Mobility Helps Learning: Unsupervised Model Adaptation for Object Recognition via Movement



YananMa[1], Yihang Tao[1], Zhengru Fang[1], Senkang Hu[1], Xianhao Chen[2], and Yuguang Fang[1,*]
[1]Department of Computer Science, City University of Hong Kong
yananma8-c@my.cityu.edu.hk, yihang.tommy@my.cityu.edu.hk, zhefang4-c@my.cityu.edu.hk,
senkang.forest@my.cityu.edu.hk, my.fang@cityu.edu.hk
[2]Department of Electrical and Computer Engineering, The University of Hong Kong
xchen@eee.hku.hk

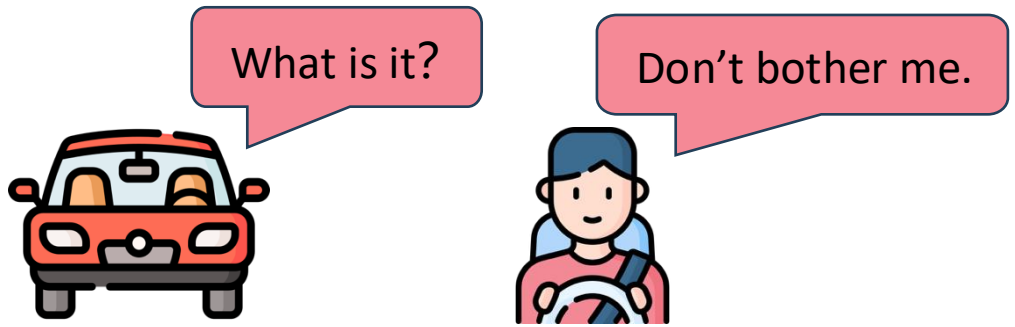
Corresponding author

Mobility Helps Learning: Background

- When encountering new data from local environments, agents face substantial real-world generalization challenges. It is imperative to equip agents with autonomous learning (training) capabilities to ensure adaptability while minimizing human supervision.



Uber's self-driving vehicle caused a fatal accident since it classified a pedestrian as an “other object”



Obtaining sufficient human-annotated labels can often be challenging at the edge!

Mobility Helps Learning: Background

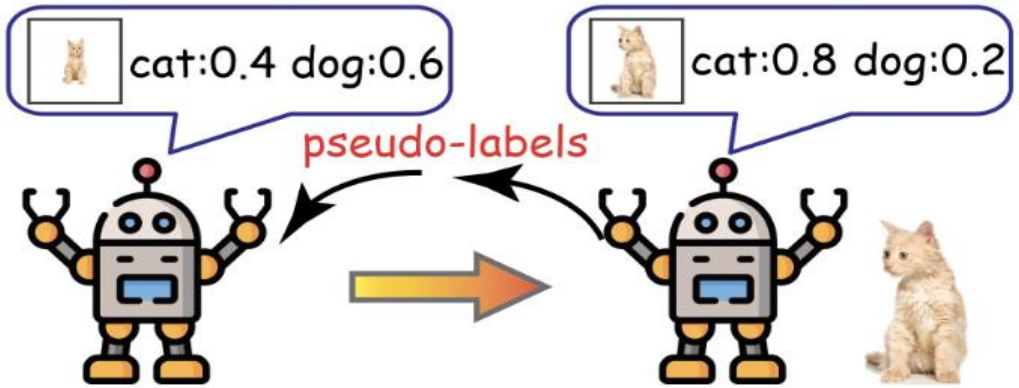


- How can we adapt the model **without** *human annotations*?

