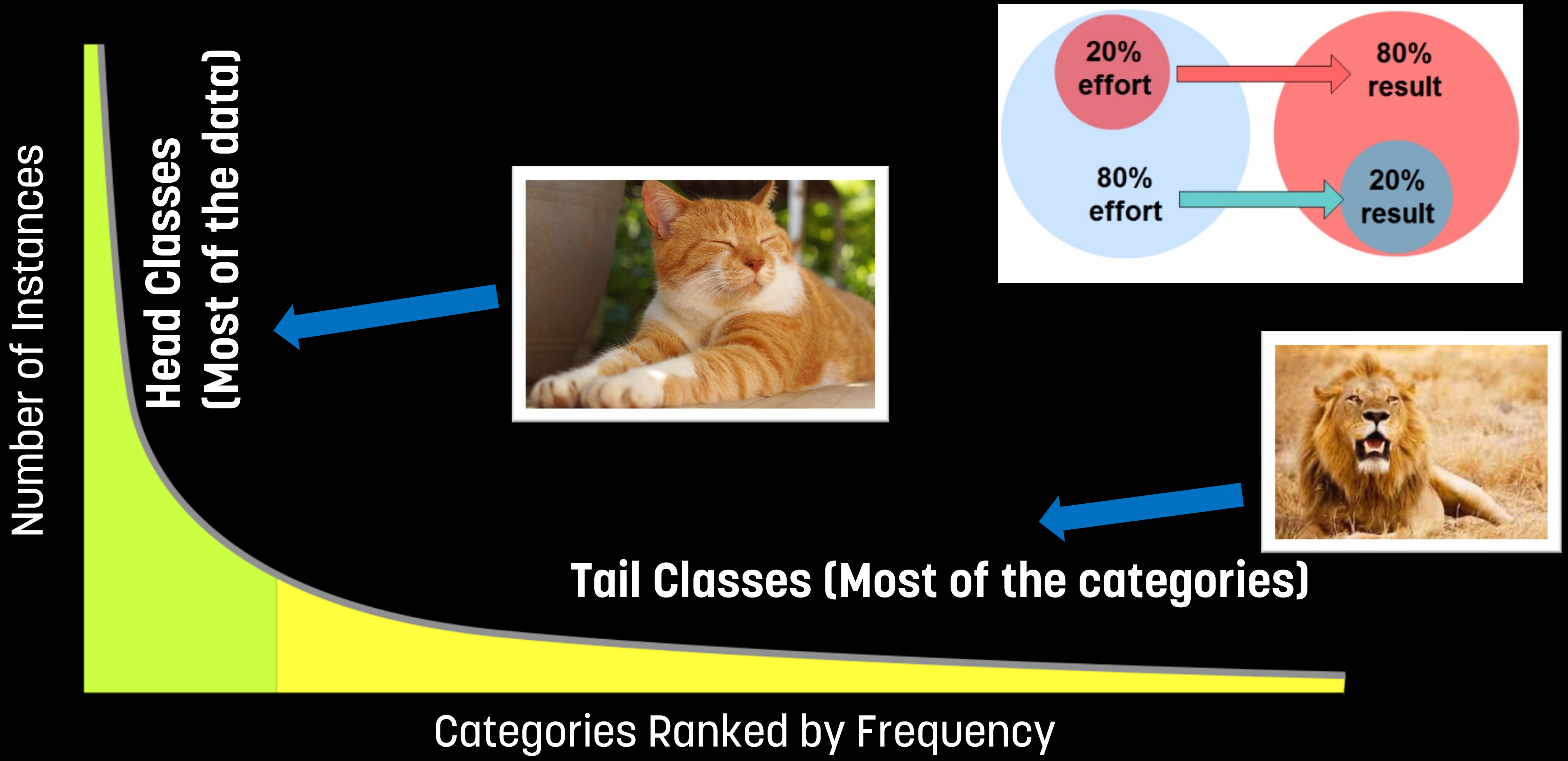
Flexible Distribution Alignment: Towards Long-Tailed Semi-supervised Learning with Proper Calibration

Emanuel Sánchez Aimar
Computer Vision Laboratory, Linköping University

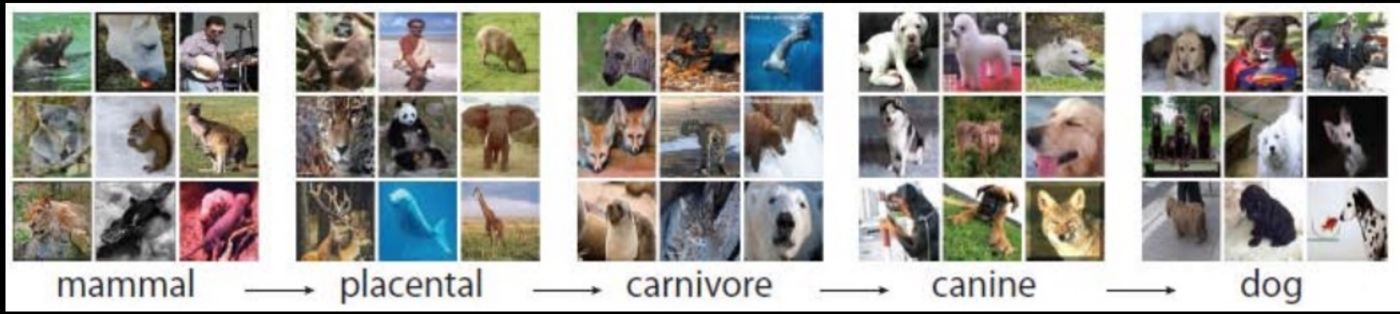
16-01-2025

What is long-tailed recognition?

