Supplementary Material for WACV 2026 ID 1955
Logit-Adjusted Test-Time Adaptation under Partial Class Imbalance

Thilina Weerasinghe, Ruwan Tennakoon, WeiQin Chuah, Alireza Bab-Hadiashar
RMIT University, Australia
{thilina.weerasinghe,ruwan.tennakoon,wei.qin.chuah,alireza.bab-hadiashar}@rmit.edu.au

1. Theoretical Analysis

Lemma 1 (Gradient of entropy loss)**.** *Let*

$$\mathcal{L}(z) = -\sum_{i=1}^{C} p_i \log p_i, \quad p_i = \frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}},$$

be the entropy loss applied to softmax probabilities. Then the gradient w.r.t. the classifier input u, where $z = Wu + b$ and, $\omega_{i:}$ denotes the i-th row of W.

$$\frac{\partial \mathcal{L}}{\partial u} = \sum_{i=1}^{C} p_i \left(\mathbb{E}_p[z] - z_i\right) \omega_{i:},$$

Proof. We first compute the gradient w.r.t. logits z. Using gradients of softmax:

$$\frac{\partial p_i}{\partial z_j} = p_i(\delta_{ij} - p_j) = \begin{cases} p_i(1-p_i) & \text{if } i = j \\ -p_i(p_j) & \text{if } i \neq j \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial z_j} = -\sum_{i=1}^{C} (\log p_i + 1) p_i (\delta_{ij} - p_j).$$

Splitting into the $i = j$ and $i \neq j$ parts, rearranging, and simplifying yields

$$\frac{\partial \mathcal{L}}{\partial z_j} = p_j \left(\mathbb{E}_p[\log p] - \log p_j\right).$$

Substituting $\log p_i = z_i - \log Z$ with $Z = \sum_k e^{z_k}$ cancels $\log Z$, giving

$$\frac{\partial \mathcal{L}}{\partial z_j} = p_j \left(\mathbb{E}_p[z] - z_j\right).$$

In vector form, $\frac{\partial \mathcal{L}}{\partial z} = p \odot (\mathbb{E}_p[z]\mathbf{1} - z)$.

Finally, backpropagating through the linear classifier $z = Wu + b$,

$$\frac{\partial \mathcal{L}}{\partial u} = W^\top \frac{\partial \mathcal{L}}{\partial z} = \sum_{i=1}^{C} p_i \left(\mathbb{E}_p[z] - z_i\right) \omega_{i:}.$$

Lemma 2. *Let $\beta_\ell^{(t)}$ be the batch-normalization shift parameter at iteration t, and let the update rule be given by*

$$\beta_\ell^{(t+1)} = \beta_\ell^{(t)} - \eta \upsilon^{(t)} v_c, \tag{1}$$

where η is nonzero scalar constants, $\upsilon = \upsilon^{(t)} \forall t$, and v_c is a nonzero constant vector. If $\upsilon < 0$, then as the number of iterations T approaches infinity, the vector $\beta_\ell^{(T)}$ aligns with the direction of v_c. That is,

$$\lim_{T\to\infty} \frac{\beta_\ell^{(T)} \cdot v_c}{\|\beta_\ell^{(T)}\| \|v_c\|} = 1.$$

Proof. Iterating the update rule T steps gives

$$\beta_\ell^{(T)} = \beta_\ell^{(0)} - T\eta\upsilon v_c$$

Compute the cosine similarity between $\beta_\ell^{(T)}$ and v_c:

$$\begin{aligned} \cos\theta_T &= \frac{\beta_\ell^{(T)} \cdot v_c}{\|\beta_\ell^{(T)}\| \|v_c\|} \\ &= \frac{\beta_\ell^{(0)} \cdot v_c - T\eta\upsilon\|v_c\|^2}{\|v_c\| \sqrt{\|\beta_\ell^{(0)}\|^2 - 2T\eta\upsilon(\beta_\ell^{(0)} \cdot v_c) + (T\eta\upsilon)^2 \|v_c\|^2}}. \end{aligned}$$

