A Details of the Datasets

A.1 Description of Datasets

We evaluated our method on two datasets: CelebA (Liu et al. 2015) and PlantData. As shown in Table 1, CelebA consists of 202,599 images annotated with four binary classification tasks. PlantData, which comes from our construction based on the PlantVillage dataset (Mohanty et al. 2016) and Rice Plant Disease Dataset (Tunio et al. 2021), includes 19,219 images with three classification tasks. Each dataset was split into training and testing sets in a 70% / 30% ratio. CelebA involves four binary classification tasks, which we choose to follow M-TENT (Chatterjee et al. 2024): Attractive, Male, Smiling, and Lipstick. PlantData comprises three multi-class classification tasks: Crop Type (8 classes), Pathogen Type (5 classes), and Disease Name (18 classes).

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

Table 1: Dataset statistics. Train/Test denotes a 70%/30% split. CelebA includes 4 binary attribute tasks; PlantData covers 3 multiclass plant health tasks.

A.2 The Reasons for Selecting the Datasets

CelebA. We follow the experimental setup of M-TENT to select CelebA for evaluation.

PlantData. Two authoritative datasets were chosen to construct PlantData: Rice Plant Disease Dataset (Tunio et al. 2021) and PlantVillage (Mohanty, Hughes, and Salathé 2016). The reason why we merge the two datasets is that the disease types covered by the two datasets are complementary; for example, PlantVillage lacks comprehensive data related to rice diseases. Therefore, we can create a more comprehensive dataset. Combining different datasets provides richer training samples and diverse image qualities. Conducting experiments with samples of varying quality allows for a more accurate evaluation of our method.

A.3 Data Preprocessing

PlantData. In PlantVillage, we removed repeated crop species and selected crop types with two or more label classes (including healthy). For classes with an excessive amount of data (such as tomato), we employed random downsampling operations to achieve dataset balance. In the Rice Plant Disease Dataset, data augmentation techniques were employed to increase the diversity of the data and mitigate extreme data imbalances. Ultimately, we merged the two processed datasets to construct the Plant Leaf Disease Dataset. The quantitative information is shown in Table 1.

Constructing Domain-Shifted Data. Following M-TENT, we select six image corruptions to investigate as shown in Figure 1. As an example, Gaussian Noise injects additive noise sampled from a zero-mean Gaussian distribution to each pixel of the original image, where the standard deviation of the distribution scales with the corruption severity. To ensure practical and challenging evaluation conditions, we focus our analysis on the most severe corruption level, i.e., severity 5.

Figure 1: Domain shifts on CelebA.

Symbol	Description
M	Total number of tasks
$(\cdot)_{(m)}$	The m-th task of $\cdot$
N_S	Number of source samples
n_s	n_s-th source sample
N_T	Number of target samples
n_t	n_t-th target sample
$(\cdot)^t$	t-th adaptation step
T	Total adaptation steps
$P(\mathbf{x})$	Source domain distribution
$Q(\mathbf{x})$	Target domain distribution
$y_{(m),n}$	Label for the m-th task of the n-th sample
Θ	Model parameters
N_p	Total number of model parameters
Θ_l	Learnable parameters
N_l	Number of learnable parameters
θ	Specific learnable parameter
$\mathcal{W}$	Probability simplex
F	Fisher information matrix
$\hat{\mathbf{y}}_n$	Hard pseudo-labels
N_{SC}	Number of samples for Fisher matrix computation
w_f	Parameter regularization weight
η	Learning rate
d	Parameter update direction
β	Regularization coefficient
$C_{(m)}$	Number of classes for the m-th task

Table 2: Notation summary.

B Details of the Mathematical Derivation

In this section, we provide the full proof of Thm 1.

