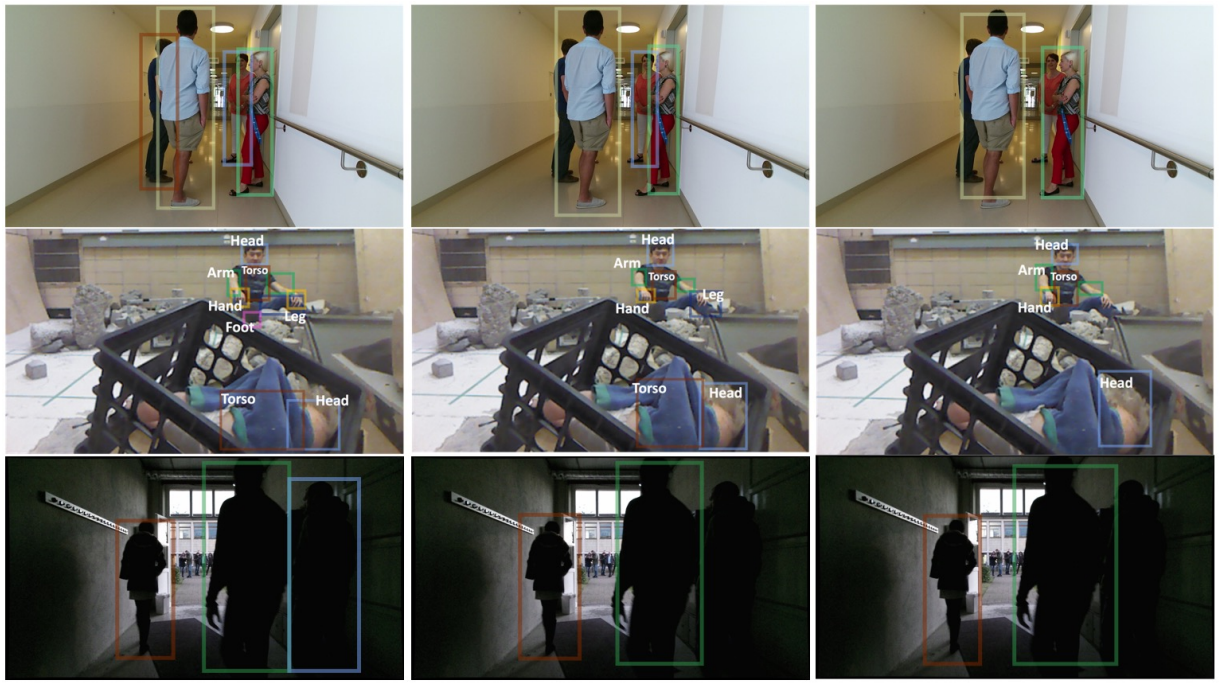
Robots Autonomously Detecting People: A Multimodal Deep Contrastive Learning Method Robust to Intraclass Variations

Angus Fung, Beno Benhabib, and Goldie Nejat
Autonomous Systems and Biomechatronics Lab (ASBLab), University of Toronto, Canada

1. Mobile robots face significant challenges in accurately detecting people in crowded (*with people*) and cluttered (*with objects*) human-centered environments under variable lighting conditions.
2. We developed a novel multimodal deep learning person detection architecture which uses a two-stage training approach consisting of Temporal Invariant Multimodal Contrastive Learning (TimCLR) and Multimodal YOLOv4 (MYOLOv4). TimCLR is a new pretraining method which incorporate intraclass variations from sampling video frames within a short temporal interval.
3. Comparison and ablation studies show our method outperforms existing person detection approaches in detecting people with *body occlusions* and *pose deformations* in different *lighting conditions*.



(a) TimCLR + MYOLOv4 (**ours**) (b) RGB-D CJ-MYOLOv4 (c) RGB-D C-FRCNN

Figure: Multimodal detection results from: (a) TimCLR + MYOLOv4 (ours), (b) RGB-D CJ-MYOLOv4, (c) RGB-D C-FRCNN; overlaid on RGB images.