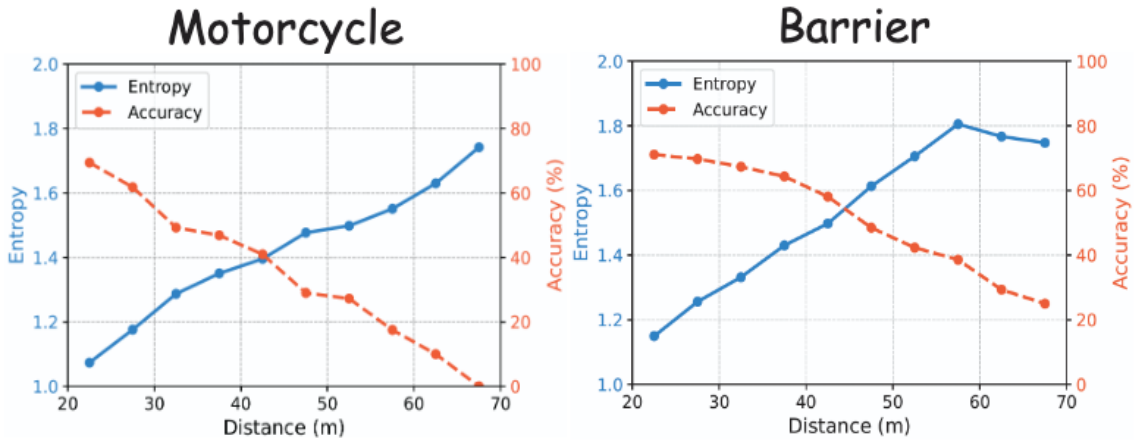
Mobility Helps Learning: Motivation

"We must perceive in order to move, but we must also move in order to perceive."

--- J.J Gibson



(a) Illustration of mobility helps learning.



(b) Accuracy and entropy vs. distance.

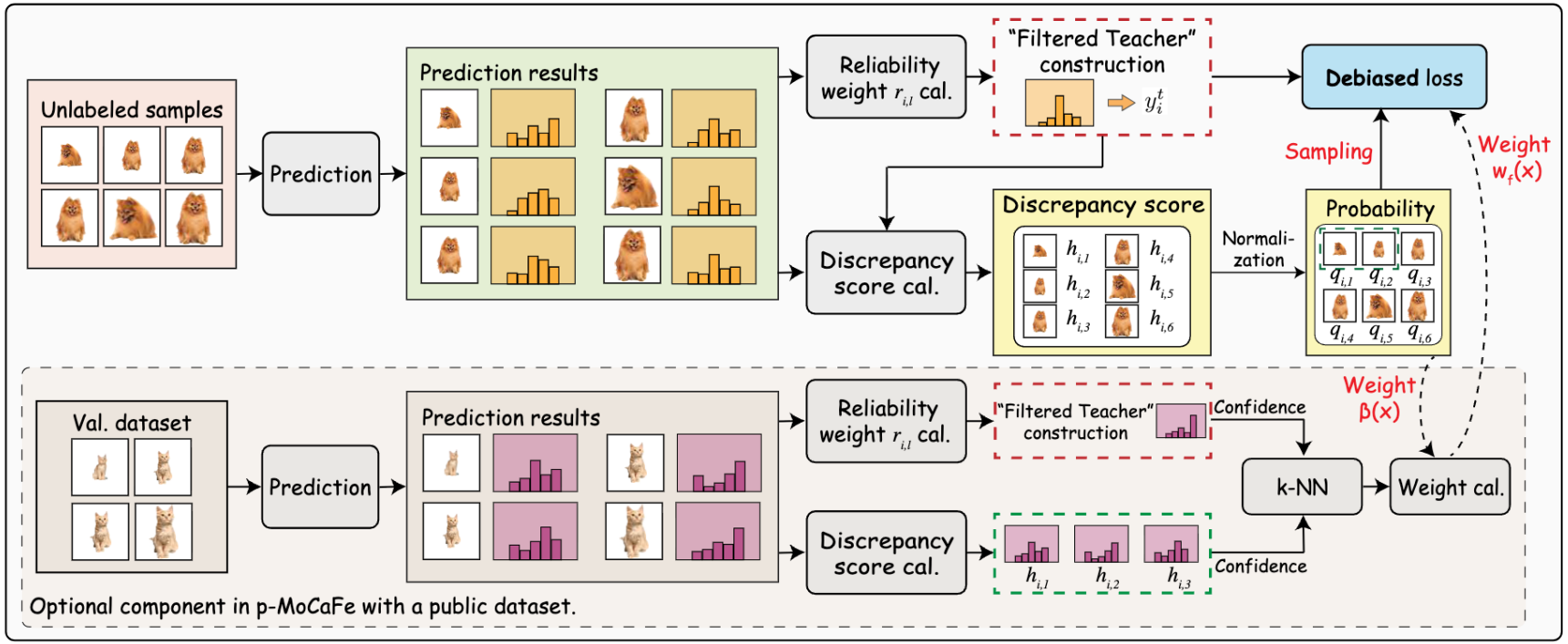
- The high-confidence predictions (on the right) can serve as pseudo labels to guide model adaptation over samples with lower confidence (on the left).
- Average entropy and classification accuracy versus the distance for two classes of objects, i.e., barriers and motorcycles, in the nuScenes dataset with a pre-trained ResNet-50 model.

