22nd Conference on Robots and Vision

Taylor Institute for Teaching and Learning, University of Calgary
434 Collegiate Blvd NW, Calgary, AB T2N 1N4 Calgary, Alberta

May 27 - May 29, 2025

Detailed Program

Contents

1 Program at a Glance

Main Room: Taylor Institute Room 140/148

Legend	
🟧	CRV Session
🟩	Breaks & Social

	Workshop Day	Conference Day 1	Conference Day 2	Conference Day 3
Time	Monday 26-May	Tuesday 27-May	Wednesday 28-May	Thursday 29-May
9:00		Welcome Remarks		
9:15				
9:30		Oral Session 1 Chair: Mrigank Rochan	Keynote Speaker Julie A. Adams	Keynote Speaker Leonid Sigal
9:45				
10:00				
10:15				
10:30	Coffee	Coffee	Coffee	Coffee
10:45				
11:00	Hosted by Canadian AI	Oral Session 2 Chair: Angel Chang	Oral Session 4 Chair: Bernadette Bucher	Oral Session 5 Chair: Vahab Khoshdel
11:15				
11:30				
11:45				
12:00				
12:15				
12:30	Lunch Dinning Centre (UCalgary Dining Centre - The Landing)	Lunch Dinning Centre (UCalgary Dining Centre - The Landing)	Lunch Dinning Centre (UCalgary Dining Centre - The Landing)	Lunch Dinning Centre (UCalgary Dining Centre - The Landing) CIPPRS AGM (CRV Main Room)
12:45				
13:00				
13:15				
13:30				
13:45				
14:00	Hosted By Canadian AI	Oral Session 3 Chair: Melissa Greeff	Invited Talks: CIPPRS Dissertation Awards in Computer Vision and Robotics	Oral Session 6 Chair: Marie Charbonneau
14:15				
14:30				
14:45				
15:00	Coffee	Coffee	Coffee	Coffee
15:15				
15:30		Poster Lighting Talks + Poster Session 1	Poster Lighting Talks + Poster Session 2	Workshop: From AI Research to Real-World Impact: Navigating the Health Innovation Journey Speaker: Yuliya Fakhr, Alberta Innovates
15:45				
16:00				
16:15				
16:30				Closing Remarks
16:45				
17:00				
17:15				
17:30				
17:45				
18:00			Banquet and Awards Ceremony MacEwan Student Centre - Hall AB	
18:15				
18:30		Welcome Reception Atrium - Taylor Instiute for Teaching and Learning		
18:45				
19:00				
19:15				
19:30				
19:45				
20:00				
20:15				
20:30				
20:45				

2 Keynote Speakers

In alphabetical order

Julie A. Adams

Oregon State University

Talk Title: The Human-Robot Ratio (m:N) Theory: Limitations and Considerations

Abstract: The traditional human-to-robot ratio, or m:N theory states that the number of robots limits humans ability to manage and maintain overall team performance. This theory was developed primarily based on ground robot capabilities 10-15 years ago. While some traditional m:N limitations persist, both applied research and commercial systems debunk this traditional theory, particularly for very large numbers of robots (m¡¡N). This keynote will discuss the limitations of the theory, provide evidence that contradicts the theory, and discuss human factors aspects that will have an impact on the number of robots a single human can safely deploy. Results and examples will include simulated large autonomous uncrewed aircraft with associated necessary interactions with air traffic control, heterogeneous swarms deployed in urban environments, and commercial delivery uncrewed aircraft.

Biography: Dr. Adams is the founder of the Human-Machine Teaming Laboratory and the Associate Director of Research of the Collaborative Robotics and Intelligent Systems (CoRIS) Institute. Adams has focused on human-machine teaming and distributed artificial intelligence for thirty-five years. Throughout her career she has focused on unmanned systems, but also focused on crewed civilian and military aircraft at Honeywell, Inc. and commercial, consumer and industrial systems at the Eastman Kodak Company. Her research, which is grounded in robotics applications for domains such as first response, archaeology, oceanography, and the U.S. military, focuses on distributed artificial intelligence, swarms, robotics and human-machine teaming. Dr. Adams is an NSF CAREER award recipient, a Human Factors and Ergonomics Society Fellow as well as a member of the National Academies Board on Army Research and Development and the DARPA Information Science and Technology Study Group.

Leonid Sigal

University of British Columbia

Talk Title: The Curious Case of Foundational and VLM Models

Abstract: The capabilities and the use of foundational (FM) and vision-language (VLM) models in computer vision have exploded over the past few years. This has led to a broad paradigm shift in the field. In this talk I will focus on the recent work from my group that navigates this quickly evolving research landscape. Addressing challenges such as building foundational models with better generalization, increasing their context length, adopting them to ever evolving task landscape and routing information among them for more complex reasoning visual problems. I will also discuss some curious benefits and challenges of working with such models, including emergent (localization) capabilities and in-consistency in their responses.

Biography: Prof. Leonid Sigal is a Professor at the University of British Columbia (UBC). He was appointed CIFAR AI Chair at the Vector Institute in 2019 and an NSERC Tier 2 Canada Research Chair in Computer Vision and Machine Learning in 2018. Prior to this, he was a Senior Research Scientist, and a group lead, at Disney Research. He completed his Ph.D at Brown University in 2008; received his B.Sc. degrees in Computer Science and Mathematics from Boston University in 1999, his M.A. from Boston University in 1999, and his M.S. from Brown University in 2003. Leonid's research interests lie in the areas of computer vision, machine learning, and computer graphics; with the emphasis on approaches for visual and multi-modal representation learning, recognition, understanding and generative modeling. He has won a number of research awards, including Killam Accelerator Fellowship in 2021 and has published over 100 papers in venues such as CVPR, ICCV, ECCV, NeurIPS, ICLR, and Siggraph.

