

23rd Conference on Robots and Vision

Student Union Building Ballroom 5000, Simon Fraser University
8888 University Dr W, Burnaby
Vancouver, British Columbia V5A 1S6

May 26 – May 28, 2026

Detailed Program

Contents

1	Program at a Glance	2
2	Keynote Speakers	3
2.1	Mac Schwager	3
2.2	Ian Stavness	3
3	Symposium Speakers	4
3.1	Nils Wilde	4
3.2	Nandita Vijaykumar	5
3.3	Xinxin Zuo	5
3.4	Risto Ojala	6
3.5	Yani Ioannou	6
3.6	Maciek Popik presenting on behalf of Mahdis Bisheban	7
4	Oral Sessions	7
5	Poster Sessions & Lightning Talks	16
6	AI-CRV Nectar Track	24

1 Program at a Glance

Main CRV Room: Student Union Building Ballroom 5000

Legend						
		CRV Session				
		Breaks & Social				
	Workshop Day	Conference Day 1	Conference Day 2	Conference Day 3		
Time	Monday 25-May	Tuesday 26-May	Wednesday 27-May	Thursday 28-May		
9:00	Canadian AI - CRV Workshops	Opening Remarks (DFA 300)				
9:15						
9:30		Keynote Speaker Mac Schwager	Keynote Speaker Ian Stavness	Invited Talk: CIPPRS Dissertation Award in Computer Vision		
9:45						
10:00						
10:15	Coffee (Halpern Centre 126)	Coffee (DFA 300)	Coffee (DFA 300)	Coffee (DFA 300)		
10:30						
10:45						
11:00	Canadian AI - CRV Workshops	Oral Session 1 Chair: Nils Wilde	Oral Session 3 Chair: Xinxin Zuo	Oral Session 5 Chair: Yani Ioannou		
11:15						
11:30						
11:45						
12:00						
12:15						
12:30	Lunch Dining Commons	Lunch Dining Commons	Lunch Dining Commons	Lunch Dining Commons CIPPRS AGM (CRV Main Room)		
12:45						
13:00						
13:15						
13:30						
13:45						
14:00	Canadian AI - CRV Workshops	Oral Session 2 Chair: Nandita Vijaykumar	Oral Session 4 Chair: Risto Ojala	Oral Session 6 Chair: Mahdis Bisheban		
14:15						
14:30						
14:45						
15:00						
15:15						
15:30	Coffee (Halpern Centre 126)	Coffee (DFA 300)	Coffee (DFA 300)	Coffee (DFA 300)		
15:45						
16:00	Canadian AI - CRV Workshops	Poster Lighting Talks + Poster Session 1	Poster Lighting Talks + Poster Session 2	Nectar Track		
16:15						
16:30				Closing Remarks		
16:45						
17:00						
17:15						
17:30						
17:45						
18:00						
18:15						
18:30						
18:45						
19:00	Welcome Reception (Diamond Alumni Centre)	Banquet and Awards Ceremony Executive Plaza & Suites				
19:15						
19:30						
19:45						
20:00						
20:15						
20:30						
20:45						
21:00						
22:00						

2 Keynote Speakers

Mac Schwager

Stanford University

Talk Title: How general are generalist robot policies? Data scaling, diagnostic tools, and memorization in VLAs

Abstract: Vision-Language-Action (VLA) policies have recently emerged as a promising paradigm for generalist robot autonomy. However, VLAs have several challenges that must be overcome before they can achieve their potential. Firstly, these models require fine tuning with human-teleoperation demonstrations, which can be tedious, expensive, and time consuming to collect. Secondly, policy performance is limited to teleop demonstration quality, which can be highly variable depending on the human teleoperator’s skill and the dexterity barrier of the teleop interface. Lastly, VLA models, with the current state of practice, appear to suffer from strong overfitting to the fine-tuning data. All of these issues lead to “generalist” policies that do not generalize very well. In this talk I will describe recent work in my lab to address each of these problems. I will describe techniques we have developed to scale up demonstration data by leveraging 3D Gaussian Splatting models and optimization-based planning experts to generate arbitrary volumes of high-quality visual demonstrations to augment or replace human teleop data. I will describe our work on multi-task progress models that can track, based on visual inputs and text prompts, the progress of a demonstration. This can be used to filter human teleop data for high quality training data, and can be used as an online performance monitor during policy execution for fault detection, recovery guidance, and diagnostics. Finally, I will describe our work on memorization vs generalization in visuo-motor policies, where we find that current fine tuning practices cause overfitting to the training data, limiting a VLA’s generalization capabilities. I will explore some remedies for this problem. The talk will include experimental results for drone navigation policies, drone aerial manipulation policies, and table-top manipulation policies.

Biography: Dr. Schwager is an Associate Professor of Aeronautics and Astronautics at Stanford University, with a courtesy appointment in Computer Science. He directs the Multi-robot Systems Lab (MSL) where he studies robot autonomy. He is interested in learning-based autonomy for UAVs, manipulators, and robotic vehicles, 3D mapping and SLAM, analytical and statistical tools for verifiable safety in learning-based autonomy, and collaborative intelligence in groups of robots and human-robot teams. He obtained his BS degree from Stanford, his MS and PhD degrees from MIT, and he was a postdoctoral researcher at the University of Pennsylvania and MIT. He received the NSF CAREER award in 2014, the DARPA YFA in 2018, and has received numerous best paper awards including the IEEE Transactions on Robotics Best Paper Award (2016), Best Paper Award in Robot Manipulation (ICRA 2018), and Best Paper Award in Multi-Robot Systems (ICRA).

Ian Stavness

University of Saskatchewan

Talk Title: Advances in 3D capture and feedforward modeling for visual perception

Abstract: Breakthroughs in 3D radiance-field rendering and feedforward models that directly infer 3D structure are rapidly reshaping what’s possible in 3D visual perception for metrology, robotics, and beyond. In this talk, I will survey the fast-moving frontier of 3D Gaussian splatting, 3D tokenization for transformer architectures, and emerging perceptual pipelines that lift

2D semantic understanding into fully reconstructed 3D worlds with semantic labels. These new paradigms deliver striking gains precisely where traditional photogrammetry struggles most: in highly cluttered environments, scenes rich in fine-scale structure, and objects that are slender, flat, or otherwise difficult to capture with conventional geometry pipelines. I will ground these advances in the demanding real-world challenge of measuring agronomic plants that are densely packed, self-occluded, and often highly self-similar. Accurate 3D plant capture unlocks new opportunities for plant breeding, digital agriculture, and large-scale plant phenotyping. I will conclude with a discussion of the human-factor considerations that shape how people perceive, interact with, and interpret the large-scale 3D information that is enabled these new 3D capture and modeling methods.

Biography: Dr. Stavness is a Professor and the Head of the Department of Computer Science at the University of Saskatchewan. He holds a Research Chair at the Global Institute for Food Security and is the Director of the CREATE in Computational Agriculture training program. He obtained his PhD from the University of British Columbia and was a Postdoctoral Fellow in Bio-Engineering at Stanford University prior to joining the University of Saskatchewan in 2012. His research focuses on machine learning, computer vision, and computer graphics with applications in biology, agriculture, and medicine.

