香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

Trustworthy Deep Learning from Open-set Corrupted Data

PI : **Dr. HAN Bo**

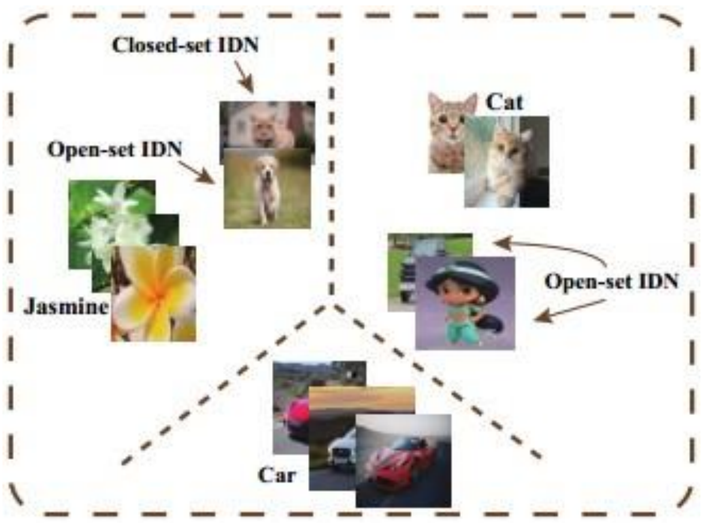
Funding Scheme: **Early CAREER Scheme**
Project Ref. No.: **22200720**
Amount Awarded (to HKBU): **HK$ 534,288**
Project Period: **Sep 2020 - Aug 2022**

OBJECTIVES

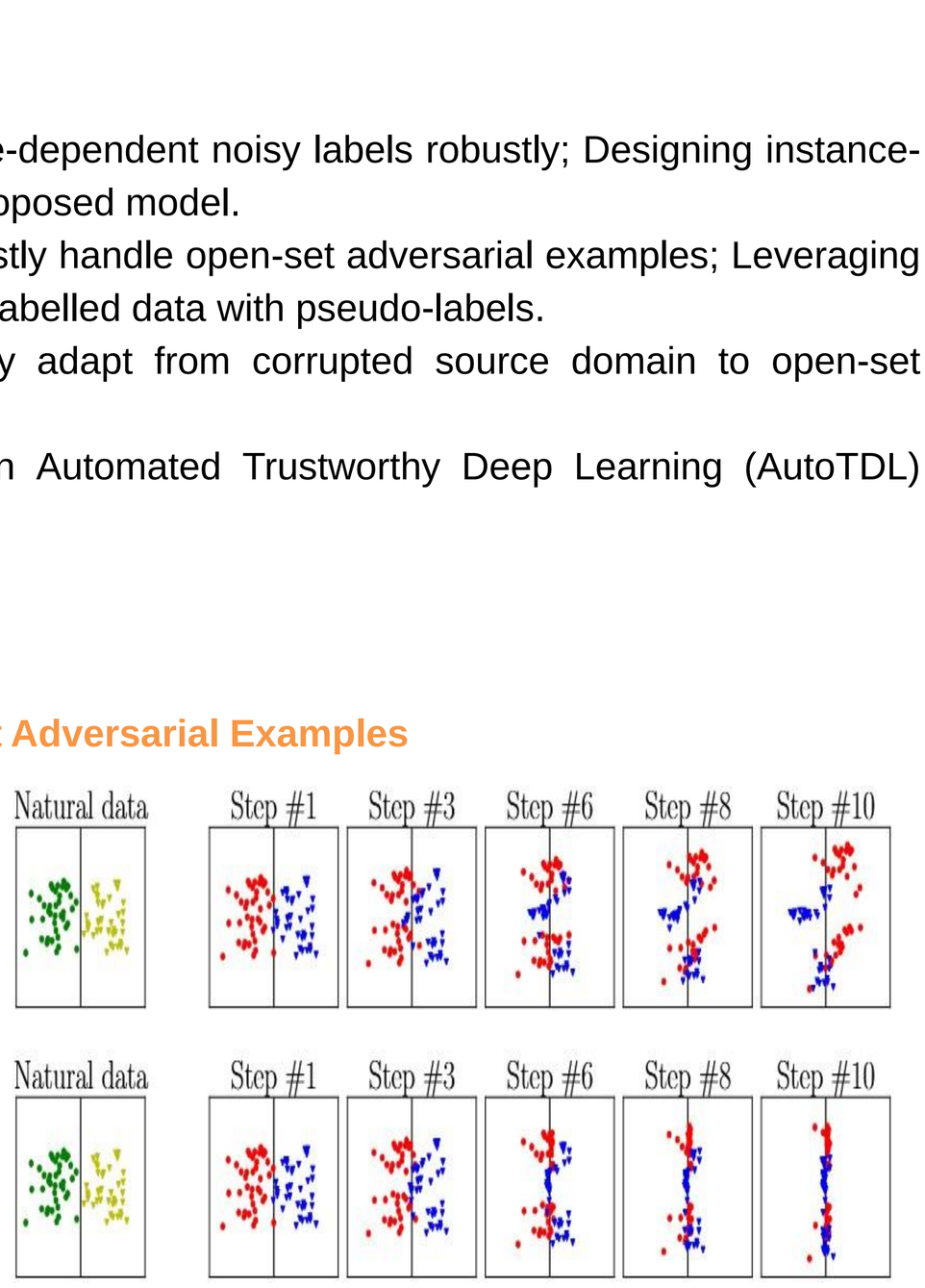
1. Developing a dual-scored methodology to model open-set instance-dependent noisy labels robustly; Designing instance-level learning algorithms with theoretical guarantees to solve the proposed model.
2. Exploiting generalized unlabelled data as auxiliary medium to robustly handle open-set adversarial examples; Leveraging adversarial robust loss to jointly train on original training set and unlabelled data with pseudo-labels.
3. Designing an adversarial dual checking methodology to robustly adapt from corrupted source domain to open-set unlabelled target domain.
4. Automating and integrating above orthogonal techniques into an Automated Trustworthy Deep Learning (AutoTDL) system; Testing this system using real-world corrupted data.

