ICLR International Conference On Learning Representations

3RD WORKSHOP ON TEST-TIME UPDATES

ProtoTTA: Prototype-Guided Test-Time Adaptation

Mohammad Mahdi Abootorabi[1,3], Parvin Mousavi[2,3], Purang Abolmaesumi[1], Evan Shelhamer[1,3]

[1] University of British Columbia, [2] Queen's University, [3] Vector Institute

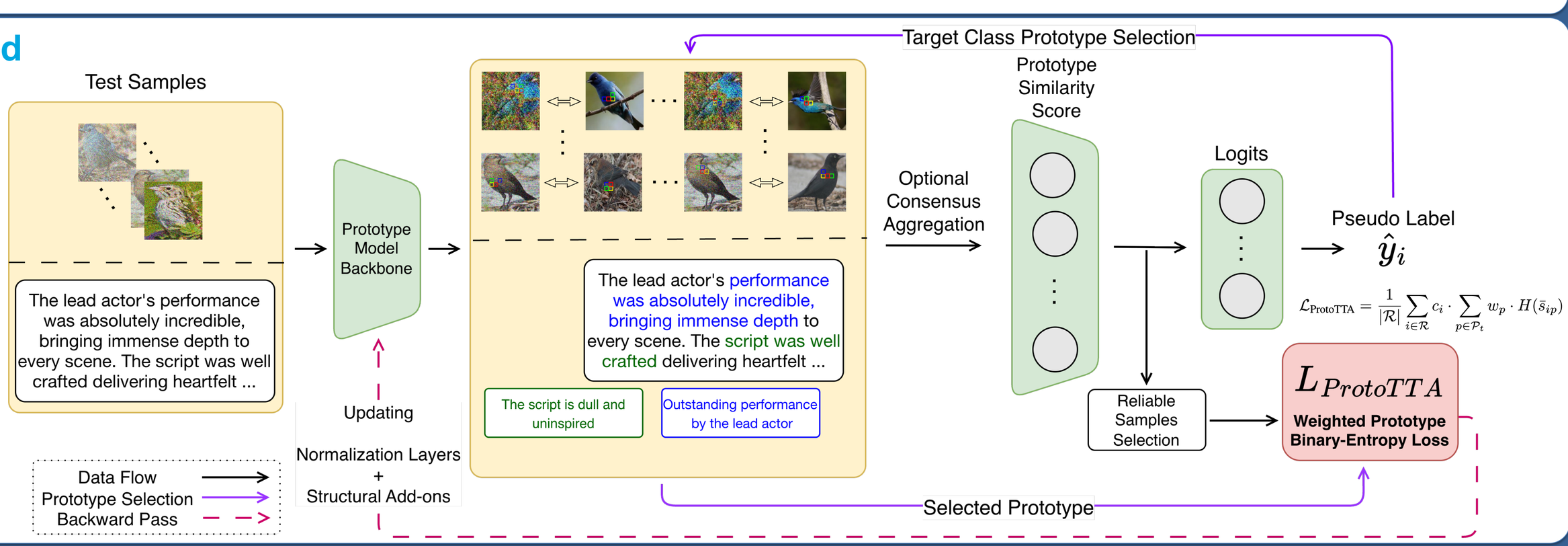
Motivation

- **Prototype Models**: Interpretable but still brittle under distribution shift.
- **Black-Box TTA**: Existing Test-Time Adaptation (TTA) ignoring rich, interpretable intermediate signals treats models as black boxes.

Contribution

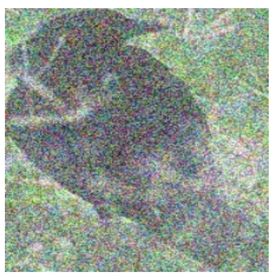
- **Prototype-Guided TTA Framework:** Leverages prototype activations and interpretability signals for semantically grounded TTA.
- **Interpretability-Aligned Metrics:** Introduces new interpretability metrics to quantify semantic stability and prototype alignment; validated through strong correlation with VLM-rated reasoning quality.
- **VLM-Driven Explainability for TTA:** Provides the first language-grounded evaluation of TTA dynamics using reasoning boards, enabled uniquely by prototype transparency.