Assumption 1. *Assume the m-th loss functions at the t-th step $\mathcal{L}^t_{(m)}, \forall m \in [1, M]$ are differentiable for all M tasks and its gradients $g_{(m)}$ are all H-Lipschitz, i.e., $||g^t_{(m)} - g^{t-1}_{(m)}|| \leq H \cdot ||\Theta^t - \Theta^{t-1}||$, 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 can yield:*

$$\Delta\mathcal{L}^{t-1}_{(m)} \leq 0, \forall m. \tag{1}$$

Proof. We use a second-order Taylor expansion. For brevity, we denote $\nabla_{\Theta^{t-1}}\mathcal{L}_{(m)}$ as $g^{t-1}_{(m)}$.

$$\begin{aligned}\Delta\mathcal{L}^t_{(m)} &= \mathcal{L}^t_{(m)} - \mathcal{L}^{t-1}_{(m)} \\ &\approx {g^{t-1}_{(m)}}^\top \left[(\Theta^t - \eta \cdot d^*) - \Theta^t\right] \\ &+ \frac{1}{2}\nabla^2\mathcal{L}^{t-1}_{(m)} \left[(\Theta^t - \eta \cdot d^*) - \Theta^t\right]^2 \\ &= -\eta \cdot {g^{t-1}_{(m)}}^\top \cdot d^* + \frac{\eta^2}{2} {d^*}^\top \nabla^2 \mathcal{L}^{t-1}_{(m)} d^*,\end{aligned} \tag{2}$$

where d^* is the update direction of the gradient consensus strategy. $\nabla^2(\cdot)$ means the second-order derivative. Since $d^\top H d \leq \lambda_{max}(H)\|d\|^2$, it follows that:

$$\begin{aligned}\Delta\mathcal{L}^t_{(m)} &\leq -\eta \cdot {g^{t-1}_{(m)}}^\top \cdot d^* + \frac{H\eta^2}{2} \cdot \|d^*\|^2 \\ &\leq -\eta \cdot \min_{w_{(m)}} < g^{t-1}_{(w)}, d^* > + \frac{H\eta^2}{2} \cdot \|d^*\|^2.\end{aligned} \tag{3}$$

For the optimization problem $\min_m < g_{(m)}, d^* >= \min_w < g^{t-1}_{(w)}, d^* >$, invoking the method of Lagrange multipliers yields the closed-form expression $\min_{w_{(m)}} < g^{t-1}_{(w)}, d^* >= {g^{t-1}_{(w)}}^\top g^{t-1}_{(0)} + c \cdot \|g^{t-1}_{(0)}\| \cdot \|g^{t-1}_{(w)}\|$. Substituting this result into Eq. (3), we arrive at:

$$\begin{aligned}\Delta\mathcal{L}^t_{(m)} \leq &-\eta \cdot \left({g^{t-1}_{(w)}}^\top g^{t-1}_{(0)} + c \cdot \|g^{t-1}_{(0)}\| \cdot \|g^{t-1}_{(w)}\|\right) \\ &+ \frac{H\eta^2}{2} \cdot \|d^*\|^2.\end{aligned} \tag{4}$$

Given $d^* = g^{t-1}_{(0)} + \frac{\sqrt{\phi}}{\|g^{t-1}_{(w)}\|} g^{t-1}_{(w)}$, where $\phi = c^2\|g^{t-1}_{(0)}\|^2$. The derivation proceeds as follows.

$$\begin{aligned}\|d^*\|^2 &= \|g^{t-1}_{(0)}\|^2 + \frac{\phi}{\|g^{t-1}_{(w)}\|^2}\|g^{t-1}_{(w)}\|^2 \\ &+ 2 \cdot \frac{\phi}{\|g^{t-1}_{(w)}\|} \cdot {g^{t-1}_{(w)}}^\top g^{t-1}_{(0)} \\ &= (1 + c^2)\|g^{t-1}_{(0)}\|^2 + \frac{2\phi}{\|g^{t-1}_{(w)}\|} {g^{t-1}_{(w)}}^\top g^{t-1}_{(0)} \\ &= (1 - c^2)\|g^{t-1}_{(0)}\|^2 \\ &+ \frac{2\phi}{\|g^{t-1}_{(w)}\|}\left({g^{t-1}_{(w)}}^\top g^{t-1}_{(0)} + c\|g^{t-1}_{(w)}\| \cdot \|g^{t-1}_{(0)}\|\right).\end{aligned} \tag{5}$$