HIGHLIGHTS

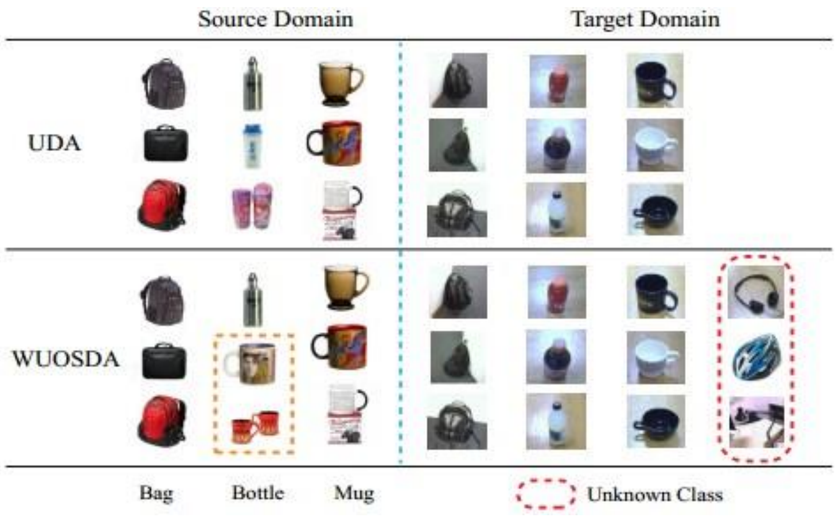
Open-set Instance-dependent Noisy Labels



Open-set Adversarial Examples



Open-set Domain Adaptation



Automated Trustworthy Deep Learning

Algorithm 2 *Search to Exploit* (S2E) algorithm for the minimization of the relaxed objective $\mathcal{J}$ in (6).

Initialize $\boldsymbol{\theta}^1 = \mathbf{1}$ so that $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is uniform distribution.
for $m = 1, \ldots, M$ **do**
  for $k = 1, \ldots, K$ **do**
    draw hyperparameter $\boldsymbol{x}$ from distribution $p_{\boldsymbol{\theta}^m}(\boldsymbol{x})$;
    using $\boldsymbol{x}$, run Algorithm 1 with $R(\cdot)$ in (4);
  end for
  use the K samples in steps 3-6 to approximate $\nabla\mathcal{J}(\boldsymbol{\theta}^m)$ in (7) and $\nabla^2\mathcal{J}(\boldsymbol{\theta}^m)$ in Proposition 1;
  update $\boldsymbol{\theta}^m$ by (8);
end for

SELECTED PUBLICATIONS

1. B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I.W. Tsang and M. Sugiyama. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *NeurIPS*, 2018.
2. B. Han, J. Yao, G. Niu, M. Zhou, I.W. Tsang, Y. Zhang and M. Sugiyama. Masking: A New Perspective of Noisy Supervision. In *NeurIPS*, 2018.
3. B. Han, G. Niu, X. Yu, Q. Yao, M. Xu, I.W. Tsang and M. Sugiyama. SIGUA: Forgetting May Make Learning with Noisy Labels More Robust. In *ICML*, 2020.
4. J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama and M. Kankanhalli. Attacks Which Do Not Kill Training Make Adversarial Learning Stronger. In *ICML*, 2020.
5. Q. Yao, H. Yang, B. Han, G. Niu and J.T. Kwok. Searching to Exploit Memorization Effect in Learning from Noisy Labels. In *ICML*, 2020.
6. J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama and M. Kankanhalli. Geometry-aware Instance-reweighted Adversarial Training. In *ICLR*, 2021.