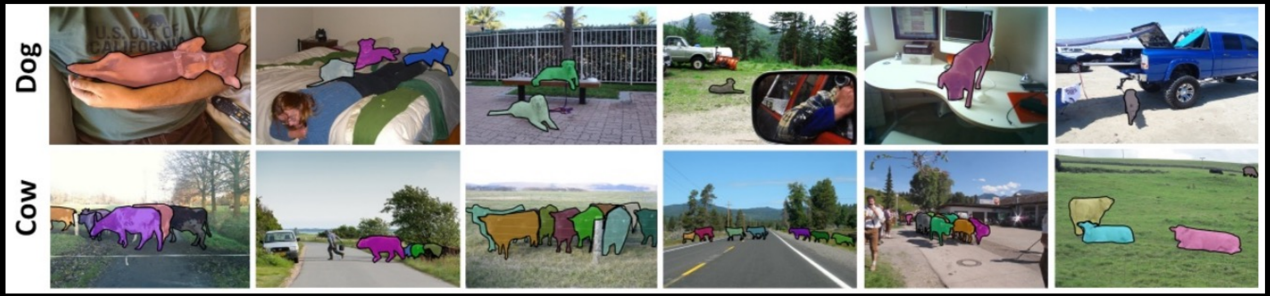


Common setting in “realistic” Computer Vision

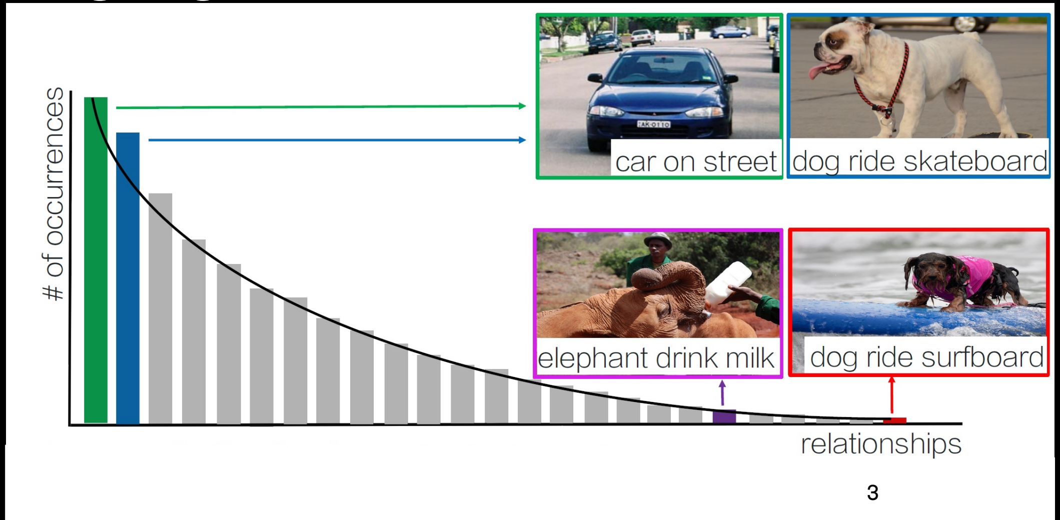
Classification (Hierarchical)