3 Symposium Speakers

Nils Wilde

Dalhousie University

Talk Title: User Preferences and Trade-offs in Robot Planning

Abstract: Real-world robot deployment requires adaptation to end-user needs. This often involves finding trade-offs between opposing criteria to align with user preferences. We explore two sides of the problem. First, we study how human-in-the-loop learning, i.e., repeated, simple interactions such as choosing among two presented robot trajectories, enables inexperienced users to quickly refine planning algorithms to their needs. Second, we study the problem of designing planning algorithms that attain all relevant trade-offs. Using direct treatment as multi-objective optimization, such problems are converted into single-objective formulations with tunable parameters, e.g., a cost function balancing trajectory length and risk with adjustable weights. We derive fundamental methods for exploring relevant weights based on error-approximations as well as novel formulations for scalar objectives with provable theoretical advantages. The presented methods are showcased in the context of path and motion planning and multi-robot coordination.

Biography: Dr. Nils Wilde is an Assistant Professor in Computer Science at Dalhousie University in Halifax, Canada, where he leads the Laboratory for Interactive Systems and Adaptive Robotics. Prior, he was a Postdoctoral Fellow at TU Delft and the University of Waterloo where he also completed his PhD in Electrical and Computer Engineering. Dr. Wilde’s research focuses on cognitive robotics, in particular human-robot interaction, preference learning and learning from human feedback, motion planning and multi-objective planning, as well as multi-robot coordination and task assignment. He is a member of the Atlantic AI Institute and an IEEE member. Dr. Wilde’s research is currently supported by NSERC and has been published in top-tier robotics journals and conferences, including T-RO, RA-L, IJRR, CoRL, WAFR, ICRA, IROS and CDC. Further, he co-organized workshops on Human Multi-Robot Interaction at IROS 2023 and on Multi-Objective

Optimization and Planning in Robotics at RSS 2025.

Nandita Vijaykumar

University of Toronto

Talk Title: Architecting Efficient and Scalable Systems for Physical Intelligence and Visual Computing

Abstract: A next frontier in intelligent systems lies in enabling machines to perceive and interact with the physical world—powering applications from robotics and self-driving cars to AR/VR and content creation. These systems must perceive and represent complex 3D environments, render photorealistic content, and generate interactive outputs—all under tight constraints on latency, memory, and scalability. In this talk, I will explore the systems challenges that arise in emerging visual computing pipelines, and how they push the limits of today’s abstractions for memory, compute, and programmability. I will then discuss some of our recent research on building across-the-stack frameworks that offer better primitives for 3D vision, differentiable rendering, and generative pipelines—spanning hardware architecture support, compiler and runtime design, and memory and storage hierarchies.

Biography: Nandita Vijaykumar is an Assistant Professor in the Department of Computer Science at the University of Toronto, where she leads the embARC research group. She is also a faculty member at the Vector Institute for Artificial Intelligence and a Research Scholar at Amazon. She received her Ph.D. from Carnegie Mellon University and has previously worked at AMD, Intel, Microsoft, and Nvidia. Her research explores the intersection of computer systems/architecture with visual computing, including computer vision, robotics, and machine learning. She is particularly interested in building efficient, scalable, and programmable systems that enable machines to perceive, interpret, and interact with the physical world. Her research has been supported by industry and academic partners including Intel, AMD, LG, Nvidia, Sony, Rebellions, NSERC, MITACS, CentML, and CSE Canada. She is a recipient of the Connaught New Researcher Award, the Benjamin Garver Lamme Fellowship, was a Qualcomm Fellowship Finalist, and has been inducted into the ISCA Hall of Fame.

Xinxin Zuo

Concordia University

Talk Title: Generative Models Should Adapt: Fast Test-Time Personalization for Video and 3D Editing

Abstract: Generative models have recently achieved remarkable advances in video synthesis and 3D editing. Despite this progress, most current methods remain fundamentally static at inference time, relying on a single frozen model to handle diverse subjects, scenes, and geometric configurations. In this talk, I will present a broader research perspective that challenges this assumption. Rather than expecting one pretrained model to generalize perfectly to every test instance, we can develop generative systems that are trained to adapt rapidly, efficiently, and reliably at inference time. This viewpoint opens new opportunities for personalized video generation, geometry-aware 3D editing, and, more generally, adaptive generative models that are better aligned with the demands of real-world deployment.

Biography: Dr. Xinxin Zuo is an Assistant Professor in the Department of Electrical and Computer Engineering at Concordia University, where she leads the X-Lab. Before joining Concordia,

she was a Staff Researcher at Huawei Canada and a Postdoctoral Fellow at the University of Alberta. Her research interests span machine learning and computer vision, with a recent focus on 3D AI-generated content (AIGC), embodied AI, human motion generation, and 3D reconstruction. She currently serves as an Associate Editor for IEEE Transactions on Multimedia. She has published over 50 papers and has received more than 3,000 citations.

Risto Ojala

Aalto University

Talk Title: Perception solutions for enabling automated driving in winter conditions

Abstract: This talk presents methods and findings from research on perception solutions for automated driving in winter conditions, carried out at the Autonomy & Mobility Laboratory, Aalto University. Winter conditions pose several challenges for automated vehicle perception pipelines, which currently limit the applicability of the technology in adverse weather conditions. The talk focuses on two main research directions: denoising snowflakes from LiDAR data and road segmentation in snowy conditions. Airborne snowflakes introduce significant noise into LiDAR scans, which can hinder downstream perception tasks. To address this challenge, the talk presents deep learning approaches for point cloud denoising based on both supervised and self-supervised learning. In addition, snowy conditions drastically alter the visual appearance of the environment and the road, rendering road segmentation methods trained on traditional datasets unreliable. To overcome this, trajectory-based approaches leveraging vision foundation models are presented for learning varied road appearance without requiring manual labeling.

Biography: Risto Ojala, DSc (Tech) is an Assistant Professor at Aalto University, Finland, where he leads the Autonomy & Mobility Laboratory within the Mechatronics research group. His research focuses on intelligent vehicles and mobile robotics, with particular emphasis on perception, sensor fusion, and applied machine learning for autonomous systems. He is also currently a Visiting Scholar at Simon Fraser University, Canada, collaborating with the Multi-Agent Robotic Systems Laboratory on research in semantic understanding for mobile robotics. His work develops perception solutions that enable robust autonomous operation in challenging environments. A central application of his research is automated driving in winter conditions, addressing problems such as road understanding, situational awareness, and perception reliability. He has published extensively in leading robotics and intelligent transportation venues and collaborates closely with both academic and industrial partners.