Method

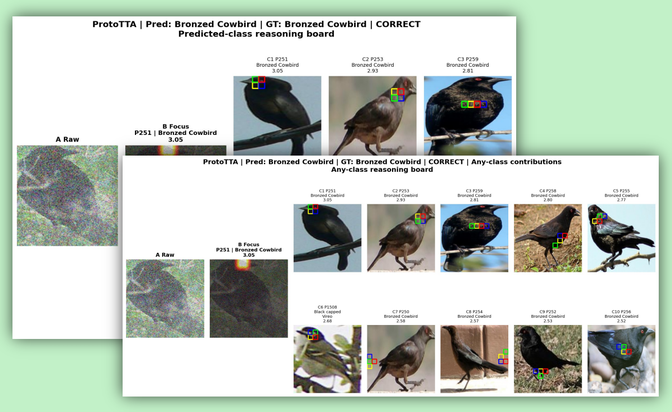


Explainable TTA via Prototype Reasoning Boards

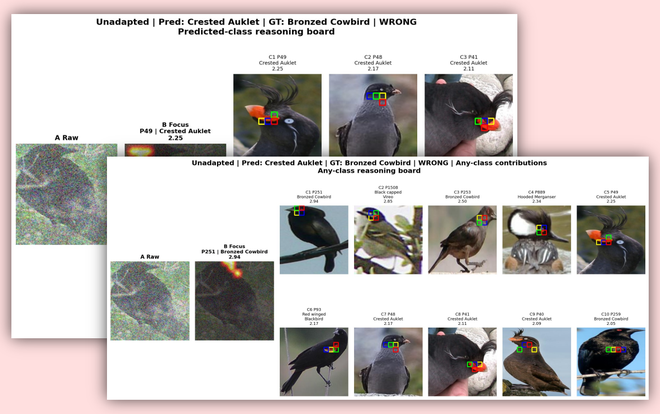
Test Image



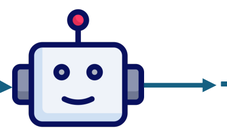
Adaptation Reasoning Boards



Unadapted Reasoning Boards



Reasoning VLM Agent



✓ **Unadapted Analysis:** The model misidentified the bird as a Crested Auklet by focusing on noisy artifacts in the beak/crown region and matching incorrect prototypes, overwhelming correct evidence.
✓ **Adaptation Analysis:** Correct identification achieved by shifting attention from the ambiguous beak to the highly species-discriminative red eye, retrieving accurate Bronzed Cowbird prototypes.
✓ **Change Impact:** Suppression of spurious wrong-class (Auklet) evidence in favor of correct-class (Cowbird) matches.

Comparative Checklist: ✓ Attention location improved ✓ Wrong-class evidence suppressed ✓ Prototype match improved ✓ Semantic part focus improved

Findings

- **Interpretability Metrics:**

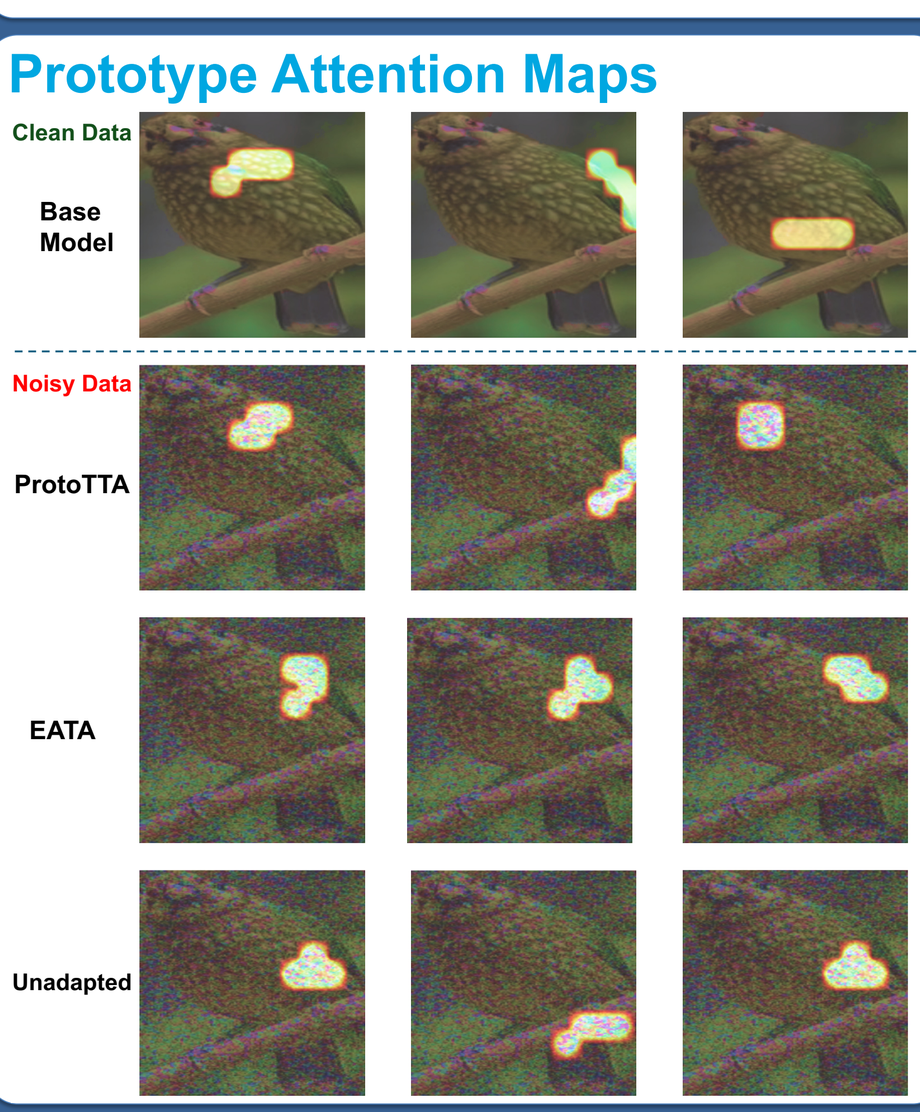
PAC (Prototype Activation Consistency) $\text{PAC} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{a}_i^{\text{clean}} \cdot \mathbf{a}_i^{\text{adapted}}}{\|\mathbf{a}_i^{\text{clean}}\|_2 \|\mathbf{a}_i^{\text{adapted}}\|_2}$