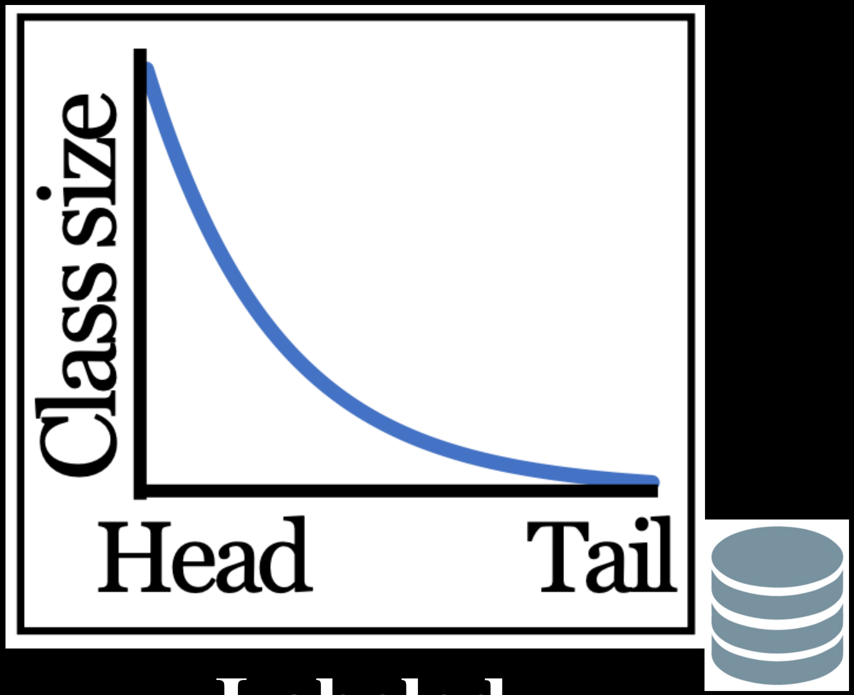
Object Detection & Segmentation



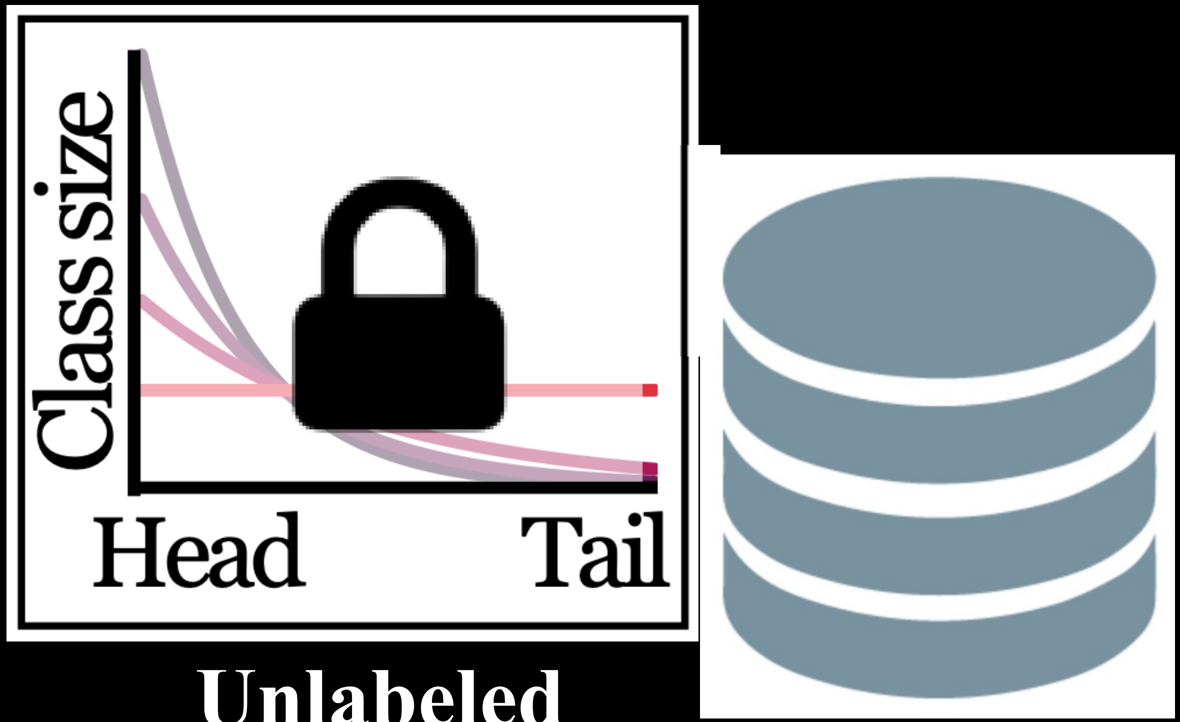
Language, Scene Graphs (Compositional)



Long-tailed Semi-supervised Learning (LTSSL)