Divide numerator and the expression under the square root by T and T^2 respectively:

$$\cos\theta_T = \frac{\frac{1}{T}(\beta_\ell^{(0)} \cdot v_c) - \eta\upsilon\|v_c\|^2}{\|v_c\| \sqrt{\frac{1}{T^2}\|\beta_\ell^{(0)}\|^2 - \frac{2}{T}\eta\upsilon(\beta_\ell^{(0)} \cdot v_c) + (\eta\upsilon)^2 \|v_c\|^2}}.$$

Taking the limit $T \to \infty$, the $\frac{1}{T}$ and $\frac{1}{T^2}$ terms vanish, leaving

$$\lim_{T\to\infty} \cos\theta_T = \frac{-\eta\upsilon\|v_c\|^2}{\|v_c\| \sqrt{(\eta\upsilon)^2 \|v_c\|^2}} = \frac{-\eta\upsilon}{|\eta\upsilon|}.$$

Since we assume $\eta\upsilon < 0$, we have $-\eta\upsilon > 0$ and thus $-\eta\upsilon/|\eta\upsilon| = 1$. Therefore

$$\lim_{T\to\infty} \cos\theta_T = 1,$$

which means $\beta_\ell^{(T)}$ becomes perfectly aligned with v_c as $T \to \infty$.

Lemma 3 (Logit bias under β-alignment)**.** *Assume a frozen downstream Jacobian J_ψ and let the BN shift update follow $\beta_\ell^{(t+1)} = \beta_\ell^{(t)} - \eta\, \upsilon\, v_c$ with step size $\eta > 0$ and scalar $\upsilon < 0$ from the entropy gradient. Then the logits evolve as*

$$z^{(t+1)} \;\approx\; z^{(t)} \;-\; \eta\, \upsilon\, W\, J_\psi J_\psi^\top \omega_{c:}.$$

In particular,

$$\Delta z_c^{(t)} \;=\; -\eta\, \upsilon\, \|J_\psi^\top \omega_{c:}\|_2^2 \;>\; 0,$$

$$\Delta z_k^{(t)} \;=\; -\eta\, \upsilon\, \langle J_\psi^\top \omega_{k:},\, J_\psi^\top \omega_{c:} \rangle \quad (k \neq c).$$

Proof. We assume frozen Jacobian J_ψ (no training in ψ) which makes the update for β^ℓ as $\beta_\ell^{(T)} = \beta_\ell^{(0)} - T\eta\upsilon v_c$, where $\eta > 0$, and $\upsilon < 0$.

The effect of this alignment on the logits for a specific unseen data-point can be understood by considering a first-order Taylor expansion of the network's function. For a specific test example, linearizing ψ about the current $h_\ell^{(t)}$ gives

$$\psi(h_\ell^{(t+1)}) \approx \psi(h_\ell^{(t)}) + J_\psi(h_\ell^{(t)}) \left(h_\ell^{(t+1)} - h_\ell^{(t)}\right)$$

Since we are looking at the effect of only β, we can write:

$$\psi(h_\ell^{(t+1)}) \approx \psi(h_\ell^{(t)}) + J_\psi(h_\ell^{(t)}) \left(\beta_\ell^{(t+1)} - \beta_\ell^{(t)}\right)$$

This linear approximation shows that the change in the BN output directly contributes an additive bias to the input of the final classifier. Substituting this into the logit equation, we find that the logits evolve as:

$$z^{(t+1)} \approx z^{(t)} + W J_\psi(h_\ell^{(t)}) \left(\beta_\ell^{(t+1)} - \beta_\ell^{(t)}\right)$$

From equation (1), we get $\left(\beta_\ell^{(t+1)} - \beta_\ell^{(t)}\right) = -\eta\upsilon v_c = -\eta\upsilon J_\psi^\top \omega_{c:}$. Substituting this gives the additive bias

$$z^{(t+1)} \approx z^{(t)} - \eta\upsilon W J_\psi J_\psi^\top \omega_{c:}.$$

In particular, the change to the dominant class logit is

$$\Delta z_c = -\eta\upsilon \omega_{c:}^\top (J_\psi J_\psi^\top) \omega_{c:} = -\eta\upsilon \|J_\psi^\top \omega_{c:}\|^2 > 0,$$

