O) across 6 domain shifts over a non-adapted model and classic baselines, respectively.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Nos. 2017YFB1400801 and 2022YFB3305500) and the National Natural Science Foundation of China (No. 62273089). We also acknowledge funding from the Guangdong Basic and Applied Basic Research Foundation (2025A1515010114), the Science and Technology Program of Guangzhou (2025A04J4334), and the Shenzhen Science and Technology Program (SGDX20230116091244004). Additional support was provided by the China University Industry-University-Research Innovation Fund (2024WA062), the "Climbing Program" Special Funds (pdjh2024a136), and the Cyberspace Security Engineering Technology Research Center of Guangdong University of Technology.

References

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Chatterjee, S.; Ghosh, A.; Kawsar, F.; and Malekzadeh, M. 2024. Analysing Softmax Entropy Minimization for Adaptating Multitask Models at Test-time. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Chen, G.; Niu, S.; Chen, D.; Zhang, S.; Li, C.; Li, Y.; and Tan, M. 2024. Cross-device collaborative test-time adaptation. *Advances in Neural Information Processing Systems*, 37: 122917–122951.

Chen, H.; Wu, S.; Wang, Z.; Yin, Y.; Jiao, Y.; Lyu, Y.; and Liu, Z. 2025. Causal-Inspired Multitask Learning for Video-Based Human Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2052–2060.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7482–7491.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Liu, B.; Liu, X.; Jin, X.; Stone, P.; and Liu, Q. 2021. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34: 18878–18890.

Liu, P.; Qiu, X.; and Huang, X. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

Mirza, M. J.; Soneira, P. J.; Lin, W.; Kozinski, M.; Possegger, H.; and Bischof, H. 2023. Actmad: Activation matching to align distributions for test-time-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24152–24161.

Misra, I.; Shrivastava, A.; Gupta, A.; and Hebert, M. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3994–4003.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, 16888–16905. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-time Adaptation in Dynamic Wild World. In *The Eleventh International Conference on Learning Representations*.

Ranjan, R.; Patel, V. M.; and Chellappa, R. 2017. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 121–135.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *nature*, 323(6088): 533–536.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Shapiro, A.; and Scheinberg, K. 2000. Duality and optimality conditions. In *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, 67–110. Springer.

Sun, T.; Lu, C.; and Ling, H. 2023. Domain adaptation with adversarial training on penultimate activations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 9935–9943.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7201–7211.