Mobility Helps Learning: Challenges & Diagram



Our idea: during mobility, "good" observations can serve as pseudo labels for model adaptation.

Challenges: highly influenced by pseudo-label quality.



Pseudo-Label Construction



- How can we construct good pseudo-labels?

❑ Reliability-Weighted Pseudo-Label Construction

❖ **Assumption 1.** We assume that the logits predictions for the same object i across different views can be modeled as noisy observations of a latent semantic state. Specifically, the logit vector $\mathbf{z}_{i,l} \in R^C$ for frame l is formulated as

$$\mathbf{z}_{i,l} = \mathbf{s}_i + \boldsymbol{\epsilon}_{i,l}, \quad \boldsymbol{\epsilon}_{i,l} \sim \mathcal{N}\left(\mathbf{b}_{i,l}, \sigma^2\left(d_{i,l}, u_{i,l}\right)\mathbf{I}\right),$$

where $\mathbf{s}_i$ is the true semantic state and $\boldsymbol{\epsilon}_{i,l}$ is noise. The mean of the noise $\mathbf{b}_{i,l}$ captures prediction bias, while its variance $\sigma^2\left(d_{i,l}, u_{i,l}\right)$ is conditioned on a mobility factor $d_{i,l}$ (e.g., distance) and an uncertainty score $\mathrm{u}_{i,l}$.

- Information filter-based sequence-level pseudo-label construction

$$y_i^t = \arg\max_{c\in\{1,\ldots,C\}} [\boldsymbol{\mu}_i]_c \qquad \boldsymbol{\mu}_i = \frac{\sum_{l=1}^{L_i} r_{i,l}\mathbf{z}_{i,l}}{\sum_{l=1}^{L_i} r_{i,l}}$$

Mobility Helps Learning: Theoretical Analysis

❖ **Lemma 1.** The stochastic posterior variance $\Sigma_L = \left(\sum_{l=1}^{L} \sigma_l^{-2}\right)^{-1}$ of the filtered prediction is lower than that of any individual observation: $\Sigma_L < \min_{1 \le l \le L} \sigma_l^2$.

❖ **Proposition 1.** Assume $\mathbf{b}_l = \mathbf{0}$, the misclassification upper bound of the filtered pseudo-label y^t is lower than that of the single best prediction y_{best}:

$$P(y^t \neq c^\star) \le \sum_{c \neq c^\star} \exp\left(-\frac{\Delta_{0,c}^2}{4\Sigma_L}\right) < \sum_{c \neq c^\star} \exp\left(-\frac{\Delta_{0,c}^2}{4\sigma_{best}^2}\right),$$

where $c^\star$ is the ground-truth class, and y_{best} denotes the prediction from the single best observation corresponding to the minimum variance $\sigma_{best}^2 = \min_{1 \le l \le L} \sigma_l^2$.

Informative Sampling and Weighting



- How can we select the most informative samples and ensure unbiased learning?

Informative Sampling

- Discrepancy score between the filtered pseudo-label and the individual prediction

$$h_{i,l} = l\left(y_i^t, f\left(x_{i,l}\right)\right),$$

where ℓ denotes the loss (e.g., cross-entropy).

- We then construct sampling probabilities

$$q_{i,l} = \frac{h_{i,l} + \eta}{\sum_{i'=1}^{N} \sum_{\tau=1}^{L_{i'}} \left(h_{i',\tau} + \eta\right)}, \quad \forall i, \forall l,$$

Accordingly, we select *m* samples from the dataset, forming the selected dataset $\mathcal{X}_m$.

Frames that ``disagree'' with the filtered pseudo-label

Weighting for Debiasing

- Calculating weights $w_f(x)$

$$w_f(x) = \frac{\beta(x)}{1 + p(x)\left(d_f(x) - 1\right)}, \quad \text{with } d_f(x) = \frac{\ell\left(f_{\text{false}}(x), f(x)\right)}{\ell\left(y, f(x)\right)}.$$

- The weighted empirical risk

$$R^w(f) = \frac{1}{m} \sum_{x \in \mathcal{X}_m} w_f(x) \ell(y^t(x), f(x)).$$

Weights calculation

- Weights $\hat{p}(x)$ and $\hat{d}_f(x)$ are estimated via the *k*-NN method using a public dataset.
- The *k*-NN approach introduces approximation errors.