Labeled

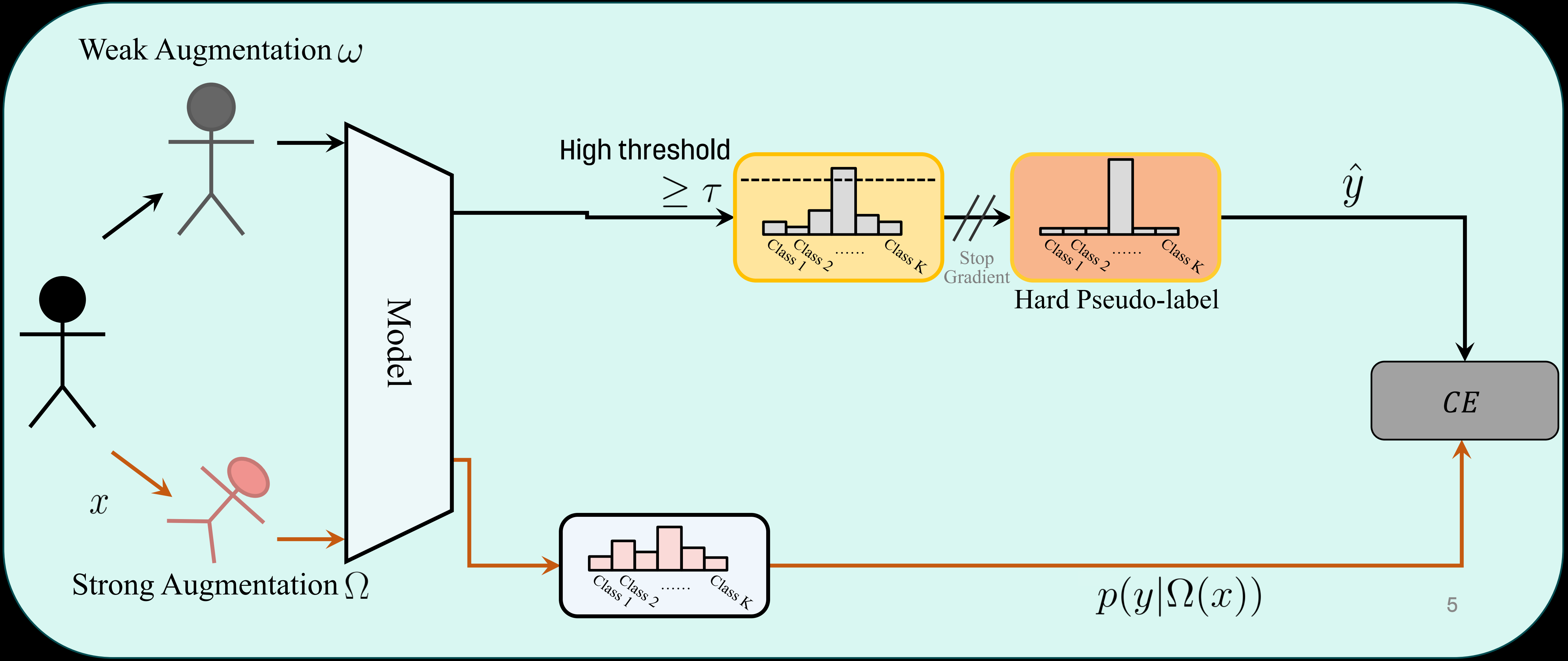


Unlabeled

FixMatch (Sohn'20)

Under the long-tailed setting:

- Many "tail" samples are discarded
- Biased classifier (exacerbated by pseudo-labels)



Distribution Alignment for (LT)SSL

- Aligns pseudo-labels with predefined priors (e.g. **uniform or labeled class distribution)**
  - Pseudo-label correction (Berthelot'20, Wei'21, Wang'22)
  - Classifier debiasing in the loss function (Wang'22, Lazarow'23)

- **Unrealistic/inaccurate assumptions?** The wrong prior can lead to:
  - **Inefficient** use of unlabeled data during training
  - **Biased classifier** during inference
  - **Poorly-calibrated** probabilities

Key observation

- What is the best classifier for the **(unknown) unlabeled distribution** $Q(y|x)$ under ***label shift***?

$$y = \arg\max_y \mathcal{Q}(y|x) = \arg\max_y \mathcal{Q}(x|y) \cdot \mathcal{Q}(y) = \arg\max_y \frac{\mathcal{P}_L(y|x)}{\mathcal{P}_L(y)} \cdot \mathcal{Q}(y)$$

Data Priors

Desired Priors

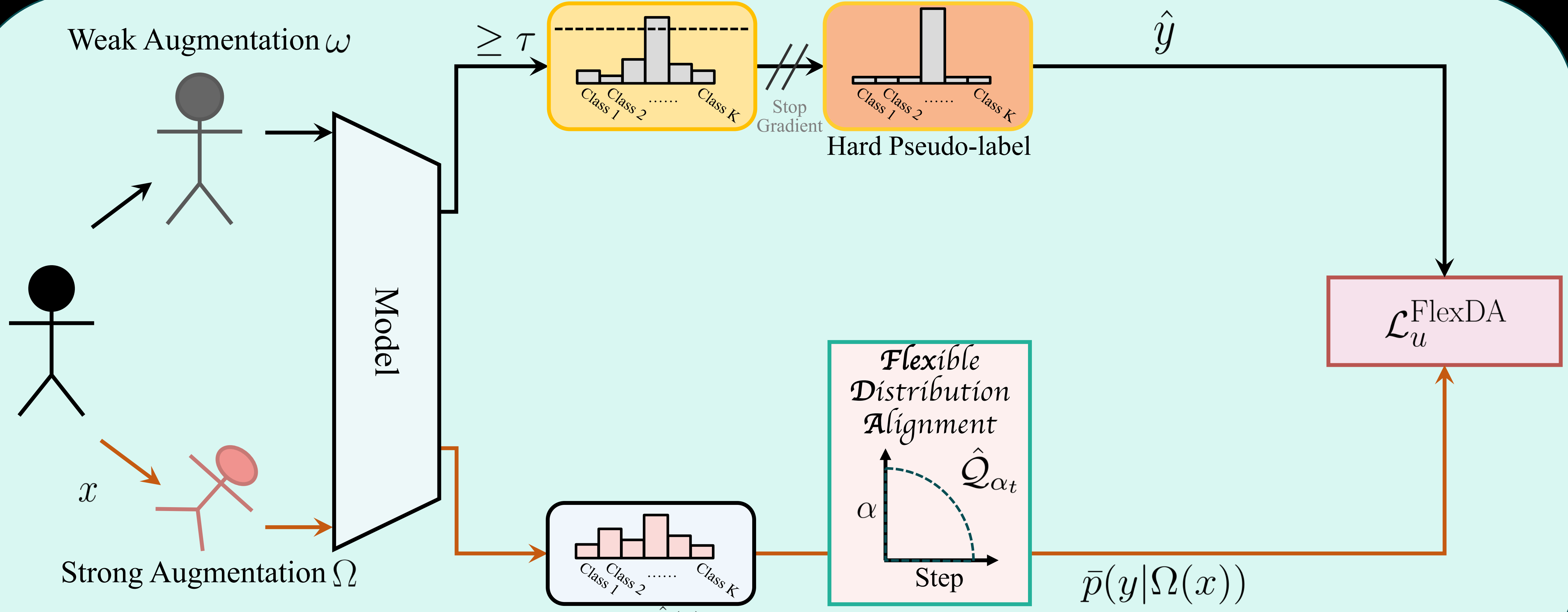
- And for test time (balanced/fairness)?