3 Symposium Speakers

In alphabetical order

Bernadette Bucher

University of Michigan

Talk Title: Building Visual Representations with Foundation Models for Mobile Manipulation

Abstract: Rapid improvements over the past few years in computer vision have enabled high performing geometric state estimation on moving camera systems in day-to-day environments. Furthermore, recent substantial improvements in language understanding and vision-language grounding have enabled rapid advancements in semantic scene understanding. In this presentation, I will demonstrate how we can build visual representations from these foundational vision-language models to enable new robotic capabilities in navigation, manipulation, and mobile manipulation. I will also discuss new robotics research directions opened up by these advancements in vision-language understanding.

Biography: Bernadette Bucher is an Assistant Professor in the Robotics Department at University of Michigan. She leads the Mapping and Motion Lab which focuses on learning interpretable visual representations and estimating their uncertainty for use in robotics, particularly mobile manipulation. Her work has been recognized by a Best Paper Award in Cognitive Robotics at ICRA 2024 and is funded by NASA and General Motors. Before joining University of Michigan this fall, she was a research scientist at the Boston Dynamics AI Institute, a senior software engineer at Lockheed Martin Corporation, and an intern at NVIDIA Research. She earned her PhD from University of Pennsylvania and bachelor's and Masters degrees from University of Alabama.

Angel Chang

Simon Fraser University — CIFAR AI Chair at AMII

Talk Title: Creating composable, interactive environments

Abstract: In popular imagination, household robots that we can instruct to "put my red mug back in the left kitchen cabinet" are common. Accomplishing this task may involve taking the mug, navigating to the kitchen, and opening the cabinet door to place the mug back in. To study whether an agent is capable of executing such a task in a simulated environment, it is important to be able to create a diverse set of interactive, composable environments. In this talk, I will describe recent trends in composable 3D indoor scene generation leveraging LLMs and VLMs, and efforts to model and create articulated objects for populating these environments.

Biography: Dr. Angel Chang is an Associate Professor at Simon Fraser University. She was previously a visiting research scientist at Facebook AI Research and a research scientist at Eloquent Labs, where she worked on dialogue systems. Dr. Chang earned her Ph.D. in Computer Science from Stanford University, where she was a member of the Natural Language Processing Group under the supervision of Professor Chris Manning. Her research lies at the intersection of language, 3D vision, and embodied AI, with a focus on connecting natural language to 3D representations of shapes and scenes. She is particularly interested in grounding language for embodied agents operating in indoor environments. Dr. Chang has developed methods for synthesizing 3D scenes and shapes from text and contributed to the creation of influential datasets for 3D scene understanding. Her broader interests include the semantics of shapes and scenes, common sense knowledge representation and acquisition, and reasoning using probabilistic models.

Marie Charbonneau

University of Calgary

Talk Title: Contact-based interaction for better human-robot collaborations

Abstract: Touch is a central component of humans' interactions with others and with the world. While robots are increasingly being developed to work alongside people, their capacity to interact with humans through touch is yet underdeveloped. This talk will explore why this may be the case, why it matters, and recent research at the Waterloo RoboHub and Calgary Human-Robot Collaboration lab towards making robots more physically interactive.

Biography: Dr. Marie Charbonneau works to make human-robot interactions safe, comfortable, and intuitive. Dr. Charbonneau joined the University of Calgary as Assistant Professor in September 2021, following postdoctoral work in humanoid robotics at the University of Waterloo and a PhD in Advanced and Humanoid Robotics from the Istituto Italiano di Tecnologia and the Università Degli Studi di Genova. Dr. Charbonneau's work in whole-body control regulates the forces between robots and their environment, towards ensuring respectful and reliable interactions with people. For instance, Dr. Charbonneau has programmed a humanoid robot to waltz with human partners, and currently works on improving a robot's awareness of and response to physical contacts.

Melissa Greeff

Queen's University

Talk Title: Robots Helping Robots: Enhancing Cross-Modal Interactions Between Aerial, Ground, and Surface Vessel Robots

Abstract: Single aerial robot systems can achieve high speed flight in challenging GPS-denied conditions enabling remote surveillance, package delivery and infrastructure inspection. However, we can further enhance robot operability in diverse environments (from air to land to marine) by augmenting the autonomous capabilities of single aerial, ground or surface vessels through cross-modal interactions. In this talk, we will discuss two different applications that benefit from cross-modality. Firstly, we explore how to leverage aerial robot imagery to enable GPS-denied, zero-shot autonomous navigation for ground vehicles in untraversed environments. Secondly, we explore how to co-ordinate autonomous aerial and surface vessels to enable the landing of aerial vehicles on surface vessels to recharge in remote marine or limnology applications. This is done by accommodating spatial and temporal uncertainties in the waves that can make landing challenging. These preliminary technologies have the potential to enable more persistent operation of robots in diverse environments.

Biography: Dr Melissa Greeff is an assistant professor in Electrical and Computer Engineering at Queen's University. She is an Ingenuity Labs Robotics and AI Institute Member and a Faculty Affiliate at the Vector Institute for Artificial Intelligence. She leads Robora Lab. Her research interests include aerial robots, vision-based navigation, and safe learning-based control. She has published in various international robotics and control systems venues including IEEE Robotics and Auto. Letters, Annual Review of Control, Robotics, and Autonomous Systems, ICRA, IROS and CDC. She has helped co-organize various workshops on safe robot learning and benchmarking at various international conferences. Her research is supported by NSERC, CFI, MITACs, the Department of National Defense (DND) and various industry collaborators. Dr. Greeff 's expertise is in building autonomous aerial systems including conducting field trials at various locations across Canada. She was listed as one of 50 women in robotics you need to know about in 2023 by the Women in Robotics organization.