Yani Ioannou

University of Calgary

Talk Title: Expert Pruning in Sparsely-activated Mixture-of-Expert Models

Abstract: Sparsely-activated Mixture-of-Experts (SMoE) models offer efficient pre-training and low latency, but their large parameter counts create significant memory overhead, motivating research into expert compression. In our recent ICLR 2026 paper, we find that expert pruning is a superior strategy for generative tasks to expert merging which has been evaluated predominantly on discriminative benchmarks. We demonstrate that existing merging techniques introduce an irreducible error due to the loss of fine-grained routing control over experts. Leveraging this insight, we propose Router-weighted Expert Activation Pruning (REAP), a novel pruning criterion that considers both router gate-values and expert activation norms to minimize the reconstruction error

bound. Across a diverse set of SMOE models ranging from 20B to 1T parameters, REAP consistently outperforms merging and other pruning methods on generative benchmarks, especially at 50% compression. Notably, our method achieves near-lossless compression on code generation tasks with Qwen3-Coder-480B and Kimi-K2, even after pruning 50% of experts.

Biography: Yani Ioannou is an Assistant Professor and Schulich Research Chair in the Department of Electrical and Software Engineering of the Schulich School of Engineering, at the University of Calgary in Canada, Alberta. Yani was previously a Visiting Researcher at Google Brain Toronto (DeepMind) with Dr. Geoffrey Hinton, and a Post-doctoral Fellow at the Vector Institute with Dr. Graham Taylor. Yani completed his PhD at the University of Cambridge in 2018 supported by a Microsoft Research Ph.D. Scholarship, where he was supervised by Dr. Roberto Cipolla and Dr. Antonio Criminisi.

Maciek Popik presenting on behalf of Mahdis Bisheban

University of Calgary

Talk Title: Intelligent Dynamics and Control for Autonomous Robotic Systems in Complex Environments

Abstract: Autonomous robotic systems are increasingly deployed in complex and uncertain environments, from infrastructure inspection to search and rescue missions. However, achieving reliable autonomy in such settings remains a significant challenge due to dynamic conditions, uncertain models, and real-time decision-making requirements. In this talk, I will present recent advances from the Intelligent Dynamics and Control Lab at the University of Calgary, focusing on the integration of control theory, machine learning, and robotics. Specifically, I will discuss our work on aerial manipulation, and modelling under uncertainty.


Biography: Dr. Mahdis Bisheban is an Assistant Professor in the Department of Mechanical and Manufacturing Engineering at the University of Calgary and the Founder and Director of the Intelligent Dynamics and Control Lab (IDCL). She earned her Ph.D. in Mechanical and Aerospace Engineering from The George Washington University and completed postdoctoral research at Queen’s University. At IDCL, her research focuses on the intersection of advanced robotics for aerospace applications, machine learning, and intelligent control systems, with an emphasis on developing autonomous aerial and ground robots that can think, adapt, and collaborate. Beyond research, Dr. Bisheban is actively engaged in the professional community as the AIAA V/STOL Technical Committee Education Chair, an Associate Editor for the Transactions of the Canadian Society for Mechanical Engineering, and a member of the Canadian Society for Mechanical Engineering Mechatronics, Robotics, and Controls Technical Committee. She is committed to training the next generation of engineers and researchers, mentoring postdoctoral fellows, graduate and undergraduate students, and hosting high school students each summer. IDCL is distinguished by its collaborative, multi-level mentoring culture, where learners at all stages teach, learn, and contribute meaningfully to research.

4 Oral Sessions

Oral Session 1 — Day 1


Probing Zero-Shot Subtask Generalization in Vision-Language-Action Models

Grigorii Guz (University of British Columbia, Vector Institute for AI), Giuseppe Carenini (University of British Columbia), Mathias Léculyer (University of British Columbia), Michiel van de Panne (University of British Columbia), Vered Shwartz (University of British Columbia, Vector Institute for AI)

Recent robotic vision-language-action (VLA) models have shown impressive zero- and few-shot capabilities when deployed in unseen environments and robot morphologies. However, while natural language is a convenient way to specify tasks, it remains unclear how reliably VLAs can follow previously unseen language instructions after adaptation to new domains. This capability is particularly important for multi-task settings where collecting data and finetuning models for each potential task is impractical. To investigate unseen instruction generalization, we evaluate how well VLAs finetuned on a set of high-level tasks (place block in drawer, stack blocks) perform on the constituent low-level subtasks (grasp block, lift grasped block), and compare this to models finetuned directly on those subtasks. This evaluation protocol isolates unseen instruction understanding from the model’s physical task execution capabilities. We find that even for bigger VLAs, the performance gap between high-level vs. subtask finetuning does not shrink consistently. Overall, our results indicate that beyond model scaling, fine-grained robot data annotation and appropriate data collection protocols are crucial for improving the multi-task capabilities of existing robotic VLA policies. 

IRIS: Learning-Driven Task-Specific Cinema Robot Arm for Visuomotor Motion Control


Qilong Cheng (New York University), Matthew Mackay (University of Toronto), Ali Berezhi (University of Toronto)

Robotic camera systems enable dynamic and repeatable motion beyond human capabilities. Yet their adoption is restricted by high costs and operational complexity of industrial-grade hardware. We present intelligent robotic imaging system (IRIS), a task-specific 6-DOF manipulator, designed for autonomous learning-driven cinematic motion control. Our system leverages a vertically integrated stack that combines a lightweight, 3D-printed hardware design with a visuomotor imitation learning framework. By employing a goal-conditioned adaptation of Action Chunking with Transformers (ACT), IRIS learns to execute object-aware, perceptually smooth trajectories directly from human expert demonstrations, without the need for explicit geometric programming. The complete system costs under \$1,000 USD, supports a 1.5 kg payload, and achieves approximately 1 mm repeatability. Real-world experiments demonstrate accurate tracking and reliable autonomous execution that generalizes across variations of targeted push-in trajectories. Implementation details for IRIS are available at: <https://github.com/thejerrycheng/iris>. 

A Vision-Based Hybrid RL–PID Control Framework for Underwater Object Tracking in 6-DOF Swimming Robots


Farnoosh Faraji (McGill University), Khalil Virji (McGill University), Faraz Lotfi (McGill University), Nicholas Dudek (Independent Robotics Inc), Gregory Dudek (McGill University)

In this paper, we investigate the feasibility of replacing conventional, decoupled PID controllers with a centralized deep RL controller for underwater target-following tasks. Autonomous underwater vehicles (AUVs) traditionally rely on Proportional-Integral-Derivative (PID) controllers to regulate motion, but these controllers require careful hand tuning and do not naturally adapt to changes in environmental conditions, vehicle dynamics, or mission objectives. In contrast, reinforcement

learning (RL) learns control policies directly through interaction with the environment, enabling complex behaviors and improved robustness to disturbances and model uncertainty. Using a highly maneuverable six-degree-of-freedom hexapod AUV, we compare a traditional PID strategy with a neural policy trained via deep Q-learning in a framework that ensures safe exploration and leverages classical control during training. Our experiments show that while well-tuned PID controllers perform reliably under nominal conditions, the RL-based controller exhibits greater adaptability to distribution shifts, environmental variability, and interactions among coupled degrees of freedom. We discuss key trade-offs between classical and learning-based control, and highlight practical considerations for integrating RL in real-world underwater systems. Although our evaluation focuses on diver following, the proposed methodology is applicable to a wide range of AUV platforms and underwater tasks. 