PCA-W (Weighted Prototype Alignment)

$$\text{PCA-W} = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{p\in\mathcal{T}_i} c_{i,p} \cdot \mathbb{1}[\text{class}(p) = y_i]}{\sum_{p\in\mathcal{T}_i} c_{i,p}}$$

- **ProtoTTA:** Source-free, highest PAC & PCA-W, low selection rate; restores correct semantic focus.
- **VLM Alignment:** Strong correlation between VLM scores and our interpretability metrics.
- **Blur Challenge:** Blur remains the hardest corruption for prototypical vision models.

Prototype Attention Maps

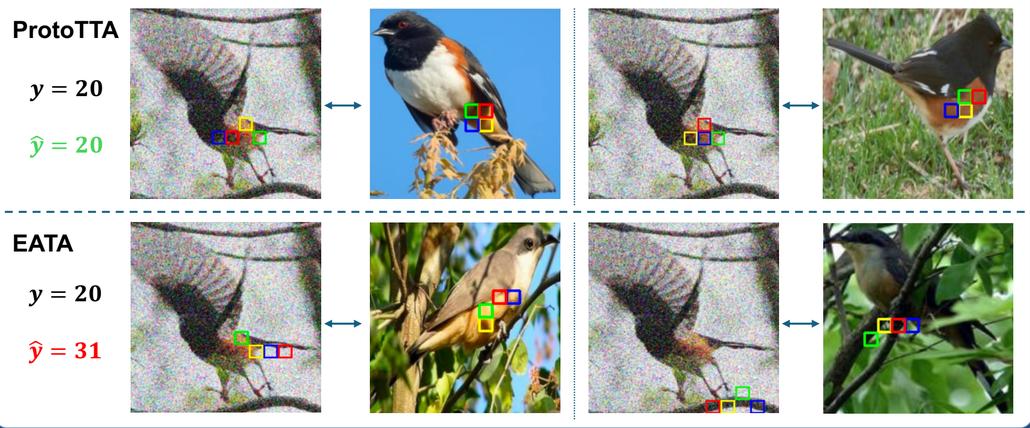


Results

Test-Time Adaptation — Mean Accuracy (%) ± std across all corruptions

Method	CUB-200-C ProtoViT · Vision	Stanford Dogs-C ProtoPFormer · Vision	SICAPv2-C ProtoPNet · Histopath.	Amazon-C ProtoLens · NLP	VLM Score Overall Quality ↑ (1–5, CUB-200-C)
Unadapted	51.9 ± 13.0	52.8 ± 10.6	31.6 ± 15.2	80.81 ± 7.43	3.53 ± 1.02
SAR	52.5 ± 12.8	42.9 ± 11.5	55.7 ± 4.5	77.65 ± 8.19	—
Tent	54.0 ± 12.8	57.1 ± 8.6	53.8 ± 5.7	80.08 ± 7.80	3.59 ± 1.02
EATA	58.9 ± 10.8	57.2 ± 8.4	55.6 ± 4.7	80.84 ± 6.91	3.75 ± 0.94
ProtoTTA (+) Ours	**60.1 ± 10.6**	**57.7 ± 8.4**	**55.8 ± 4.9**	**81.33 ± 7.45**	**3.78 ± 0.97**

Semantic Focus Restoration



Code

Paper