Vahab Khoshdel

University of Manitoba

Talk Title: From Virtual Tissues to Digital Farms—How Diffusion Models Are Shaping Medicine and Robotics

Abstract: Diffusion models now let us generate anatomically accurate 3-D data for healthcare and agriculture. I'll present a latent-diffusion pipeline that (1) creates high-fidelity breast phantoms for safer, more accessible imaging and (2) synthesizes controllable point-cloud "digital farms" to train perception and manipulation in ag-robots. I'll share qualitative gains, deployment lessons, and open challenges—uncertainty quantification, domain adaptation, real-time operation—and show how generative data can drive robotic decisions and field outcomes.

Biography: Vahab Khoshdel, PhD, P.Eng. is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Manitoba. He holds dual PhDs—one in Robotics from Ferdowsi University of Mashhad (Iran) and another in Biomedical Engineering from the University of Manitoba. His research interests lie at the intersection of machine learning, computer vision, and robotics, with applications in medical imaging, agriculture, and power systems. In addition to his academic work, Dr. Khoshdel has extensive experience in the computer industry, where he has led interdisciplinary research and development teams. He has overseen the design and deployment of AI-driven solutions in both startup and consulting environments, translating cutting-edge research into production-level systems.

Mrigank Rochan

University of Saskatchewan

Talk Title: Advancing Video Abstraction with Deep Learning

Abstract: As video data continues to grow exponentially in volume and complexity, the development of intelligent systems to manage and summarize videos has become a pressing need. Video abstraction, a key task in computer vision and video understanding, aims to create a short, informative visual summary of a video, enabling users to quickly gain valuable insights about the video without watching it entirely. With applications spanning entertainment, sports, surveillance, healthcare, and video search, this technology has the potential to transform how we interact with video content and unlock the full potential of video data. In this talk, I will discuss our recent research and innovative solutions leveraging deep learning to advance the state of the art in video abstraction.

Biography: Dr. Mrigank Rochan is an Assistant Professor in the Department of Computer Science at the University of Saskatchewan, where he leads a research group focusing on computer vision and deep learning. Prior to this, he was a Senior Researcher with the Autonomous Driving Perception team at Huawei Noah's Ark Lab in Toronto. He earned his PhD from the University of Manitoba, and his doctoral thesis was awarded the 2020 Canadian Image Processing and Pattern Recognition Society (CIPPRS) John Barron Doctoral Dissertation Award, a national award presented annually to the top PhD thesis in computer or robot vision in Canada. His research has been published in top-tier computer vision and robotics venues, including CVPR, ICCV, ECCV, ICRA, and TPAMI. Dr. Rochan's research is currently supported by the University of Saskatchewan, Google, and NSERC.

4 Oral Sessions

Oral Session 1 — Day 1

TrapCounter: Integrating Detection, Tracking and Optical Flow for the Automatic Counting of Active Fishing Traps

Ali Soltaninezhad (University of Victoria), Melissa Cote (University of Victoria), Alejandro Rico Espinosa (University of Victoria), Tunai Porto Marques (University of Victoria), Alexandra Branzan Albu (University of Victoria), Justin Paetkau (University of Victoria), Vanesa Diaz Gimeno (Archipelago Marine Research), and Jacob W. Lower (Archipelago Marine Research)

Sablefish are a commercially important species in the northeast Pacific Ocean, typically caught using longline traps in a tightly regulated fishery. Mandatory electronic monitoring (EM) systems record deck activities to support compliance, but reviewing footage for trap hauling remains a manual, time-intensive process. This paper introduces TrapCounter, a method that automatically counts active fishing traps in EM videos by combining deep learning-based detection, tracking, and optical flow. We integrate the RetinaNet object detection architecture with the StrongSORT tracker to properly detect and track active traps as they are being hauled, addressing such challenges as varied poses and appearance due to the semi-transparent nature of traps and crowded deck environments. To reduce overcounting, we add an optical flow-based counting module that improves the tracker's active trap count accuracy. We also release DTCFT, the first public dataset for trap detection, tracking, and counting, built from actual commercial and survey fishing vessel footage captured under varying conditions. Experiments show that TrapCounter outperforms approaches based on YOLOv9, Faster R-CNN, DeepSORT, and ByteTrack, streamlining EM workflows and enhancing EM scalability in commercial and survey fisheries. Dataset: `https://github.com/alisoltani7596/TrapCounter`

Attention-Mamba for Multi-Object Tracking

Dheeraj Khanna (University of Waterloo) and John S. Zelek (University of Waterloo)

Multi-Object Tracking (MOT) is fundamental to applications such as autonomous driving, video surveillance, and sports analytics. However, challenges persist in maintaining long-term identity associations, handling dynamic object counts, managing irregular motion patterns, and mitigating occlusions in complex environments. Inspired by advancements in state-space models (SSMs), particularly Mamba [1], we propose a novel learning-based motion prediction architecture that integrates Mamba's input-dependent sequence modeling with self-attention layers to effectively capture non-linear motion patterns within the Tracking-By-Detection (TBD) framework. Mamba's ability to model long-range dependencies enhances motion prediction. We further refine spatial association by improving the cost matrix traditionally based on Intersection over Union (IoU). We incorporate Height-based IoU and extend bounding boxes using adjusted buffers to account for fast motion and partial overlaps, increasing robustness in object association. Achieving accurate MOT requires a balance between precise motion modeling and effective spatial and appearance matching, leveraging both strong and weak cues in data association. Our method is evaluated on challenging benchmarks, such as DanceTrack [2] and SportsMOT [3], achieving HOTA scores of 63.16% and 77.26%, respectively. These results surpass multiple state-of-the-art methods with a 3-7% improvement over other learning-based motion models, demonstrating the effectiveness of our approach in real-world tracking scenarios.