Safe Reinforcement Learning with Contrastive Risk Prediction


Hanping Zhang (Carleton University), Yuhong Guo (Carleton University, CIFAR AI Chair, Amii)

As safety violations can lead to severe consequences in real-world applications, the increasing deployment of reinforcement learning (RL) in safety-critical domains such as robotics has propelled the study of safe reinforcement learning (safe RL). In this work, we propose a risk preventive training method for safe RL, which learns a binary classifier based on contrastive sampling to predict the probability of a state-action pair leading to unsafe states. Based on predicted risk probabilities, risk preventive trajectory exploration and reward shaping-based optimality criterion modification are simultaneously conducted to induce safe RL policies. We conduct experiments in robotic simulation environments. The results show the proposed approach outperforms existing model-free safe RL approaches, and yields comparable performance with the state-of-the-art model-based method. 

Oral Session 2 — Day 1


CascadedViT: Cascaded Chunk-FeedForward and Cascaded Group Attention Vision Transformer

Srivathsan Sivakumar (Ontario Tech University), Faisal Z. Qureshi (Ontario Tech University)

Vision Transformers (ViTs) have demonstrated remarkable performance across a range of computer vision tasks; however, their high computational, memory, and energy demands hinder deployment on resource-constrained platforms. In this paper, we propose Cascaded-ViT (CViT), a lightweight and compute-efficient vision transformer architecture featuring a novel feedforward network design called Cascaded-Chunk Feed Forward Network (CCFFN). By splitting input features, CCFFN improves parameter and FLOP efficiency without sacrificing accuracy. Experiments on ImageNet-1K show that our CViT-XL model achieves 75.5% Top-1 accuracy while reducing FLOPs by 15% and energy consumption by 3.3% compared to EfficientViTM5. Across various model sizes, the CViT family consistently exhibits the lowest energy consumption, making it suitable for deployment on battery-constrained devices such as mobile phones and drones. Furthermore, when evaluated using a new metric called Accuracy-Per-FLOP (APF), which quantifies compute efficiency relative to accuracy, CViT models consistently achieve top-ranking efficiency. Particularly, CViT-L is 2.2% more accurate than EfficientViT-M2 while having comparable APF scores. 


Parameter-Efficient Invariance via Equivariant Object Detection Networks

Kallin M. Kehrig (University of Regina), Howard J. Hamilton (University of Regina), Timothy D. Oleskiw (University of Regina)

Leading convolutional object detection neural networks, such as YOLO variants, have proven highly effective at generalizing across feature symmetries such as rotation and scale, but often do so at the cost of increased parameter counts that complicate model training and inference. Following recent work in the field of equivariant convolutions, this work seeks to exploit rotational and scale symmetries present in image data for parameter-efficient invariance. We propose three object detection models based on the YOLOv9-t (Y9t) architecture that demonstrate improved transformation invariance directly attributable to equivariant convolutions, namely for rotation C_4 -Y9t, S-Y9t for scale and C_4 S-Y9t for both, which maintain a similar number of trainable parameters to Y9t. To directly evaluate the impact of rotation- and scale-invariant methods in object detection networks, we propose the MNIST Object-Detection (MOD) datasets. Using the proposed models, mean average precision (mAP) is additively increased by 16.8%, 15.2%, and 3% on unseen rotation, scale, and rotation-scale transformed partitions of the MOD datasets, respectively. Further, the proposed models generalize to the standard dataset used to evaluate object detection (COCO 2017), with our best model increasing mAP by 4.9%. However, the models are not effective on the aerial object detection dataset HIT-UAV due to our imposed parameter limitations and the large number of feature transformations mapping to a single class that are not represented by the introduced symmetry groups, such as many additional camera angles and light levels. 

Distilling 3D Spatial Reasoning into a Lightweight Vision-Language Model with CoT


Alaa Asfour (Toronto Metropolitan University), Christopher Indris (Toronto Metropolitan University), Leihan Chen (Toronto Metropolitan University), Tejas Vyas (Toronto Metropolitan University), Guanghui Wang (Toronto Metropolitan University)

Large-scale 3D vision-language models (VLMs), such as LLaVA-3D, exhibit strong spatial reasoning capabilities but face significant deployment challenges due to their computational demands. We propose a knowledge distillation framework that transfers 3D spatial reasoning competence from a 7B-parameter teacher model to a compact 2.29B-parameter student model, achieving an $8.7\times$ reduction in inference latency and a $3\times$ reduction in model size while retaining 54-72% of the teacher’s performance on specialized spatial reasoning tasks. Our approach integrates VGGT (Visual Geometry Grounded Transformer) as the vision encoder and introduces a novel multi-task distillation pipeline with uncertainty-aware loss weighting. In addition, we introduce Hidden Chain-of-Thought (CoT): a fixed set of learnable “thinking” tokens that function as an internal scratchpad before the final answer. To the best of our knowledge, this is the first application of latent scratchpad reasoning to distilled 3D VLMs from a teacher model (LLaVA-3D-7B), requiring no CoT-capable teacher or explicit chain-of-thought data. Within a unified architecture, the student model is trained to generate spatial descriptions, estimate depth, and detect objects. Central contributions include (i) spatial feature alignment across multi-view inputs, (ii) adaptive task loss weighting, and (iii) Hidden CoT for enhanced reasoning without latering the user-facing interface. Experimental evaluation on ScanNet and 3D-FRONT datasets demonstrates that the distilled model retains robustspatial relationship understanding, achieving 68-72% accuracy in proximity and contact reasoning tasks, despite reduced text generation performance. The framework enables practical deployment of 3D VLMs on resource-constrained platforms while preserving the core spatial reasoning abilities required for robotics, augmented reality, and autonomous navigation applications. The source code will be available upon publication. 

Oral Session 3 — Day 2

NEEDL-Bench: Dataset for Swiss Needle Cast and Stomata Detection in Microscopy Images


Benjamin Blake (University of Victoria), Declan McIntosh (University of Victoria), Jürgen Ehrling (University of Victoria), Nicolas Feau (Pacific Forestry Centre Canadian Forest Service, Natural Resources Canada), Joey B. Tanney (Pacific Forestry Centre Canadian Forest Service, Natural Resources Canada), Alexandra Branzan Albu (University of Victoria)

Swiss Needle Cast (SNC) is a fungal disease that affects Douglas-fir trees, a keystone species of major ecological and economic importance as a softwood timber resource. The fungus forms sexual reproductive structures (pseudothecia) that emerge through the gas exchange pores (stomata) of the needles, thereby blocking gas exchange and compromising needle function. To date, there is no dataset for automatic computer vision detection of these structures, despite computer vision being well poised to standardize and viably scale severity measurements. To address this, we present NEEDL-Bench, a dataset of 3250 annotated images from 1082 Douglas-fir needles, annotated for both keypoints and bounding-box detectors. This dataset exhibits a challenging collection of features, including blur, poor object contrast, small objects of interest, and occlusions. To better capture both the nominal distribution of the data and the full breadth of rare structures, we present two distinct evaluation splits: either random sampling from the collected images or sequential sampling to maximize structural diversity. We evaluate multiple popular keypoint and bounding box methods for detection on this dataset as a baseline and observe a maximum F1 score of 0.8479, suggesting significant potential for gains from future development on this problem. Further, we find that larger models generally do not show commensurate gains in performance on this dataset, indicating that improvements on this problem will not come from scaling laws but rather from domain-specific inductive biases. 