$$y = \arg\max_y \mathcal{P}_{\text{bal}}(y|x) = \arg\max_y \mathcal{P}_{\text{bal}}(x|y) \cdot \mathcal{P}_{\text{bal}}(y) = \arg\max_y \frac{\mathcal{Q}(y|x)}{\mathcal{Q}(y)} \cdot \frac{1}{K}$$

=> Trade-off between training and inference requirements

ADELLO: <u>**A**</u>lign and <u>**D**</u>istill <u>**E**</u>verything A<u>**ll**</u> at <u>**O**</u>nce

ADELLO: Align (ECCV'24)



$$\hat{\mathcal{Q}}_{\alpha_t}(y) = \frac{\hat{\mathcal{Q}}(y)^{\alpha_t}}{\sum_j \hat{\mathcal{Q}}(j)^{\alpha_t}}$$: estimated unlabeled class distribution (smoothed)

Flexible Distribution Alignment (FlexDA)

Supervised loss:

$$\mathcal{L}_s^{\text{FlexDA}} = \frac{1}{B}\sum_{b=1}^{B} \mathcal{H}(y_b, \sigma(f(\omega(x_b))\boxed{+ \log \frac{\mathcal{P}_L}{\hat{Q}_{\alpha_t}}}))$$

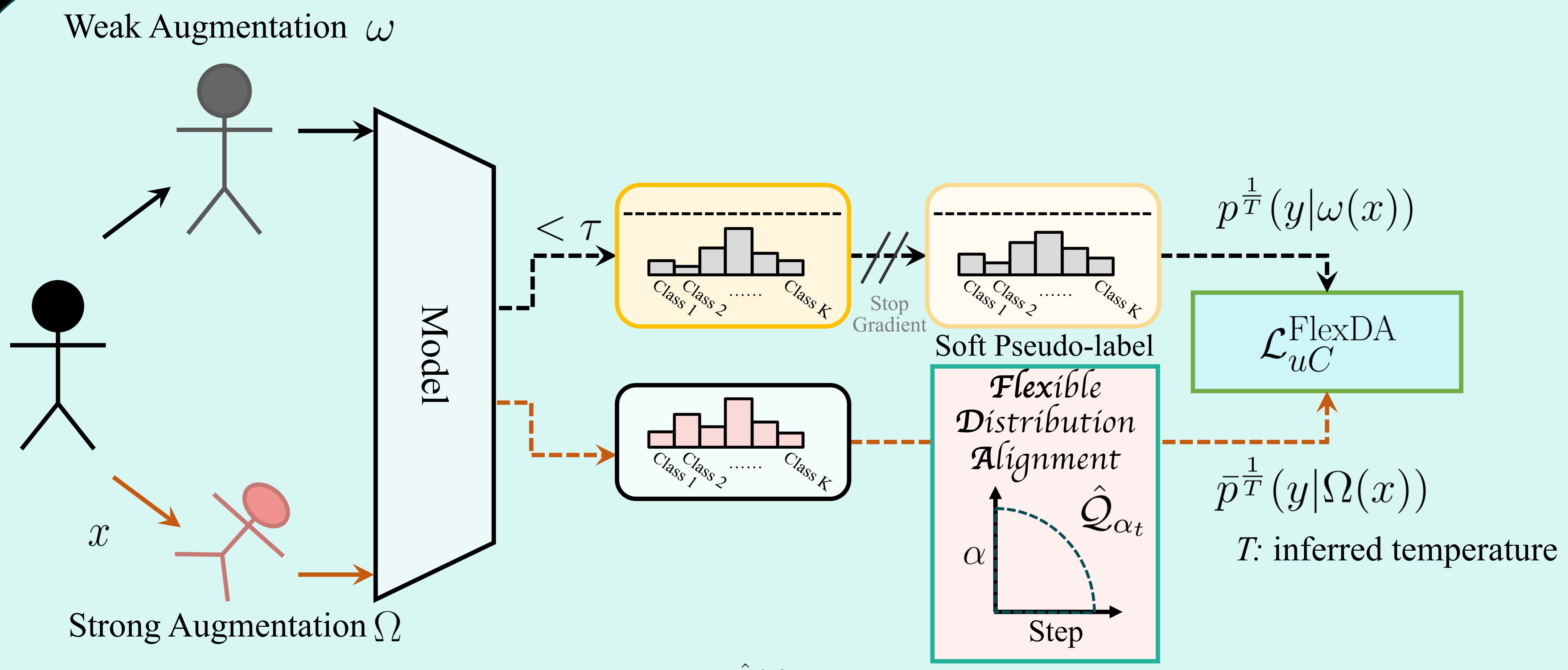
Logit Adjustments

Consistency loss:

$$\mathcal{L}_u^{\text{FlexDA}} = \frac{1}{\mu B}\sum_{b=1}^{\mu B} \boxed{\mathcal{M}(u_b)} \cdot \mathcal{H}(\hat{y}_b, \sigma(f(\Omega(u_b))\boxed{+ \log \frac{\hat{Q}}{\hat{Q}_{\alpha_t}}}))$$

Hard PLs Mask

ADELLO: Align and Distill (ECCV'24)