Oral Session 2 — Day 1

SemanticOBB: Semantic Front Estimation for Indoor 3D Objects

Armin Kavian (Simon Fraser University) and Manolis Savva (Simon Fraser University)

We introduce the SemanticOBB task: identifying the semantically meaningful front side of 3D objects in indoor scenes. Given a 3D scene point cloud as input, we output the front direction for each object as a 3D vector. Knowing the front side of objects has many applications in robotics, augmented reality, and 3D shape generation. This is a challenging task due to the diversity of object categories, the large intra-category variance in objects, and partial observations in 3D reconstructions. We design and systematically benchmark three families of approaches for this task based on classification, regression, and an anchor-based hybrid of classification and regression. We also study set-based reasoning to aggregate features for multiple object instances of the same category, and the impact of 3D-only vs combined 2D and 3D object features. Our experiments on two 3D reconstruction datasets show that there is much space for improvement on this task, with the best performing methods having mAP values in the 30% range.

CHOSEN: Contrastive Hypothesis Selection for Multiview Depth Refinement

Di Qiu (Google XR), Yinda Zhang (Google XR), Thabo Beeler (Google XR), Vladimir Tankovich (Google XR), Christian Häne (Google XR), Sean Fanello (Google XR), Christoph Rhemann (Google XR), and Sergio Orts-Escolano (Google XR)

We propose CHOSEN, a simple yet flexible, robust and effective multi-view depth refinement framework. It can provide significant improvement in depth and normal quality, and can be integrated in existing multi-view stereo pipelines with minimal modifications. Given an initial depth estimation, CHOSEN iteratively re-samples and selects the best hypotheses. The key to our approach is the application of contrastive learning in an appropriate solution space and a carefully designed hypothesis feature, based on which positive and negative hypotheses can be effectively distinguished. We integrated CHOSEN in a basic multi-view stereo pipeline, and show that it can deliver impressive quality in terms of depth and normal accuracy compared to many other top deep learning based multi-view stereo pipelines.

Weeds and Crops: The Stem Emergence Point Dataset Collection

Dieudonné N. D'Arnall (Acadia University), Andrew R. McIntyre (Acadia University), and Lydia Bouzar-Benlabiod (Acadia University)

Weeds present a significant threat to agriculture by competing with crops for vital resources like water, nutrients, and sunlight, ultimately reducing crop yields. Detecting weeds at an early stage is crucial for implementing effective control measures such as herbicide application, mechanical removal, or other treatments, which become less efficient and more costly as crops grow. In this paper, a dataset of 2,382 images with 1,389 images from a carrot field and 993 images from a lettuce field is presented. The visual challenge lies in accurately distinguishing weeds from crops. The paper also presents three deep neural network models for the weed detection task: a standard CNN, a hybrid CNN-LSTM, and a pure LSTM model—to assess the impact of temporal information on classification performance. Model performances are analyzed under identical training and testing conditions, evaluating trade-offs in accuracy, false detection rates, and inference speed. Our findings suggest that while the pure CNN model achieves the lowest false weed detection rate, it also has the highest false crop identification rate (13%), whereas the pure LSTM model converges faster and provides the best overall weed classification performance.

Oral Session 3 — Day 1

An SE(3) Noise Model for Range-Azimuth-Elevation Sensors

Thomas Hitchcox (McGill University) and James Richard Forbes (McGill University)

Scan matching is a widely used technique in state estimation. Point-cloud alignment, one of the most popular methods for scan matching, is a weighted least-squares problem in which the weights are determined from the inverse covariance of the measured points. An inaccurate representation of the covariance will affect the weighting of the least-squares problem. For example, if ellipsoidal covariance bounds are used to approximate the curved, "banana-shaped" noise characteristics of many scanning sensors, the weighting in the least-squares problem may be overconfident. Additionally, sensor-to-vehicle extrinsic uncertainty and odometry uncertainty during submap formation are two sources of uncertainty that are often overlooked in scan matching applications, also likely contributing to overconfidence on the scan matching estimate. This paper attempts to address these issues by developing a model for range-azimuth-elevation sensors on matrix Lie groups. The model allows for the seamless incorporation of extrinsic and odometry uncertainty. Illustrative results are shown both for a simulated example and for a real point-cloud submap collected with an underwater laser scanner.

Pointing the Way: Refining Radar-Lidar Localization Using Learned ICP Weights

Daniil Lisus (University of Toronto), Johann Laconte (University of Toronto), Keenan Burnett (University of Toronto), Ziyu Zhang (University of Toronto), Timothy Barfoot (University of Toronto)

This paper presents a novel deep-learning-based approach to improve localizing radar measurements against lidar maps. This radar-lidar localization leverages the benefits of both sensors; radar is resilient against adverse weather, while lidar produces high-quality maps in clear conditions. However, owing in part to the unique artefacts present in radar measurements, radar-lidar localization has struggled to achieve comparable performance to lidar-lidar systems, preventing it from being viable for autonomous driving. This work builds on ICP-based radar-lidar localization by including a learned preprocessing step that weights radar points based on high-level scan information. To train the weight-generating network, we present a novel, stand-alone, open-source differentiable ICP library. The learned weights facilitate ICP by filtering out harmful radar points related to artefacts, noise and even vehicles on the road. Combining an analytical approach with a learned weight reduces overall localization errors and improves convergence in radar-lidar ICP results run on real-world autonomous driving data. Our code base is publicly available to facilitate reproducibility and extensions. Paper code is available at: `https://github.com/utiasASRL/mm_masking`

Oral Session 4 — Day 2

The Finer Points: A Systematic Comparison of Point-Cloud Extractors for Radar Odometry

Elliot Preston-Krebs (University of Toronto), Daniil Lisus (University of Toronto), and Timothy Barfoot (University of Toronto)

A key element of many odometry pipelines using spinning frequency-modulated continuous-wave radar is the extraction of a point-cloud from the raw signal intensity returns. This extraction greatly impacts the overall performance of point-cloud-based odometry, but a consensus on which extractor performs best in which circumstances is missing. This paper provides a first-of-its-kind, comprehensive comparison of 13 common radar point-cloud extractors for the task of iterative closest point-based odometry in autonomous driving environments. Each extractor's parameters are tuned and tested on two FMCW radar datasets using approximately 176 km of data from public roads. We find that the simplest, and fastest extractor, K-strongest, performs the best overall, outperforming the average by 13.59% and 24.94% on each dataset, respectively. In addition to an overall extractor recommendation, we highlight trends and note the substantial impact that the choice of extractor can have on the accuracy of odometry. ↗