Data-Geometry-Aware Ensemble for Reducing Uncertainty in Skin Lesion Classification


Farnaz Sheikhi (University of Calgary)

While deep learning models achieve increasingly high accuracy, their reliability remains a critical concern for high-risk clinical decision-making. In this work, we propose a *Data-Geometry-Aware (DGA) Ensemble* that reduces predictive uncertainty caused by sparse feature-space coverage and improves calibration, particularly for underrepresented classes. The proposed approach integrates feature-space geometric modeling, class-adaptive decision boundary estimation, and diffusion-based synthetic data generation to selectively augment uncertain regions of class manifolds. We evaluate the DGA ensemble on the benchmark HAM10000 skin lesion dataset and conduct a comprehensive comparison against models with varying architectural complexity, including a single ResNet-50, ensembles of ResNet-50 models (with and without metadata), mixture-of-experts variants, multimodal fusion models, and state-of-the-art baselines. Evaluation metrics include the standard classification performance (accuracy, precision, recall, and F1-score), uncertainty and calibration measures (ECE, Brier score, NLL, and predictive entropy), and bias analysis across patient age and sex. Experimental results demonstrate that the proposed DGA ensemble consistently outperforms competing methods in both predictive performance and calibration quality. Moreover, while baseline models exhibit systematic performance disparities favoring non-male patients under 50, the DGA ensemble significantly mitigates demographic bias across age and sex groups. These findings highlight the effectiveness of data-geometry-aware ensembling in improving accuracy, uncertainty

estimation, and fairness for reliable clinical decision support. 

Weld Joint Generation from 3D Point Clouds Using Gaussian-Weighted Potential Fields


Chaimae Belmarouf (University of Quebec at Rimouski), Tan Sy Nguyen (University of Quebec at Rimouski), Noureddine Barka (University of Quebec at Rimouski)

In this paper, a Gaussian-weighted potential field approach has been developed to effectively automate the detection of weld joint trajectories from 3D point cloud data. This method employs principal component analysis (PCA) for coordinate alignment, followed by cross-sectional slicing perpendicular to the weld direction. A novel potential function that combines lateral position weighting with elevation analysis has been integrated to identify optimal path points along the weld seam. Distance-based percentile outlier removal and local polynomial smoothing have been used to produce a clean final trajectory. Some experiments on multiple weld geometries are tested to explore the efficiency of the proposed approach. The results show that the method provides a potential tool for generating weld trajectories across varying joint configurations. It also provides a flexible framework for future development in automated welding path planning and related multi-axis manufacturing applications. 

Oral Session 4 — Day 2


Synthetic-to-Real Pipeline for Safe Landing Zone Detection

Shrikant Banerjee (Toronto Metropolitan University), Reza Faieghi (Toronto Metropolitan University)

As Uncrewed Aerial Vehicles (UAVs) transition toward higher levels of autonomy, the ability to perform unassisted recovery in non-cooperative, unstructured environments becomes critical. Achieving safe autonomous landing requires high-fidelity semantic resolution to distinguish navigable terrain from hazardous obstacles, yet development is often hindered by the scarcity of annotated aerial datasets. This work proposes a comprehensive perception and data generation pipeline designed to bridge the sim-to-real gap for autonomous landing tasks. We introduce a procedural synthetic data engine that generates photorealistic urban environments with automated semantic annotations through domain randomization. A Transformer-based OneFormer architecture is fine-tuned exclusively on this synthetic data, leveraging multi-head self-attention mechanisms for global context resolution. To ensure operational safety, a deterministic landing module utilizes a Euclidean Distance Transform (EDT) and dynamic inference logic to identify the largest inscribed safe landing zones while maintaining strict clearance buffers around obstacles. Quantitative benchmarking against the UAVid dataset demonstrates robust semantic segmentation performance, while qualitative validation on real-world UAV footage confirms the system's ability to identify collision-free landing sites in unseen environments. Our results highlight the potential of high-fidelity procedural simulation to eliminate the need for manual annotation while providing robust, edge-deployable situational awareness for autonomous UAV recovery. 


Zero-Shot Grasping from Local Surface Geometry

Jayalekshmi Jayakumar (Ingenuity Labs, Queen's University), Eric Haden (Ingenuity Labs, Queen's University), Matthew Pan (Ingenuity Labs, Queen's University), Michael Greenspan (Ingenuity Labs, Queen's University)

Robotic grasping of novel objects remains challenging, especially in unstructured environments without object models or prior knowledge. Many learning-based methods rely on global representations or large datasets, limiting generalization. We present a zero-shot grasp-prediction framework that operates solely on local surface geometry. A parallel-jaw grasp is represented as a pair of 3D surface patches corresponding to gripper contacts, and grasp feasibility is formulated as a binary classification over patch pairs. Our network is trained entirely on procedurally generated synthetic data using geometric constraints, without object meshes, CAD models, or RGB information. The model learns to identify graspable contact configurations directly from point cloud geometry. Experiments demonstrate high accuracy on synthetic data, effective transfer to standard object benchmarks, and reliable real-world grasping of unseen objects. These results highlight the power of local geometry-driven representations for zero-shot, transferable grasp prediction. 

Cooperative 3D LiDAR-Based Target Tracking and Visibility-Aware Following for Multi-UAV Systems


Hugo Aranha (Instituto Superior Técnico), Rodrigo Ventura (Instituto Superior Técnico), Meysam Basiri (Instituto Superior Técnico)

This paper presents a cooperative target tracking and visibility-aware following pipeline for multi-UAV systems equipped with compact 3D LiDAR sensors. The proposed system combines a lightweight point-cloud detector, a distributed Unscented Kalman Filter for sharing compact target estimates between neighbouring UAVs, and a nonlinear model predictive controller for target following while maintaining line-of-sight and avoiding obstacles and neighbouring UAVs. The approach is designed for scenarios where compact LiDAR measurements are sparse and intermittent, especially at longer ranges. Simulation results show that cooperative estimate sharing improves tracking robustness by reducing peak errors and increasing measurement availability. In a cluttered following scenario, two UAVs track and follow a moving person while maintaining safe separation and visibility constraints. A real-world single-UAV experiment with a Livox Mid-360 further demonstrates the practical feasibility of compact LiDAR-based aerial target tracking. The results support the use of cooperative perception and visibility-aware planning for robust UAV target following under sparse sensing conditions. 