For brevity, we denote ${g^{t-1}_{(w)}}^\top g^{t-1}_{(0)} + c\|g^{t-1}_{(0)}\| \cdot \|g^{t-1}_{(w)}\|$ as A. Plugging Eq. (5) into the Eq. (4) yields:

$$\begin{aligned}\Delta\mathcal{L}^t_{(m)} &\leq -\eta \cdot A + \frac{H\eta^2}{2}(1 - c^2)\|g^{t-1}_{(0)}\|^2 + \frac{H\eta^2 c\|g_{(0)}\|}{\|g_{(w)}\|} A \\ &= -\eta(1 - cH\eta\frac{\|g_{(0)}\|}{\|g_{(w)}\|})A + \frac{H\eta^2(1 - c^2)}{2}\|g^{t-1}_{(0)}\|^2 \\ &= -\eta(1 - cH\eta\frac{\|g_{(0)}\|}{\|g_{(w)}\|})(c + \cos\xi)\|g^{t-1}_{(0)}\| \cdot \|g^{t-1}_{(w)}\| \\ &+ \frac{H\eta^2(1 - c^2)}{2}\|g^{t-1}_{(0)}\|^2,\end{aligned} \tag{6}$$

where ξ indicates the angle between $g^{t-1}_{(0)}$ and $g^{t-1}_{(w)}$. Depending on the sign of $-\eta(1 - cH\eta\frac{\|g_{(0)}\|}{\|g_{(w)}\|})$, the scaling direction differs, i.e., $\cos\xi = 1$ or $\cos\xi = -1$.

1) Assuming $\eta \in \left[\frac{\|g_{(w)}\|}{Hc\|g_{(0)}\|}, \frac{K\|g_{(w)}\|}{H(c+1)\|g_{(0)}\|}\right] \subseteq (0, \frac{1}{H}]$, the following derivation holds.

$$\begin{aligned}\Delta\mathcal{L}^t_{(m)} &\leq -\eta(1 - cH\eta\frac{\|g^{t-1}_{(0)}\|}{\|g^{t-1}_{(w)}\|})(1 + c)\|g^{t-1}_{(0)}\| \cdot \|g^{t-1}_{(w)}\| \\ &+ \frac{H\eta^2(1 - c^2)}{2}\|g^{t-1}_{(0)}\|^2 \\ &= -\eta(1 + c)\|g^{t-1}_{(0)}\| \cdot \|g^{t-1}_{(w)}\| \\ &+ \frac{H\eta^2(1 + c)^2}{2}\|g^{t-1}_{(0)}\|^2 \\ &\leq -\frac{1}{H}\|g^{t-1}_{(w)}\|^2 + \frac{K^2}{2H}\|g^{t-1}_{(w)}\|^2 \\ &= \frac{1}{H}\|g^{t-1}_{(w)}\|^2\left[\frac{K^2}{2} - 1\right],\end{aligned} \tag{7}$$

where $\max\{\frac{1}{c} - 1, 0\} < K < \min\{\sqrt{2}, \frac{(c+1)\|g_{(0)}\|}{\|g_{(w)}\|}\}$.

2) Under the assumption that $\eta < \frac{\|g_{(w)}\|}{cH\|g_{(0)}\|} < \frac{\|g_{(w)}\|}{(c-1)H\|g_{(0)}\|}$ holds, the derivation unfolds as:

$$\begin{aligned}\Delta\mathcal{L}^t_{(m)} &\leq -\eta(1 - cH\eta\frac{\|g_{(0)}\|}{\|g_{(w)}\|})(c - 1)\|g^{t-1}_{(0)}\| \cdot \|g^{t-1}_{(w)}\| \\ &+ \frac{H\eta^2(1 - c^2)}{2}\|g^{t-1}_{(0)}\|^2 \\ &= -\eta(c - 1)\|g^{t-1}_{(0)}\| \cdot \|g^{t-1}_{(w)}\| \\ &+ \frac{H\eta^2}{2}(c - 1)^2\|g^{t-1}_{(0)}\|^2 \\ &< -\frac{1}{H}\|g^{t-1}_{(w)}\|^2 + \frac{1}{2H}\|g^{t-1}_{(w)}\|^2 \\ &= -\frac{1}{2H}\|g^{t-1}_{(w)}\|^2.\end{aligned} \tag{8}$$

Therefore, from Eq. (7) and Eq. (8), $\Delta\mathcal{L}_{(m)} \leq 0$ holds for all m tasks. □