TRIT-Net: Triplet-based Railway Instance Tracing Network Using Attraction Field Representation

Mohammadjavad Ghorbanalivakili (York University) and Gunho Sohn (York University)

Driverless trains with Grade of Automation 4 (GoA4) are reliably implemented in metro systems but lack local situational awareness, requiring human intervention during safety-critical events. In contrast, autonomous trains must perceive their environment and make real-time decisions similar to a human operator responses. This study focuses on rail route identification from RGB images, a key aspect of visual intelligence for autonomous railway. While existing methods have successfully achieved rail detection and segmentation, they struggle to generalize rail associations for route formation due to relying on hard-coded heuristics tailored to specific datasets. To address this, we propose TRIT-Net that simultaneously predicts left-right rail associations and route instances. Our architecture integrates convolutional layers with transformers, featuring a rail triplet regression branch and an Attraction Field Map (AFM) for instance-based route tracing, inspired by trajectory tracking theory in closed-loop linear control systems. Our experiments, ablation study, and stability analysis confirm the effectiveness of our approach in achieving comparatively higher precision, recall, and inference speed. We also demonstrate that our versatile model selection strategy is applicable to any off-the-shelf semantic segmentation baseline. Future work will focus on recognizing track switch configuration to extract a unique train route rather than multiple ego-path candidates. ↗

Oral Session 5 — Day 3

Visual Concept-driven Image Generation with Text-to-Image Diffusion Model

Tanzila Rahman (University of British Columbia, Vector Institute for AI), Shweta Mahajan (University of British Columbia, Vector Institute for AI), Hsin-Ying Lee (Snap Inc.), Jian Ren (Snap Inc.), Sergey Tulyakov (Snap Inc.), and Leonid Sigal (University of British Columbia, Vector Institute for AI, Canada CIFAR AI Chair)

Text-to-image (TTI) diffusion models have demonstrated impressive results in generating high-resolution images of complex and imaginative scenes. Recent approaches have further extended these methods with personalization techniques that allow them to integrate user-illustrated concepts (e.g., the user him/herself) using a few sample image illustrations. However, the ability to generate images with multiple interacting concepts, such as human subjects, as well as concepts that may be entangled in one, or across multiple, image illustrations remains illusive. In this work, we propose a concept-driven TTI personalization framework that addresses these core challenges. We build on existing works that learn custom tokens for user-illustrated concepts, allowing those to interact with existing text tokens in the TTI model. However, importantly, to disentangle and better learn the concepts in question, we jointly learn (latent) segmentation masks that disentangle these concepts in user-provided image illustrations. We do so by introducing an Expectation Maximization (EM)-like optimization procedure where we alternate between learning the custom tokens and estimating (latent) masks encompassing corresponding concepts in user-supplied images. We obtain these masks based on cross-attention, from within the U-Net parameterized latent diffusion model and subsequent DenseCRF optimization. We illustrate that such joint alternating refinement leads to the learning of better tokens for concepts and, as a by-product, latent masks. We illustrate the benefits of the proposed approach qualitatively and quantitatively with several examples and use cases that can combine three or more entangled concepts. ↗

SinSim: Sinkhorn-Regularized SimCLR

M. Hadi Sepanj (University of Waterloo) and Paul W. Fieguth (University of Waterloo)

Self-supervised learning has revolutionized representation learning by eliminating the need for labeled data. Contrastive learning methods, such as SimCLR, maximize the agreement between augmented views of an image but lack explicit regularization to enforce a globally structured latent space. This limitation often leads to suboptimal generalization. We propose SinSim, a novel extension of SimCLR that integrates Sinkhorn regularization from optimal transport theory to enhance representation structure. The Sinkhorn loss, an entropy-regularized Wasserstein distance, encourages a well-dispersed and geometry-aware feature space, preserving discriminative power. Empirical evaluations on various datasets demonstrate that SinSim outperforms SimCLR and achieves competitive performance against prominent self-supervised methods such as VICReg and Barlow Twins. UMAP visualizations further reveal improved class separability and structured feature distributions. These results indicate that integrating optimal transport regularization into contrastive learning provides a principled and effective mechanism for learning robust, well-structured representations. Our findings open new directions for applying transport-based constraints in self-supervised learning frameworks. ↗

TrackPGD: Efficient Adversarial Attack using Object Binary Masks against Robust Transformer Trackers

Fatemeh Nourilenjan Nokabadi (Université Laval, IID - Institute Intelligence and Data, Mila – Quebec AI Institute), Yann Pequignot (Université Laval, IID - Institute Intelligence and Data), Jean-Francois Lalonde (Université Laval, IID - Institute Intelligence and Data), and Christian Gagné (Université Laval, IID - Institute Intelligence and Data, Mila – Quebec AI Institute, Canada-CIFAR AI Chair)

Adversarial perturbations can deceive neural networks by adding small, imperceptible noise to the input. Recent object trackers with transformer backbones have shown strong performance on tracking datasets, but their adversarial robustness has not been thoroughly evaluated. While transformer trackers are resilient to black-box attacks, existing white-box adversarial attacks are not universally applicable against these new transformer trackers due to differences in backbone architecture. In this work, we introduce TrackPGD, a novel white-box attack that utilizes predicted object binary masks to target robust transformer trackers. Built upon the powerful segmentation attack SegPGD, our proposed TrackPGD effectively influences the decisions of transformer-based trackers. Our method addresses two primary challenges in adapting a segmentation attack for trackers: limited class numbers and extreme pixel class imbalance. TrackPGD uses the same number of iterations as other attack methods for tracker networks and produces competitive adversarial examples that mislead transformer and non-transformer trackers such as MixFormerM, OSTrackSTS, TransT-SEG, and RTS on datasets including VOT2022STS, DAVIS2016, UAV123, VOT2018 and GOT-10k. Additional information and resources are available on the project webpage at `https://lvsn.github.io/TrackPGD/`.