Thus, in the presence of class imbalance, each entropy minimisation update contributes an additive bias that monotonically raises the logit of the over-represented class. For other classes $k \neq c$, the change is

$$\Delta z_k = -\eta\upsilon \langle J_\psi^\top \omega_{k:}, J_\psi^\top \omega_{c:} \rangle$$

Method	Domain Shift			Avg
	R → S	R → P	R → C	
AllAcc (%) ↑				
EATA	44.5	58.9	51.6	51.7
DEYO	47.2	57.9	51.3	52.2
TENT	48.6	61.0	54.0	54.5
SAR	49.3	62.9	54.4	55.6
EATA + LA	44.8	60.1	52.0	52.3
DEYO + LA	49.2	59.9	53.1	54.1
TENT + LA	48.6	62.4	54.1	55.0
SAR + LA	50.3	63.4	55.3	56.4
PeAR (%) ↓				
DEYO	6.32	9.44	7.05	7.60
EATA	4.42	5.81	3.99	4.74
TENT	2.14	4.21	2.05	2.80
SAR	0.38	0.42	0.48	0.43
DEYO + LA	2.96	6.65	4.12	4.57
EATA + LA	4.40	3.70	3.01	3.71
TENT + LA	1.54	2.04	1.78	1.79
SAR + LA	1.64	0.34	0.42	0.80

Table 1. Comparison of baseline methods and their logit-adjusted counterparts (LA) with CIR = 0.1 for DomainNet-126 dataset

2. Experimental Results and Discussion

2.1. Comprehensive Results

In this section, we present comprehensive results for DomainNet-126 [2] dataset (Tab. 1) and Vision Transformer (Vit-B) based experiments (Tab. 2).

2.2. Discussion

Hyperparameter Sensitivity Analysis We analyse the effect of τ and α on logit adjustment and adaptation stability. The parameter τ controls the strength of logit adjustment: lower values yield weaker adjustment, leading to high PeAR, while larger values apply stronger adjustment and result in lower PeAR with only minor reductions in AllAcc (see Fig. 1).

The parameter α governs the update rate of the target prior. Small α values enable the prior to adapt quickly to the current minibatch, accelerating convergence to the oracle prior and improving performance, as reflected by low PeAR. Conversely, large α values slow adaptation, causing the prior to rely more heavily on historical estimates and resulting in higher PeAR.

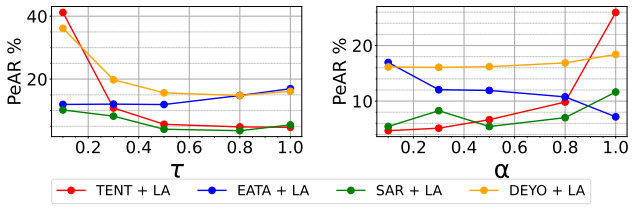

Figure 1. PeAR % variation with τ (left) and α (right)