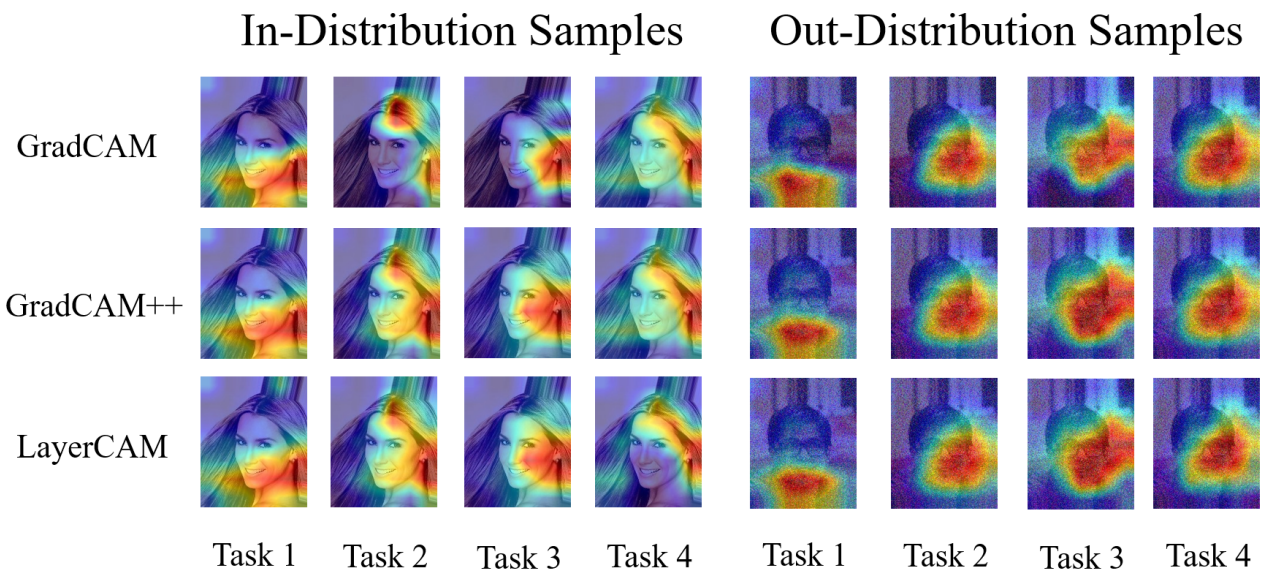

Figure 2: Visualization of model's output acorss different tasks.

C Detailed Environment Settings and Training Configuration.

C.1 Environment Settings.

Experiments were conducted on a workstation with a GeForce RTX 4090 D (22 GB) GPU and an AMD Ryzen 9 9950X CPU.

C.2 Training Configuration.

We adopt the 18-layer Residual Network (ResNet-18) (He et al. 2016) as the backbone and append multiple task-specific output heads for multi-task learning. 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 epochs on the CelebA dataset and for 50 epochs on the PlantData dataset, respectively. 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.

D Attribution-Based Analysis

We employ GradCAM (Selvaraju et al. 2017), GradCAM++ (Lerma and Lucas 2022), and LayerCAM (Jiang et al. 2021) to visualize the model outputs, as shown in Fig. 2. Across the four CelebA tasks, the model consistently attends to similar facial regions. For in-distribution samples, it mainly focuses on the ear region in all tasks, indicating that the CelebA tasks selected under the M-TENT protocol do not exhibit strong gradient conflict. This observation helps explain why the Gradient Consensus (GC) module in CoCo offers only limited gains compared with M-TENT.

E Limitations.

Based on the analysis in Section D, we argue that the field of multi-task test-time adaptation (MT-TTA) still lacks a standardized benchmark with well-defined evaluation criteria. Furthermore, in practice, the GC module requires solving an optimization problem at each updating step, which reduces computational efficiency. In future work, we aim to develop more efficient MT-TTA algorithms that achieve an optimal balance between mitigating gradient conflict and maintaining high efficiency.

References

Jiang, P.-T.; Zhang, C.-B.; Hou, Q.; Cheng, M.-M.; and Wei, Y. 2021. Layercam: Exploring hierarchical class activation maps for localization. *IEEE transactions on image processing*, 30: 5875–5888.

Lerma, M.; and Lucas, M. 2022. Grad-CAM++ is equivalent to Grad-CAM with positive gradients. *arXiv preprint arXiv:2205.10838*.

Mohanty, S. P.; Hughes, D. P.; and Salathé, M. 2016. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Tunio, M. H.; Jianping, L.; Butt, M. H. F.; and Memon, I. 2021. Identification and Classification of Rice Plant Disease Using Hybrid Transfer Learning. In *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 525–529.