Ultra-wideband Time Difference of Arrival Indoor Localization: From Sensor Placement to System Evaluation

Wenda Zhao (University of Toronto), Abhishek Goudar (University of Toronto), Mingliang Tang (University of California Berkeley), Angela P. Schoellig (TU Munich, MIRMI)

Ultra-wideband (UWB) time difference of arrival (TDOA)-based localization has emerged as a scalable positioning solution for mobile robots, consumer electronics, and wearable devices, featuring good accuracy and reliability. While UWB TDOA-based localization systems rely on the deployment of UWB radio sensors as positioning landmarks, existing works often assume these placements are predetermined or study the sensor placement problem alone without evaluating it in practical scenarios. In this article, we bridge this gap by approaching the UWB TDOA localization from a system-level perspective, integrating sensor placement as a key component and conducting practical evaluation in real-world scenarios. Through extensive real-world experiments, we demonstrate the accuracy and robustness of our localization system, comparing its performance to the theoretical lower bounds. Using a challenging multi-room environment as a case study, we illustrate the full system construction process, from sensor placement optimization to real-world deployment. Our

evaluation, comprising a cumulative total of 39 minutes of real-world experiments involving up to five agents and covering 2608 meters across four distinct scenarios, provides valuable insights and guidelines for constructing UWB TDOA localization systems. 

Oral Session 5 — Day 3

Frequency-based View Selection in Gaussian Splatting Reconstruction


Monica M.Q. Li (Polytechnique Montreal), Pierre-Yves Lajoie (Polytechnique Montreal), Giovanni Beltrame (Polytechnique Montreal)

Recent advances in photorealistic 3D reconstruction—ranging from Neural Radiance Fields (NeRF) to Gaussian Splatting—have demonstrated impressive results, but their heavy computational demands pose challenges for online robotics applications. In this paper, we address the critical problem of active view selection to minimize the number of input images required for high-quality Gaussian Splatting reconstructions as a robot traverses an environment. Our key insight lies in a novel frequency-domain ranking strategy that identifies the most informative viewpoints, enabling the system to capture essential scene features without redundant observations. Experimental results show that our method not only reduces computational overhead but also preserves high-fidelity reconstructions, making it a promising solution for efficient 3D modeling in robotics and other resource-constrained applications. The code is available at <https://github.com/lemonci/gS-fft>.



MatEval: Evaluating Indoor 3D Scene Material Recovery from a Single Image


Dongchen Yang (Simon Fraser University), Manolis Savva (Simon Fraser University)

Converting a single image to a 3D scene with geometry, materials, and lighting is a challenging problem. While geometry reconstruction from a single view has been extensively studied, material recovery for the single-photo-to-scene task remains underexplored. Recent advances in differentiable procedural materials, inverse rendering, and texture generation can be potentially applied to this task. However, they have not been systematically applied to this task and evaluated in a benchmark. In this project, we establish a comprehensive benchmark for material recovery in the single-image-to-scene task. We evaluate three families of methods inspired by recent state-of-the-art approaches in inverse rendering, texture generation, and single-image-to-scene. Our results show that single-view inverse rendering baselines outperform procedural material baselines (19.40 vs 13.60 in PSNR for albedo on original views), highlighting the strong potential of methods based on single-view inverse rendering for material recovery in the single-image-to-scene task. We will release the full dataset, evaluation code, and baseline implementations to support future work. 

Toward explainable monocular depth estimation


Seyed Mohammad Hossein Hosseini (York University), James H Elder (York University), David N. White (York University), Amin Alizadeh Naeini (York University)

Self-supervised large-scale training on highly diverse datasets has led to a dramatic improvement in the accuracy of monocular depth estimation systems. Unfortunately, these performance improvements have not been concomitant with increased clarity in the principles these systems rely upon. This is potentially a problem for certifying these systems for performance-critical applications and for generalizing to new scenarios. Here we address this problem by assessing the role of three

principles that are thought to be important for human monocular depth estimation. The first of these is ground theory, the principle that monocular depth perception is anchored on recovery of the ground surface. The second is perceptual organization, the spatial chunking of image information to locally regularize depth estimation. The third is linear perspective, the recovery of depth information derived from regularities in the projection of parallel lines. We show that each of these three principles is important and that combining them yields a completely explainable monocular depth estimation system that in some respects can rival the performance of completely opaque deep network systems. We further show that many of these systems can benefit from fusion with our system, suggesting that training, while extensive, does not always succeed in fully encoding these principles. 

CtxPL: Context-based Prototype Learning for Weakly-Supervised Temporal Action Localization

Moayadeldin Hussain (St. Francis Xavier University), Iker Gondra (St. Francis Xavier University)


Weakly-Supervised Temporal Action Localization (WS-TAL) aims to identify action classes and locate action boundaries in untrimmed videos by only using video-level labels during training. Prototype-based learning is a promising approach that has recently been used to solve this challenging task. However, current solutions overlook the temporal context structure that surrounds the action of interest, leading to an increased number of false positive predictions and high action-background confusion. To address this, we present Context-based Prototype Learning (CtxPL), a novel framework that models both the class-specific characteristics and class-specific ambiguity into two different prototype sets, then concurrently utilizes the mined information for achieving an enhanced action-background separation. Our experiments show that CtxPL achieves superior performance compared to the previous prototype-based solutions in WS-TAL, providing more precise localization boundaries and better action comprehension. Our code is available through: <https://github.com/moayadeldin/CtxPL> 

Oral Session 6 — Day 3

Adaptive PD Gains for Energy-Conscious Control in Physical Human-Robot Interaction


Danyal Saqib (University of British Columbia), Francisco Javier Andrade Chavez (Thompson Rivers University), Marie Charbonneau (University of Calgary)

Compliant force or torque control are approaches often investigated to achieve safe physical human-robot interaction (pHRI). However, these approaches have limitations. Force control requires a robot to be equipped with external force sensors to track the amplitude and direction of applied forces. Torque control requires torque sensing or estimation in each joint. As this is not available on every robot, energy-based approaches offer a promising alternative. Such approaches aim to achieve safe pHRI by limiting the mechanical energy of the robot. Current schemes leveraging an energy-based approach tend to have a complex implementation, and some may require further stability verification. We hence propose an adaptive proportional derivative (PD) controller that can limit a robot's energy under any given limit to achieve safe pHRI. The proposed controller can limit both the kinetic and potential energy of a robot, and the behaviour of the controller gains can be shaped using various parameters, defining precisely the cutoff limit and sharpness. We construct a stability

proof for the controller and define a condition to ensure the controller’s stability. The proposed controller’s behaviour and compliance are tested on the TALOS robot from PAL Robotics both in simulation and on hardware, verifying the expected compliant and energy-limiting behaviour of the controller. 