$\hat{\mathcal{Q}}_{\alpha_t}(y) = \frac{\hat{\mathcal{Q}}(y)^{\alpha_t}}{\sum_j \hat{\mathcal{Q}}(j)^{\alpha_t}}$: estimated unlabeled class distribution (smoothed)

FlexDA + Complementary Consistency Regularization

Complementary Consistency loss:

$$\mathcal{L}_{uC}^{\text{FlexDA}} = \frac{1}{\mu B}\sum_{b=1}^{\mu B} \boxed{\mathcal{M}^C(u_b)} \cdot \mathcal{H}(p^{\frac{1}{T}}(y|\omega(u_b)), \underline{p^{\frac{1}{T}}(y|\Omega(u_b))})$$

Soft PLs Mask

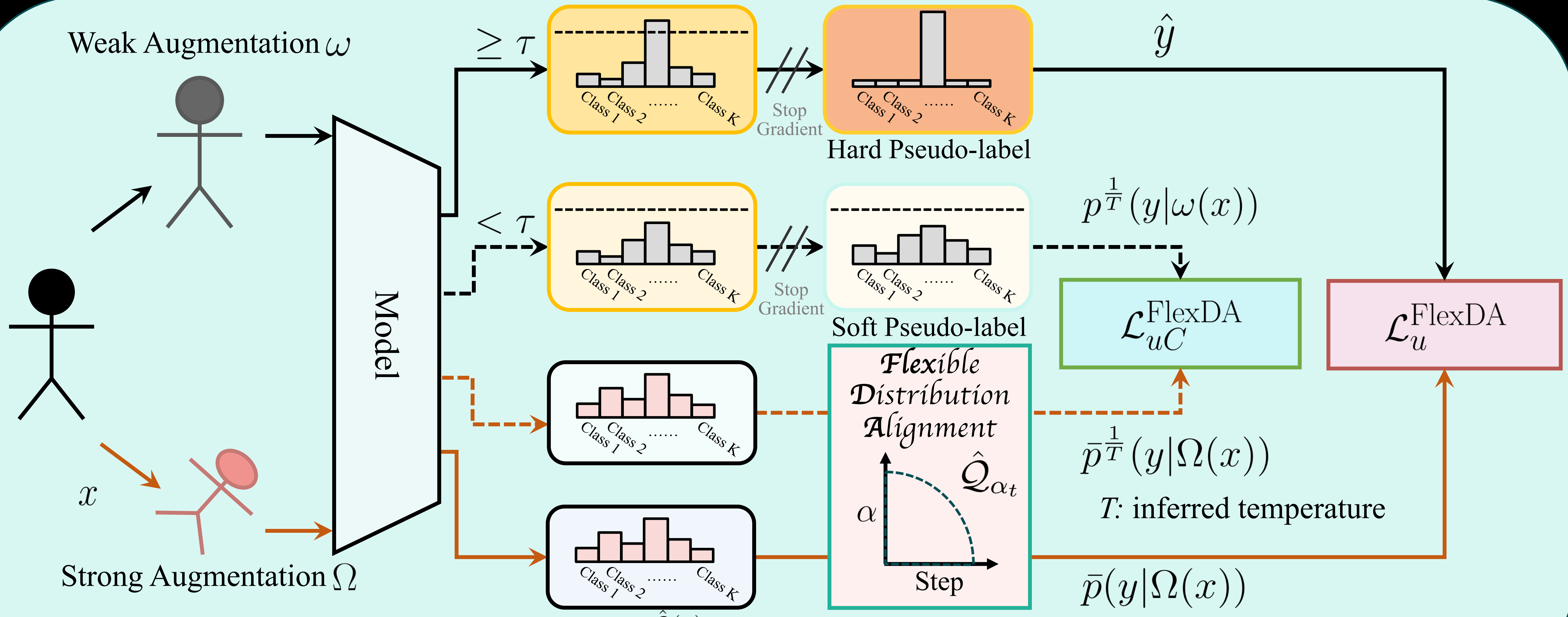
where $\underline{\bar{p}^{\frac{1}{T}}(y|\Omega(u_b))} = \sigma(\frac{1}{T}(f(\Omega(u_b)) \boxed{+ \log \frac{\hat{Q}}{\hat{Q}_{\alpha_t}}}))$

Logit Adjustments

Imbalance-aware temperature (inferred after warmup):