Method	Noise			Blur				Weather				Digital				Avg
	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixel	JPEG	
AllAcc (%) ↑																
EATA	52.0	56.3	55.5	46.6	54.4	53.3	55.7	68.4	62.0	51.4	80.1	2.54	64.4	75.5	71.7	56.7
TENT	59.8	61.1	61.4	58.6	9.31	63.9	59.3	68.3	64.8	0.13	79.7	66.4	59.9	73.9	70.3	57.1
DEYO	52.4	53.7	53.0	47.5	49.4	59.4	55.7	66.1	61.6	70.1	77.7	59.5	61.9	72.4	69.5	60.7
SAR	57.4	58.5	58.5	55.6	54.0	61.3	57.1	66.2	62.8	70.0	79.2	30.9	59.3	73.1	69.8	60.9
TENT + LA	59.9	61.3	61.7	57.7	55.4	63.1	59.3	67.5	63.4	0.16	79.3	65.1	59.1	73.0	69.5	59.7
EATA + LA	57.9	59.6	59.5	56.1	53.7	61.9	57.6	68.2	63.3	22.9	79.8	57.0	58.8	72.5	69.3	59.9
SAR + LA	59.5	61.2	61.1	57.3	56.2	63.2	59.1	67.5	64.7	71.4	79.8	43.1	60.7	73.8	70.0	63.2
DEYO + LA	55.7	57.9	58.3	56.2	56.4	62.3	59.3	68.1	65.7	69.5	78.8	62.1	64.4	73.8	71.2	64.0
PeAR (%) ↓																
EATA	12.0	7.31	7.44	21.2	4.46	20.5	8.08	1.50	6.51	36.8	0.42	1.95k	3.09	1.28	1.51	139
TENT	0.50	0.31	0.47	1.09	399	1.55	1.92	2.22	2.00	1.01k	0.52	1.96	3.71	2.25	1.98	95.2
DEYO	7.10	8.09	8.69	15.4	10.4	2.90	4.74	0.95	2.12	2.01	2.00	2.88	1.64	2.22	1.41	4.83
SAR	2.89	3.86	3.83	1.83	1.41	1.28	0.89	0.81	1.35	1.00	0.23	5.08	3.97	1.56	1.04	2.07
SAR + LA	0.67	0.47	0.56	0.48	1.83	0.95	1.61	1.08	1.50	1.00	0.50	50.1	4.21	2.08	1.40	4.56
DEYO + LA	5.21	4.69	3.20	1.64	1.66	1.92	2.02	1.06	1.72	0.65	1.75	1.61	0.99	1.96	0.62	2.05
TENT + LA	0.37	0.34	0.67	0.52	3.12	1.24	1.99	1.42	1.81	0.00	0.21	1.60	3.98	1.46	1.42	1.34
EATA + LA	0.00	0.00	0.00	0.00	0.00	0.00	0.36	2.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21

Table 2. Comparison of baseline methods and their counterparts enhanced with logit adjustment (LA) method, under partial class imbalance with CIR = 0.1 for Vision Transformers (Vit-B). Numbers with k denotes $\times 10^3$.

Class Imbalance Variation Comparison with Baselines In the main manuscript, we discussed that the Class Inclusion Ratio (CIR), in other words, the proportion of observable classes, has an impact on the adaptation process. Higher CIR (milder imbalance) yields higher AllAcc and lower PeAR for both baselines and LA variants. As evident in the Fig. 2, our method consistently outperforms the baseline at every CIR level in terms of both AllAcc and PeAR. For experiments associated with CIR values greater than 0.1, we apply a prior warmup with 300 steps before applying the logit adjustment.

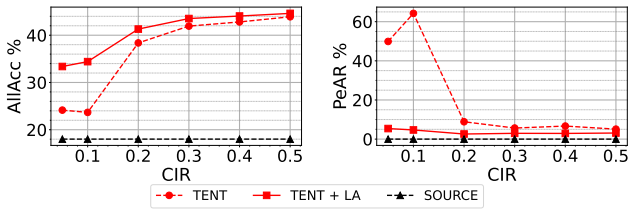

Figure 2. AllAcc % (left) and PeAR % (right) with CIR variation for TENT baseline and its logit-adjusted version. Results are obtained for the ImageNet-C [1] dataset with level 5 corruption

Ablation Study for the Prior π^{batch} We conduct an ablation study to evaluate different strategies for estimating the target prior π^{batch} at the minibatch level. As shown in Tab. 3, the softmax output-based prior, which aggregates class probabilities across samples, provides a performance gain over the baseline. This improvement stems from its ability to capture uncertainty in model predictions and to reduce the impact of noisy individual samples.

However, we observe that the prediction-based prior, computed directly from the hard class assignments (argmax of model outputs) consistently yields the best results. Unlike the softmax prior, the prediction-based prior produces sharper class distributions that better approximate the underlying target prior. This sharper estimate accelerates the convergence of the exponential moving average prior update, which in turn leads to more effective logit adjustment. Although the softmax prior is inherently more robust to noisy predictions, its smoother distribution tends to underestimate the dominance of frequently occurring target classes, causing slower adaptation and ultimately lower performance.

LA	✗	✓	
Prior Type	N/A	Softmax	Prediction
AllAcc	23.7	28.5	34.4
PeAR	64.2	23.9	4.68

Table 3. Comparison of performance for different priors. Results are obtained for TENT at CIR = 0.1 for ImageNet-C dataset.

References

[1] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3

[2] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 2