Robust Integrated Planning and Control for Quadrotors in Dynamic Environments via NMPC with CBF Penalties

Zeinab Shayan (Toronto Metropolitan University), Mohammadreza Izadi (Toronto Metropolitan University), Reza Faieghi (Toronto Metropolitan University)

This paper presents a new robust integrated planning and control (IPC) strategy for multirotor uncrewed aerial vehicles. We propose a nonlinear model predictive control (NMPC) formulation that embeds control barrier functions (CBFs) as exponential penalties, improving feasibility while ensuring smooth obstacle avoidance under tight input bounds. The penalty weights provide a practical tuning knob to trade off tracking accuracy against avoidance aggressiveness. We enhance the system robustness by employing a high-gain disturbance observer (HGDO) to estimate and compensate for external disturbances. We also incorporate a Kalman filter (KF) for computationally efficient, real-time prediction of obstacle motion, enabling avoidance of moving obstacles. Comparative studies against both conventional NMPC and NMPC with hard CBF constraints, validated in Gazebo and hardware experiments, demonstrate superior feasibility, safety, and robustness. To the best of our knowledge, this is the first hardware-validated NMPC–CBF IPC framework, offering a practical step toward safe quadrotor deployment in dynamic environments. 

5 Poster Sessions & Lightning Talks

Robotics – Session 1

xModel-KD: Cross-modal Knowledge Distillation for 3D Scene Perception using LiDAR

Thenukan Pathmanathan (Lakehead University), Kanchan Keisham (Vellore Institute of Technology), Thangarajah Akilan (Lakehead University)

Point cloud segmentation is a fundamental task in 3D scene understanding. Its progress is constrained by the high cost and time required for dense 3D annotations, making labeled samples difficult to obtain. Beyond annotation scarcity, different sensing modalities face inherent limitations. 2D images provide rich texture and appearance cues, yet they lack explicit depth and geometric structure. In contrast, 3D point clouds capture accurate spatial geometry but are sparse and contain no texture information. As a result, relying on a single modality restricts the richness of learned representations and weakens generalization. Although recent multi-modal methods that combine 3D point clouds with 2D images have demonstrated strong performance in tasks such as classification and retrieval, they typically depend on large-scale labeled datasets and have not been fully exploited for data-efficient dense prediction. To address these limitations, we propose a novel cross-modal knowledge distillation framework, xModel-KD, for 3D point cloud segmentation. Our method exploits the complementary strengths of 2D texture and 3D geometry by learning unified per-point representations through cross-modal alignment. Specifically, we design a cross-modal fusion encoder trained with a contrastive objective that enforces feature consistency between cor-

responding 2D and 3D representations across multiple views. By integrating powerful pre-trained backbones with a targeted fusion strategy, the proposed framework effectively transfers appearance cues from images to geometry-aware point features. Experimental results show that cross-modal fusion achieves a 2% absolute improvement in mIoU over a LiDAR-only baseline, demonstrating the benefit of leveraging complementary multi-modal information for scalable and annotation-efficient 3D scene understanding. [↗](#)

Safe Local Navigation for Ackermann-Steered Robots in Unmapped Environments

Christian Schaible (University of Waterloo), Shahin Sirouspour (McMaster University)

A control framework is proposed for safe local navigation of mobile robots equipped with Ackermann steering in unmapped environments where a global goal is absent. Based on local obstacle detections, the safest heading angle is determined along the direction of the largest open space ahead of the vehicle. Guided by this direction, bounding lines are constructed on the left and right sides of the vehicle to achieve obstacle separation. These bounding lines are obtained by solving a convex quadratic optimization that maximizes vehicle-to-obstacle clearance. Optionally, conditions are imposed on the bounding lines to preserve parallelism and smooth abrupt changes from prior control steps. A feedback-linearizing controller is then used to regulate the vehicle's distance from one or both bounding lines, effectively enabling tracking of a local reference path that preserves safety through obstacle clearance maximization. Open-source code is included for the application of this control scheme. Experimental results demonstrate that the proposed method produces safer navigation paths with significantly shorter computation times, compared to some existing exploration-based planners. [↗](#)

Beyond Point Tracking: Learning Local Affine Deformations for Robust and Geometrically Consistent Feature Tracking


Nigel Dias (Federal University of Goias), GUSTAVO TEODORO LAUREANO (Federal University of Goias), Ronaldo Martins da Costa (Federal University of Goias)

In this paper, we revisit the canonical Kanade-Lucas-Tomasi (KLT) feature tracker. We reformulate this classical algorithm by integrating deep neural mechanisms for spatio-temporal and geometric learning. The proposed model is a fully convolutional neural network that regresses the parameters of an affine transformation that links patches across consecutive frames. The system architecture integrates a spatio-temporal feature extractor with a Cost Volume Block and a Regression Block. To train the model, we introduce a versatile protocol for synthesizing feature-tracking annotations from arbitrary datasets, suitable for VO and V-SLAM. The process yields a set of tracking states in which keypoints remain strictly geometrically consistent with the camera trajectory. Quantitative analysis on the TUM RGB-D benchmark demonstrates the consistent superiority of the proposed method in estimating relative camera motion over both the KLT baseline and Pips++. Although our method exhibits a lower inlier ratio, the resulting correspondence subset possesses significantly higher geometric fidelity. These results establish the proposed method as a robust solution suitable for deployment on resource-constrained embedded systems. [↗](#)

Bench-Push: Benchmarking Pushing-based Navigation and Manipulation Tasks for Mobile Robots

Ninghan Zhong (Georgia Institute of Technology), Steven Caro (University of Waterloo), Meghath Ramesh (University of Waterloo), Rishi Bhatnagar (University of Alberta), Avraiem Iskandar (University of Waterloo), Stephen L. Smith (University of Waterloo)

Mobile robots are increasingly deployed in cluttered environments with movable objects, posing

challenges for traditional methods that prohibit interaction. In such settings, the mobile robot must go beyond traditional obstacle avoidance, leveraging pushing or nudging strategies to accomplish its goals. While research in pushing-based robotics is growing, evaluations rely on ad hoc setups, limiting reproducibility and cross-comparison. To address this, we present Bench-Push, the first unified benchmark for pushing-based mobile robot navigation and manipulation tasks. Bench-Push includes multiple components: 1) a comprehensive range of simulated environments that capture the fundamental challenges in pushing-based tasks, including navigating a maze with movable obstacles, autonomous ship navigation in ice-covered waters, box delivery, and area clearing, each with varying levels of complexity; 2) novel evaluation metrics to capture efficiency, interaction effort, and partial task completion; and 3) demonstrations using Bench-Push to evaluate example implementations of established baselines across environments. Bench-Push is open-sourced as a Python library with a modular design. The code, documentation, and trained models can be found at <https://github.com/IvanIZ/BenchPush>. 