$$T = \exp(\mathrm{KL}(\mathcal{P}_{\text{bal}} \| \hat{Q}))$$

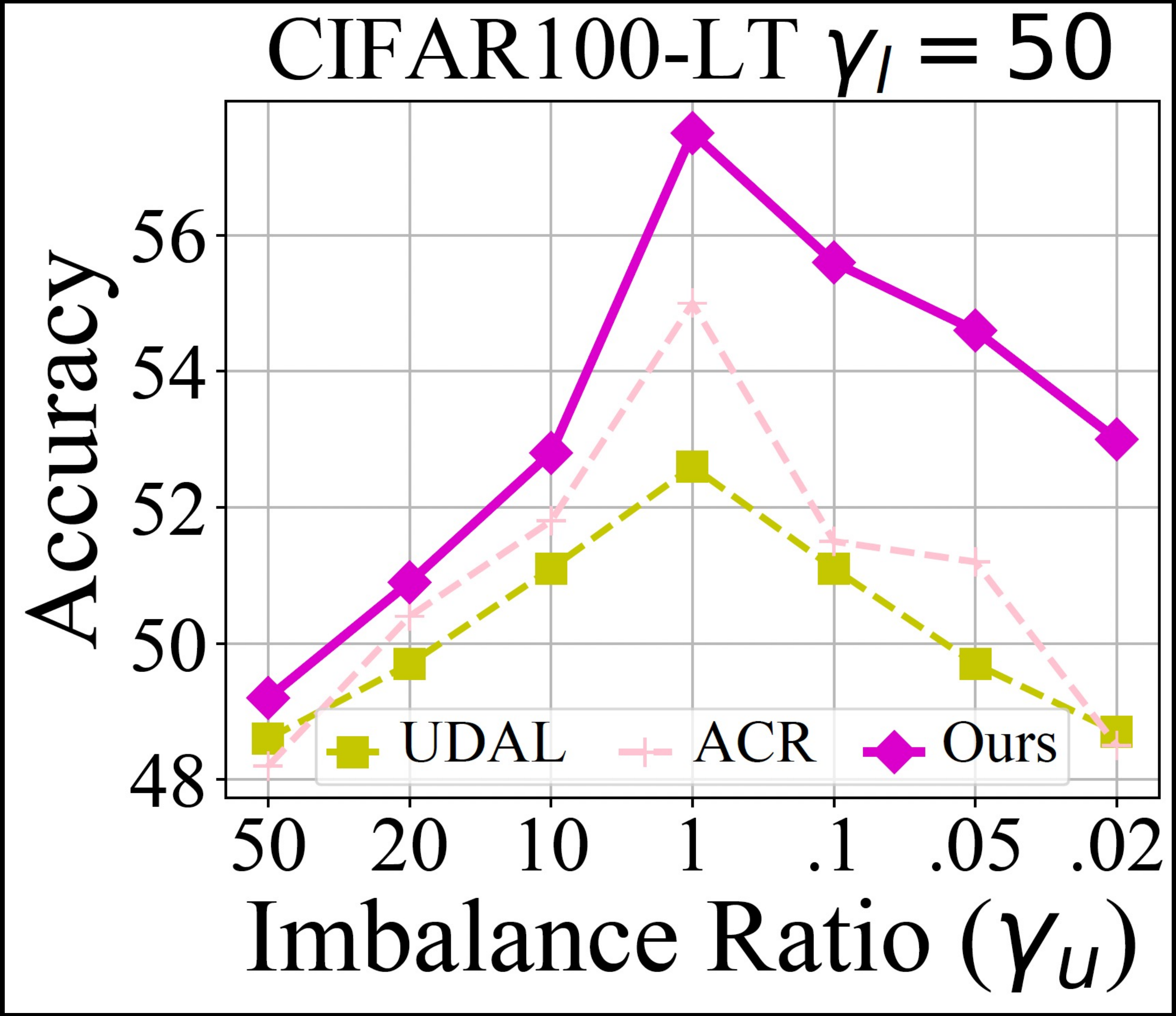
ADELLO: Align and Distill Everything All at Once (ECCV'24)



$\hat{\mathcal{Q}}_{\alpha_t}(y) = \frac{\hat{\mathcal{Q}}(y)^{\alpha_t}}{\sum_j \hat{\mathcal{Q}}(j)^{\alpha_t}}$: estimated unlabeled class distribution (smoothed)