Mobility Helps Learning: Theoretical Analysis

❖ **Proposition 2.** Assume that $\mathcal{X}_m$ is drawn from the induced distribution $\tilde{Q}_t$, then we have $E[R^w(f)] = R(f)$.

Generalize (estimation error)

❖ **Proposition 3.** The weighted empirical risk estimator, given by $\hat{R}^{\tilde{w}}(f) = \frac{1}{m}\sum_{x\in\mathcal{X}_m} \tilde{w}_f(x)\ell\big(y^t(x), f(x)\big)$ satisfies:

$$\left|E\left[\hat{R}^{\tilde{w}}(f)\right] - R(f)\right| \le E_{x\sim\tilde{Q}_t}\left[\tilde{w}_f(x)\ell\big(y, f(x)\big)\left|p(x)\big(d_f(x)-1\big) - \hat{p}(x)\big(\hat{d}_f(x)-1\big)\right|\right] + E_{x\sim\tilde{Q}_t}\left[\left|\tilde{w}_f(x) - \hat{w}_f(x)\right|\left[1+\hat{p}(x)\big(\hat{d}_f(x)-1\big)\right]\ell\big(y, f(x)\big)\right].$$

Mobility Helps Learning: Numerical Results

Table 1: Performance in the Boston Seaport scenario on the nuScenes dataset under varying pre-trained model accuracy (i.e., 75%, 70%, and 65%). The value in the bracket represents the standard deviation.

	ResNet-50			Swin-T		
Pre-trained	75%	70%	65%	75%	70%	65%
w/o MoCaFe	76.18↑(± 0.28)	72.05↑(± 0.28)	62.17↓(± 0.37)	76.69↑(± 0.26)	72.55↑(± 0.25)	40.23↓(± 0.43)
AdaContrast	77.88↑(± 0.19)	73.63↑(± 0.15)	66.51↑(± 0.18)	76.85↑(± 0.33)	72.83↑(± 0.47)	62.28↓(± 0.61)
PLUE	77.94↑(± 0.12)	73.82↑(± 0.21)	67.28↑(± 0.29)	77.43↑(± 0.24)	73.55↑(± 0.30)	59.91↓(± 0.35)
RoTTA	78.67↑(± 0.21)	75.13↑(± 0.33)	68.20↑(± 0.31)	—	—	—
C-SFDA	78.71↑(± 0.22)	75.37↑(± 0.26)	68.34↑(± 0.34)	77.75↑(± 0.18)	72.20↑(± 0.19)	66.43↑(± 0.25)
UPA	76.87↑(± 0.36)	74.13↑(± 0.34)	57.60↓(± 0.48)	78.24↑(± 0.26)	73.15↑(± 0.23)	68.12↑(± 0.28)
AnCon	79.23↑(± 0.29)	75.61↑(± 0.36)	68.23↑(± 0.38)	80.38↑(± 0.27)	76.28↑(± 0.30)	70.59↑(± 0.32)
MoCaFe	80.81↑(± 0.23)	77.25↑(± 0.25)	69.32↑(± 0.24)	89.60↑(± 0.11)	89.34↑(± 0.15)	89.27↑(± 0.16)
p-MoCaFe	**81.23↑(± 0.13)**	**77.93↑(± 0.16)**	**69.74↑(± 0.25)**	**90.19↑(± 0.17)**	**90.07↑(± 0.18)**	**90.02↑(± 0.12)**
FSL	88.66↑(± 0.12)	87.83↑(± 0.18)	86.87↑(± 0.25)	92.30↑(± 0.06)	92.12↑(± 0.15)	91.98↑(± 0.28)

Mobility Helps Learning: Numerical Results

Table 2: Per-class test accuracy in the Boston Seaport scenario on the nuScenes dataset. Pre-trained accuracy for the Boston scenario is 70%. → represents that the performance remains the same with the pre-trained model.