Oral Session 6 — Day 3

Vision-based Autonomous Blood Suction with a Concentric Tube Continuum Robot

Jinjie Sun (University of Toronto), Courtney Amm (University of Toronto), Jessica Burgner-Kahrs (University of Toronto, University of Toronto Mississauga), and Lueder Alexander Kahrs (University of Toronto, University of Toronto Mississauga)

Blood-water mixture removal is essential in many surgeries. Existing works adding a robotic assistant primarily focus on conventional robots. Limited attention has been given to using continuum robots for this task. This paper introduces a vision-based control framework for autonomous liquid suction using a concentric tube continuum robot (CTCR). The proposed method employs a controller using camera input combined with a hybrid control strategy that integrates differential inverse kinematics and a pre-computed lookup table to ensure stable and precise motion during suction. A CTCR simulator, implemented in the Unity Game Engine with photorealistic rendering and robot-liquid interaction capabilities, as well as a benchtop robot system were developed as the experiment platform. The proposed method was evaluated through simulation and real-world experiments across four scenarios, demonstrating its generalizability and stability. In 32 real-world trials, less than 0.1 g of liquid remained after the suction, while over 99% of liquid was removed across 32 simulated trials. The results highlight the potential of CTCR for autonomous surgical liquid suction, showcasing the system's adaptability and performance in dynamic environments.

MakeWay: Object-Aware Costmaps for Proactive Indoor Navigation Using LiDAR

Binbin Xu (University of Toronto), Allen Tao (University of Toronto), Hugues Thomas (Apple), Jian Zhang (Apple), and Timothy Barfoot (University of Toronto)

In this paper, we introduce a LiDAR-based robot navigation system, based on novel object-aware affordance-based costmaps. Utilizing a 3D object detection network, our system identifies objects of interest in LiDAR keyframes, refines their 3D poses with the Iterative Closest Point (ICP) algorithm, and tracks them via Kalman filters and the Hungarian algorithm for data association. It then updates existing object poses with new associated detections and creates new object maps for unmatched detections. Using the maintained object-level mapping system, our system creates affordance-driven object costmaps for proactive collision avoidance in path planning. Additionally, we address the scarcity of indoor semantic LiDAR data by introducing an automated labeling technique. This method utilizes a CAD model database for accurate ground-truth annotations, encompassing bounding boxes, positions, orientations, and point-wise semantics of each object in LiDAR sequences. Our extensive evaluations, conducted in both simulated and real-world robot platforms, highlights the effectiveness of proactive object avoidance by using object affordance costmaps, enhancing robotic navigation safety and efficiency. The system can operate in real-time onboard and we intend to release our code and data for public use.

5 Poster Session

Robotics

LeYOLO, New Embedded Architecture for Object Detection

Lilian Hollard (Université de Reims Champagne-Ardenne), Lucas Mohimont (Université de Reims Champagne-Ardenne), Luiz Angelo Steffenel (Université de Reims Champagne-Ardenne), and Nathalie Gaveau (Université de Reims Champagne-Ardenne)

Efficient computation in deep neural networks is crucial for real-time object detection. However, recent advancements primarily result from improved high-performing hardware rather than improving parameters and FLOP efficiency. This is especially evident in the latest YOLO architectures, where speed is prioritized over lightweight design. As a result, object detection models optimized for lows-resource environments like microcontrollers have received less attention. For devices with limited computing power, existing solutions primarily rely on SSDLite or combinations of low-parameter classifiers, creating a noticeable gap between YOLO-like architectures and truly efficient lightweight detectors. This raises a key question: Can a model optimized for parameter and FLOP efficiency achieve accuracy levels comparable to mainstream YOLO models? To address this, we introduce two key contributions to object detection models using MSCOCO as a base validation set. First, we propose LeNeck, a general-purpose detection framework that maintains inference speed comparable to SSDLite while significantly improving accuracy and reducing parameter count. Second, we present LeYOLO, an efficient object detection model designed to enhance computational efficiency in YOLO-based architectures. LeYOLO effectively bridges the gap between SSDLite-based detectors and YOLO models, offering high accuracy in a model as compact as MobileNets. Both contributions are particularly well-suited for mobile, embedded, and ultra-low-power devices, including microcontrollers, where computational efficiency is critical. Code: `https://github.com/LilianHollard/LeYOLO`.

Adaptformer: Sequence Models as Adaptive Iterative Planners

Akash Karthikeyan (University of Waterloo) and Yash Vardhan Pant (University of Waterloo)

Despite recent advances in sequence models for autonomous systems, adapting to harder, unseen tasks at test time while leveraging only demonstrations from simpler tasks remains a significant challenge. This limitation is particularly critical in planning and decision-making, where agents must utilize previously observed data to generate informed actions for novel scenarios, rather than resorting to random behavior. Conventional behavioral cloning techniques often fail in these contexts, as they rely heavily on well-represented demonstrations (labeled data) and struggle with coherent generation of long-horizon plans. To address these challenges, we propose Adaptformer, a stochastic and adaptive planner that leverages energy-based sequence models to enable sample-efficient exploration and exploitation. Adaptformer learns an energy landscape to optimize trajectories and adapts effectively to novel test cases. Additionally, it employs an intrinsic goal proposal module to generate achievable shorter sub-goals, enabling consistent and long-horizon action sequence generation. This modular design also facilitates reasoning about the planner's behavior. Empirical evaluations in procedurally generated maze environments demonstrate Adaptformer's effectiveness, achieving up to a 25% improvement over state-of-the-art methods in long-horizon adaptation tasks where existing models fail.