Experimental Results

Robustness under distribution mismatch



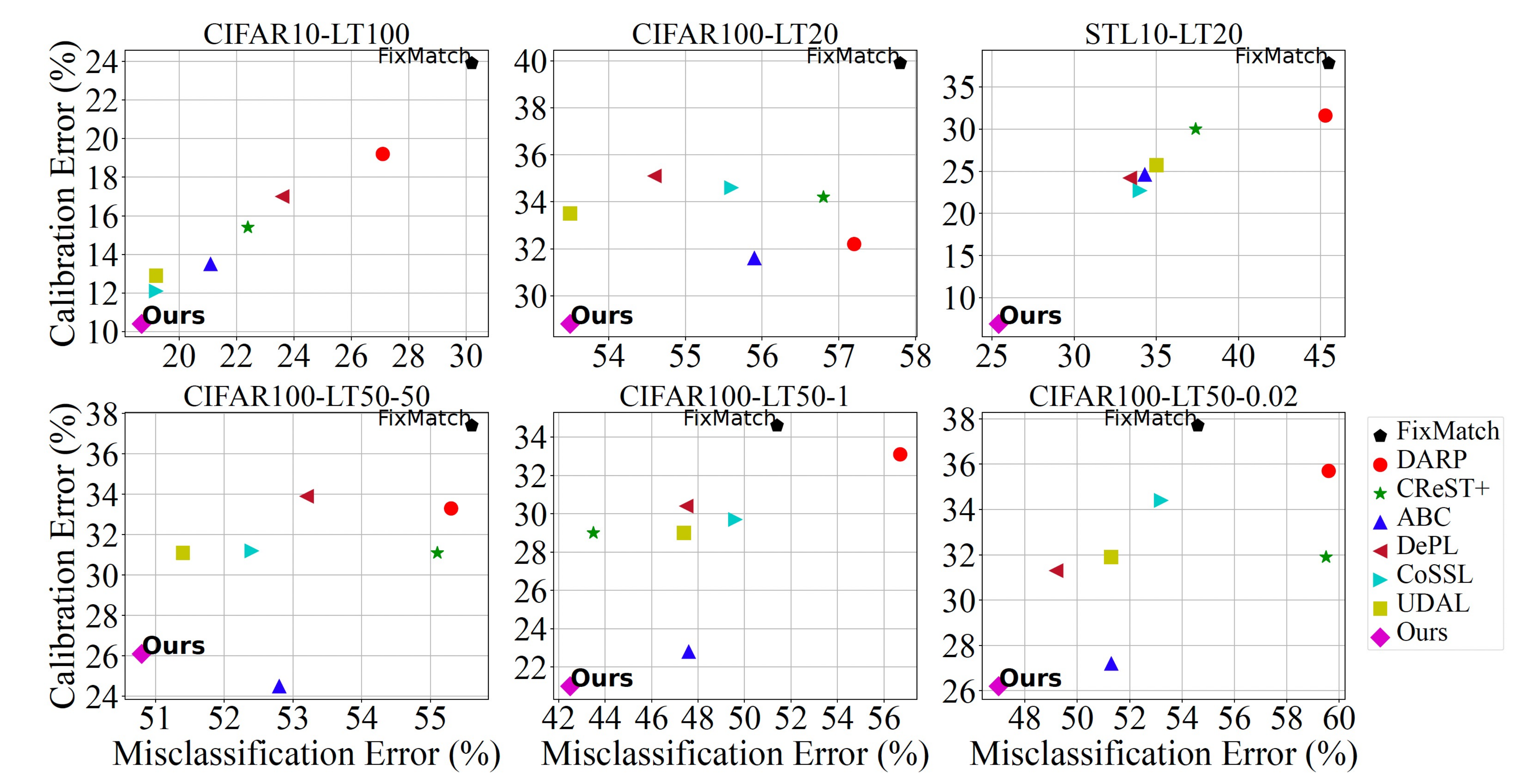
SOTA performance under consistent case

ImageNet127

Bal. Accuracy Resolution

Method	32×32	64×64
FixMatch [58][†]	29.7	42.3
+DARP [29][†]	30.5	42.5
+DARP +cRT [29][†]	39.7	51.0
+CReST+ [68][†]	32.5	44.7
+CReST+ +LA [68][†]	40.9	<u>55.9</u>
+CoSSL [16][†]	43.7	53.8
+UDAL ($\alpha_{\min}$=0.55) [37]	40.2	49.4
+UDAL ($\alpha_{\min}$=0.1) [37]	<u>44.1</u>	52.3
+ADELLO (ours)	**47.5**	**58.0**

Best accuracy-calibration performance trade-off!



ADELLO: Align and Distill Everything All at Once

- Strong SSL classifier
- Theoretically-sound
- Robust under class imbalance
- Robust under distribution mismatch
- Simple and end-to-end trainable

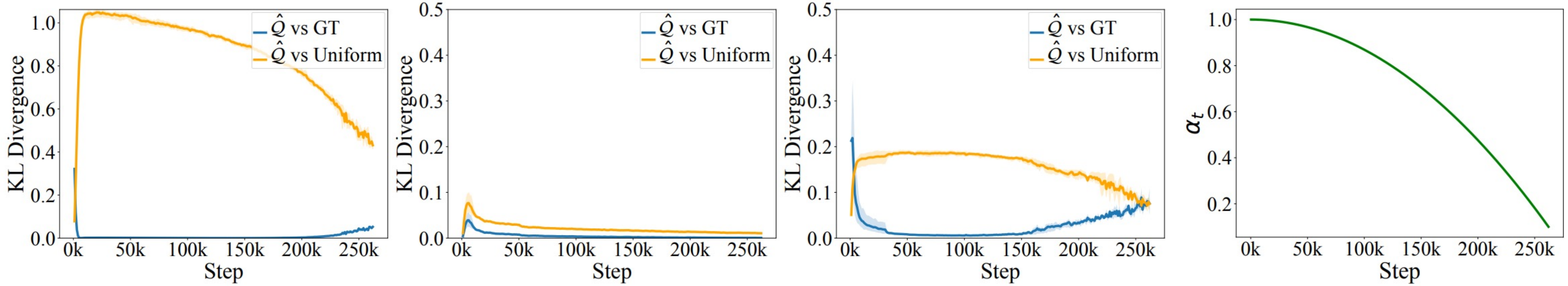
Thank you for listening!

Paper

Code

Extra Slides

Robust prior estimation



(a) KL-div for *forward* case **(b)** KL-div for *balanced* case **(c)** KL-div for *reversed* case **(d)** α_t $(d = 2, \alpha_{\text{min}} = 0.1)$

Fig. 3: Prior estimation under label shift. A comparison of KL divergence shows 1) a small difference between the estimated prior, $\hat{\mathcal{Q}}$, and the ground-truth prior, $\mathcal{Q}$, during most of the training (blue curve), and 2) a larger disparity between $\hat{\mathcal{Q}}$ and the uniform prior, $\mathcal{P}_{\text{bal}}$, (orange curve). The progression of a quadratic scheduler ($d = 2$) is shown in (d) (green curve). Label shift settings: (a) forward, (b) balanced, and (c) reversed long-tailed, computed for CIFAR10-LT100.