					ResNet-50					
Class	Car	Pedestrian	Traffic-Cone	Trailer	Bus	Motorcycle	Barrier	Con-Vehicle	Bicycle	Truck
Pre-trained	82.42	77.35	70.46	77.22	62.86	61.11	53.09	49.64	48.15	23.28
w/o MoCaFe	87.41↑	81.05↑	76.95↑	75.32↓	61.90↓	47.22↓	52.31↓	49.68↑	46.91↓	18.84↓
AdaContrast	88.38↑	85.29↑	77.42↑	**81.00**↑	68.10↑	50.05↓	53.09→	40.53↓	37.50↓	14.64↓
PLUE	88.79↑	84.98↑	75.55↑	78.67↑	64.87↑	55.89↓	54.24↑	46.15↓	39.71↓	15.51↓
RoTTA	87.41↑	83.78↑	73.12↑	74.74↓	65.39↑	52.53↓	**58.33**↑	40.53↓	37.62↓	12.87↓
C-SFDA	87.96↑	84.05↑	78.24↑	72.50↓	62.86→	**62.58**↑	54.73↑	51.16↑	35.67↓	18.50↓
UPA	86.43↑	83.02↑	77.84↑	74.97↓	69.52↑	57.67↓	48.84↓	45.86↓	28.88↓	16.73↓
AnCon	88.91↑	86.31↑	81.42↑	78.67↑	68.10↑	51.48↓	54.24↑	47.22↓	42.54↓	16.73↓
MoCaFe	90.34↑	88.68↑	**83.85**↑	76.48↓	71.50↑	52.81↓	54.31↑	50.50↑	53.14↑	19.17↓
p-MoCaFe	**91.19**↑	**89.77**↑	83.78↑	77.90↑	**73.38**↑	61.18↑	56.85↑	**56.21**↑	**54.42**↑	**20.71**↓
					Swin-T					
Class	Car	Pedestrian	Traffic-Cone	Trailer	Bus	Motorcycle	Barrier	Con-Vehicle	Bicycle	Truck
Pre-trained	85.45	73.42	74.68	72.78	60.95	65.56	52.25	51.76	49.38	24.37
w/o MoCaFe	87.41↑	78.24↑	76.05↑	76.53↑	70.47↑	75.56↑	56.32↑	49.92↓	46.58↓	19.14↓
AdaContrast	89.72↑	80.02↑	78.10↑	73.90↑	66.83↑	73.92↑	50.37↓	45.96↓	48.65↓	15.13↓
PLUE	90.08↑	78.10↑	78.26↑	76.53↑	64.43↑	68.37↑	54.18↑	47.31↓	49.19↓	16.73↓
C-SFDA	90.30↑	81.23↑	77.41↑	74.55↑	65.95↑	66.83↑	55.83↑	45.66↓	36.59↓	18.96↓
UPA	88.90↑	80.20↑	80.61↑	71.65↓	69.48↑	58.23↓	58.59↑	46.46↓	38.90↓	17.27↓
AnCon	92.80↑	84.81↑	82.47↑	75.83↑	67.16↑	62.64↓	52.68↑	43.26↓	39.18↓	17.36↓
MoCaFe	93.73↑	**91.74**↑	94.00↑	90.75↑	**90.69**↑	87.16↑	85.42↑	86.23↑	**76.12**↑	35.60↑
p-MoCaFe	**94.20**↑	91.67↑	**94.03**↑	**92.74**↑	90.13↑	**88.31**↑	**86.88**↑	**87.19**↑	75.41↑	**40.60**↑

Key Takeaways

- **Movement itself can serve as source of supervision.** Movement provides multiple predictions with varying quality for the same object across distances and viewpoints.
- We leverage *motion-induced* predictions into reliable pseudo-labels and learn from informative frames. The optional p-MoCaFe extension uses public data to yield a bounded-bias estimator of the clean loss.
- The approach delivers strong, robust adaptation. Experiments on nuScenes, KITTI, and embodied-agent datasets demonstrate gains across architectures compared to existing approaches.
- Our findings can enable autonomous systems to adapt more effectively to unfamiliar environments.

香港城市大學
City University of Hong Kong

Thank You!



YananMa[1], Yihang Tao[1], Zhengru Fang[1], Senkang Hu[1], Xianhao Chen[2], and Yuguang Fang[1,*]
[1]Department of Computer Science, City University of Hong Kong
yananma8-c@my.cityu.edu.hk, yihang.tommy@my.cityu.edu.hk, zhefang4-c@my.cityu.edu.hk,
senkang.forest@my.cityu.edu.hk, my.fang@cityu.edu.hk
[2]Department of Electrical and Computer Engineering, The University of Hong Kong
xchen@eee.hku.hk