Path-Following Controller Designs for Autonomous and Semi-Autonomous Industrial Motor Graders

Anthony Beca (Queen's University) and Joshua A. Marshall (Queen's University)

Haulage road maintenance is crucial for operational efficiency and safety in mining and construction activities. Industrial motor graders play a key role in this task, both on surface and in underground mines, where production vehicles—such as trucks and loaders—are increasingly being driven autonomously. However, motor graders have yet to be commercially automated. The redundant kinematics of motor graders, including articulation, front-axle steering, and blade operations, pose technical challenges for autonomy. In this work, we leverage the steering redundancy of motor grader designs to formulate a new path following controller that is compatible with existing approaches for the automation of articulated vehicles. The proposed methodology, coined "Single-Track Control" (STC) allows for coordination of both the front-axle steering angle and the vehicle's articulation angle to keep the front and rear wheels on a common track. This innovation mitigates the risk of collisions with drift walls and improves manoeuvrability. It can be used for semi-autonomous operations, to reduce the complexity for operators, as well as for fully autonomous operations. The approach was validated in simulation, comparing the implementation performance of two controller types.

A Target-Based Extrinsic Calibration Framework for Non-Overlapping Camera-Lidar Systems Using a Motion Capture System

Nicholas Charron (University of Waterloo), Huaiyuan Weng (University of Waterloo), Steven L. Waslander (University of Toronto), and Sriram Narasimhan (University of California Los Angeles)

We present a novel target-based lidar-camera extrinsic calibration methodology that can be used for non-overlapping field of view (FOV) sensors. Contrary to previous work, our methodology overcomes the non-overlapping FOV challenge using a motion capture system (MCS) instead of traditional simultaneous localization and mapping approaches. Due to the high relative precision of MCSs, our methodology can achieve both the high accuracy and repeatable calibrations common to traditional target-based methods, regardless of the amount of overlap in the sensors' field of view. Furthermore, we design a target-agnostic implementation that does not require uniquely identifiable features by using an iterative closest point approach, enabled by the MSC measurements. We show using simulation that we can accurately recover extrinsic calibrations for a range of perturbations to the true calibration that would be expected in real circumstances. We prove experimentally that our method out-performs state-of-the-art lidar-camera extrinsic calibration methods that can be used for non-overlapping FOV systems, while using a target-based approach that guarantees repeatably high accuracy. Lastly, we show in simulation that different target designs can be used, including easily constructed 3D targets such as a cylinder that are normally considered degenerate in most calibration formulations.

aUToPath: Unified Planning and Control for Autonomous Vehicles in Urban Environments Using Hybrid Lattice and Free-Space Search

Tanmay P. Patel (University of Toronto), Connor Wilson* (University of Toronto), Ellina R. Zhang* (University of Toronto), Morgan Tran (University of Toronto), Chang Keun Paik (University of Toronto), Steven L. Waslander (University of Toronto), and Timothy D. Barfoot (University of Toronto)*

**equal contribution*

This paper presents aUToPath, a unified online framework for global path-planning and control to address the challenge of autonomous navigation in cluttered urban environments. A key component of our framework is a novel hybrid planner that combines pre-computed lattice maps with dynamic free-space sampling to efficiently generate optimal driveable corridors in cluttered scenarios. Our system also features sequential convex programming (SCP)-based model predictive control (MPC) to refine the corridors into smooth, dynamically consistent trajectories. A single optimization problem is used to both generate a trajectory and its corresponding control commands; this addresses limitations of decoupled approaches by guaranteeing a safe and feasible path. Simulation results of the novel planner on randomly generated obstacle-rich scenarios demonstrate the success rate of a free-space Adaptively Informed Trees* (AIT*)-based planner, and runtimes comparable to a lattice-based planner. Real-world experiments of the full system on a Chevrolet Bolt EUV further validate performance in dense obstacle fields, demonstrating no violations of traffic, kinematic, or vehicle constraints, and a 100% success rate across eight trials.

Image Space Path Following Control Using Visual Servoing

Cole Dewis (University of Alberta) and Martin Jagersand (University of Alberta)

Visual servoing has been well explored in the literature for task specification and planning in image space. Planning tasks and paths in image space can be especially useful in unstructured environments, as a 3D reconstruction is not needed. However, few works have discussed following arbitrary image paths with visual servoing for robotic arms. This paper presents a path following controller for robotic arms based on image based visual servoing that can follow arbitrary paths in image space. The controller uses visual error to generate velocities that smoothly approach the path along the tangent. Additionally, the controller can optionally follow the orientation of the path, and can be applied to both eye-in-hand and eye-to-hand setups. Experiments are conducted on a Kinova Gen3 7DOF arm to evaluate the controller. Benefits of the path following controller over a trajectory-tracking approach are shown. Specifically, our path following controller displays smooth responses to physical disturbances and forced pauses. ↗

Vision

WSCurLe: Weakly Supervised Curriculum Learning for Foundational Vision and Language Architectures in Digital Soil Mapping

Vishvam Porwal (University of Guelph), Stacey D. Scott (University of Guelph), Neil D. B. Bruce (University of Guelph), and Asim Biswas (University of Guelph)

Effective soil sampling is crucial for modern agricultural practices, requiring precise identification of viable agricultural land while excluding non-agricultural areas. While existing remote sensing approaches can distinguish broad land-use categories, they often lack the precision needed for agricultural applications. Our previous work introduced SLVVA, a satellite-based framework for land viability analysis, but its performance was limited by the representation capabilities of its vision-language encoders. In this paper, we propose WSCurLe, a novel weakly supervised curriculum learning approach that progressively fine-tunes vision-language encoders to enhance their semantic understanding while preserving fine-grained image details. Through a series of alternating training stages combining contrastive learning and reconstruction-based methods, WSCurLe significantly improves the quality of land viability segmentation. We demonstrate the effectiveness of our approach through comprehensive comparisons with state-of-the-art vision and language encoders, showing substantial improvements in both efficiency and accuracy for agricultural land analysis. ↗

Unmasking Facial DeepFakes: A Robust Multiview Detection Framework for Natural Images

Sami Belguesmia (University of Quebec in Outaouais), Mohand Said Allili (University of Quebec in Outaouais), and Assia Hamadene (University of Quebec in Outaoua)

DeepFake technology has advanced significantly in recent years, enabling the creation of highly realistic synthetic face images. Existing DeepFake detection methods often struggle with pose variations, occlusions, and artifacts that are difficult to detect in real-world conditions. To address these challenges, we propose a multi-view architecture that enhances DeepFake detection by analyzing facial features at multiple levels. Our approach integrates three specialized encoders—a global view encoder for detecting boundary inconsistencies, a middle view encoder for analyzing texture and color alignment, and a local view encoder for capturing distortions in expressive facial regions such as the eyes, nose, and mouth, where DeepFake artifacts frequently occur. Additionally, we incorporate a face orientation encoder, trained to classify face poses, ensuring robust detection across various viewing angles. By fusing features from these encoders, our model achieves superior performance in detecting manipulated images, even under challenging pose and lighting conditions. Experimental results on challenging datasets demonstrate the effectiveness of our method, outperforming conventional single0view approaches.

WheatSAM: A Two-Stage Wheat Head Automatic Segmentation Framework

Md Jaber Al Nahian (University of Calgary), Tapotosh Ghosh (University of Calgary), Farnaz Sheikhi (University of Calgary), Ian McQuillan (University of Saskatchewan), and Farhad Maleki (University of Calgary)

The Segment Anything Model (SAM) has demonstrated impressive zero-shot segmentation capability. However, it requires accurate prompts. In this paper, we benchmark SAM on wheat head segmentation for the first time, evaluating its performance using manually and automatically generated prompts. Our findings indicate that SAM provides good performance with high-quality manually annotated bounding boxes, and fine-tuning further enhances the segmentation performance. However, the performance of SAM significantly decreases when the bounding boxes are generated by YOLOV9 or Faster RCNN due to object detection errors. To mitigate this, we have proposed WheatSAM, a two-stage segmentation model in which SAM is fine-tuned with an error-aware module. The error-aware module synthetically adds controlled perturbations in bounding boxes during training to simulate real-world bounding box errors. By methodically regulating the error probability, shift, scaling, and overlap, we have achieved the Dice score of 89.73% with YOLO-generated bounding boxes, surpassing SAM's zero-shot performance on YOLO-generated prompts, which was 89.44%. Our results demonstrate a basic trade-off between manual annotation and segmentation precision, illustrating the potential of error-aware training to bridge the gap between manually annotated prompts and automatic detection-based segmentation.

Supervised Contrastive Learning for Ordinal Engagement Measurement

Sadaf Safa (KITE Research Institute, University of Toronto), Ali Abedi (KITE Research Institute, University of Toronto), and Shehroz S. Khan (KITE Research Institute, University of Toronto, American University of the Middle East)

Student engagement plays a crucial role in the successful delivery of educational programs. Automated engagement measurement helps instructors monitor student participation, identify disengagement, and adapt their teaching strategies to enhance learning outcomes effectively. This paper identifies two key challenges in this problem: class imbalance and incorporating order into engagement levels rather than treating it as mere categories. Then, a novel approach to video-based student engagement measurement in virtual learning environments is proposed that utilizes supervised contrastive learning for ordinal classification of engagement. Various affective and behavioral features are extracted from video samples and utilized to train ordinal classifiers within a supervised contrastive learning framework (with a sequential classifier as the encoder). A key step involves the application of diverse time-series data augmentation techniques to these feature vectors, enhancing model training. The effectiveness of the proposed method was evaluated using a publicly available dataset for engagement measurement, DAiSEE, containing videos of students who participated in virtual learning programs. The results demonstrate the robust ability of the proposed method for the classification of the engagement level. This approach promises a significant contribution to understanding and enhancing student engagement in virtual learning environments.

POC-SLT: Partial Object Completion with SDF Latent Transformers

Faezeh Zakeri (University of Tübingen), Raphael Braun (University of Tübingen), Lukas Ruppert (University of Tübingen), and Hendrik P.A. Lensch (University of Tübingen)

3D geometric shape completion hinges on representation learning and a deep understanding of geometric data. Without profound insights into the three-dimensional nature of the data, this task remains unattainable. Our work addresses this challenge of 3D shape completion given partial observations by proposing a transformer operating on a latent space representing Signed Distance Fields (SDFs). Instead of a monolithic volume, the SDF of an object is partitioned into smaller high-resolution patches leading to a sequence of latent codes. The approach relies on a smooth latent space encoding learned via a variational autoencoder (VAE), trained on millions of 3D patches. We employ an efficient masked autoencoder transformer to complete partial sequences into comprehensive shapes in latent space. Our approach is extensively evaluated on partial observations from ShapeNet and the ABC dataset where only fractions of the objects are given. The proposed POC-SLT architecture compares favorably with several baseline state-of-the-art methods, demonstrating a significant improvement in 3D shape completion, both qualitatively and quantitatively.

6 Workshop

From AI Research to Real-World Impact: Navigating the Health Innovation Journey

Yuliya Fakhr, Ph.D. (Senior Business Partner, Health Platforms, Alberta Innovates)

Turning AI research into real-world healthcare solutions takes more than algorithms—it takes strategy, validation, and system-level thinking. This talk walks through the health innovation journey, outlining what it really takes to move from discovery to adoption in clinical and wellness settings. Using examples from AI in healthcare, we explore the unique challenges of implementation: regulatory compliance, clinical validation, health system procurement, and the critical importance of designing for real-world workflows. Attendees will gain a practical framework for translating their AI research into solutions that clinicians, patients, and health systems can trust and use. The session will also include an interactive audience discussion focused on researchers' experiences, challenges, and support needs when translating health-related AI into clinical or public health solutions. Attendees will gain a practical framework for advancing their work from research to impact and insights into securing grant funding to support the journey.