Empir3D:A Framework for Multi-Dimensional Point Cloud Assessment

Yash Turkar (University at Buffalo), Pranay Meshram (University at Buffalo), Christo Aluckal (University at Buffalo), Charuvahan Adhivarahan (University at Buffalo), Karthik K Dantu (University at Buffalo)


Advancements in sensors, algorithms and compute hardware has made 3D perception feasible in real-time. Current methods to compare and evaluate quality of a 3D model such as Chamfer, Hausdorff and Earth-mover’s distance are uni-dimensional and have limitations; including inability to capture coverage, local variations in density and error, and are significantly affected by outliers. In this paper, we propose an evaluation framework for point clouds (Empir3D) that consists of four metrics - resolution (Qr) to quantify ability to distinguish between the individual parts in the point cloud, accuracy (Qa) to measure registration error, coverage (Qc) to evaluate portion of missing data, and artifact-score (Qt) to characterize the presence of artifacts. Through detailed analysis, we demonstrate the complementary nature of each of these dimensions, and the improvement they provide compared to uni-dimensional measures highlighted above. Further, we demonstrate the utility of Empir3D by comparing our metric with the uni-dimensional metrics for dense SLAM map evaluation. We believe that Empir3D advances our ability to reason about point clouds and helps better debug 3D perception applications by providing richer evaluation of their performance.



Cross-Attention-Driven Quality Assessment for Robust Camera-LiDAR Sensor Balance

Zeping Zhang (University of Ottawa), Robert Laganiere (University of Ottawa)


We propose a cross-attention-driven Quality Assessment Fusion (QAF) framework that dynamically balances information from camera (RGB) and LiDAR sensors under diverse and potentially adverse driving conditions. At its core is a Quality Assessment Unit (QAU) that learns to identify and weight higher-quality sensor features while down-weighting degraded inputs such as blurred images or sparsely sampled point clouds. This adaptive scheme contrasts with simpler early-fusion baselines by explicitly modeling input reliability through intra-modal and cross-modal quality masks. In addition, we introduce a dedicated Distance Estimation Module to improve depth prediction, which is first pretrained with ground-truth bounding boxes and later fine-tuned using predicted boxes for robust generalization. Experiments on synthetic variations of the KITTI dataset—spanning snowy, foggy, and camera-lighting scenarios, as well as sparse LiDAR conditions—demonstrate that our QAF framework consistently outperforms naive early-fusion methods. Our results underscore the

importance of quality-aware sensor fusion and specialized distance-estimation strategies to ensure reliable perception in challenging real-world environments. 

Vision – Session 1


Distill3R: A Pipeline for Democratizing 3D Foundation Models on Commodity Hardware

Brandon Leblanc (Concordia University), Charalambos Poullis (Concordia University)

While multi-view 3D reconstruction has shifted toward large-scale foundation models capable of inferring globally consistent geometry, their reliance on massive computational clusters for training has created a significant barrier to entry for most academic laboratories. To bridge this compute divide, we introduce Distill3R, a framework designed to distill the geometric reasoning of 3D foundation models into compact students fully trainable on a single workstation. Our methodology centers on two primary innovations: (1) an offline caching pipeline that decouples heavy teacher inference from the training loop through compressed supervision signals, and (2) a confidence-aware distillation loss that leverages teacher uncertainty to enable training on commodity hardware. We propose a 72M-parameter student model which achieves a 9x reduction in parameters and a 5x inference speedup compared to its 650M-parameter teacher. The student is fully trainable in under 3 days on a single workstation, whereas its teacher requires massive GPU clusters for up to a week. We demonstrate that the student preserves the structural consistency and qualitative geometric understanding required for functional 3D awareness. By providing a reproducible, single-workstation training recipe, Distill3R serves as an exploratory entry point for democratized 3D vision research and efficient edge deployment. This work is not intended to compete with state-of-the-art foundation models, but to provide an accessible research baseline for laboratories without access to large-scale compute to train and specialize models on their own domain-specific data at minimal cost. 

Contour Refinement using Discrete Diffusion in Low Data Regime

Fei Yu Guan (University of Toronto), Ian Keefe (University of Toronto), Sophie L Wilkinson (Simon Fraser University), Daniel D. B. Perrakis (Pacific Forestry Centre, Natural Resources Canada), Steven L. Waslander (University of Toronto)

Boundary detection of irregular and translucent objects is an important problem with applications in medical imaging, environmental monitoring and manufacturing, where many of these applications are plagued with scarce labeled data and low in situ computational resources. While recent image segmentation studies focus on segmentation mask alignment with ground-truth, the task of boundary detection remains understudied, especially in the low data regime. In this work, we present a lightweight discrete diffusion contour refinement pipeline for robust boundary detection in the low data regime. We use a Convolutional Neural Network(CNN) architecture with self-attention layers as the core of our pipeline, and condition on a segmentation mask, iteratively denoising a sparse contour representation. We introduce multiple novel adaptations for improved low-data efficacy and inference efficiency, including using a simplified diffusion process, a customized model architecture, and minimal post processing to produce a dense, isolated contour given a dataset of size <500 training images. Our method outperforms several SOTA baselines on the medical imaging dataset KVASIR, is competitive on HAM10K and our custom wildfire dataset, Smoke, while improving inference framerate by 3.5X. 

On the Impact of Video Compression on Vision-Based Occupancy Prediction: A Con-

trolled Preliminary Study

Jürgen Leppsalu (University of Tartu Tartu Observatory), Joosep Kivastik (University of Tartu Tartu Observatory), Rando Avarmaa (University of Tartu Tartu Observatory)

Vision-based occupancy prediction for edge deployment must operate on compressed video due to storage and bandwidth limits. We quantify how lossy compression affects occupancy accuracy in a controlled simulation study. A multi-camera CARLA dataset was encoded with MJPEG, H.264, and H.265 across chroma formats (4:2:0/4:4:4) and quality settings, and models were trained/evaluated under matched (same codec/setting) and mismatched (train compressed, evaluate raw) conditions. Performance was measured using Mean Average Precision (mAP) and analysed with respect to the compression factor and perceptual metrics (PSNR, MS-SSIM, FSIM, LPIPS). Under matched conditions, mAP remains stable up to roughly $20\times$ compression ($\sim 95\%$ storage savings). H.264 and H.265 maintain robustness at high compression, whereas MJPEG degrades earlier; chroma subsampling has only a minor effect. With domain shift, prediction performance degrades slightly faster, highlighting the importance of training–deployment alignment. Notably, we observe that mild compression can outperform raw in this controlled setup, suggesting a possible regularisation effect. Correlation analysis with image-quality metrics show a strong association with mAP across compression settings. These preliminary results indicate that training and inferring directly on compressed data, and preferring modern inter-frame codecs over MJPEG, substantially reduce storage without sacrificing occupancy prediction reliability. [↗](#)

Bootstrapped Semantic-Aware Gaussian Splatting for Geometrically Accurate Outdoor Reconstruction

Rhea Joyce Zambra (University of Calgary), Hongzhou Yang (University of Calgary)


3D representation of unbounded outdoor scenes from multi-view images remains challenging due to depth ambiguity, weak texture, and heterogeneous scene structure. While 3D Gaussian Splatting achieves high visual fidelity, it remains largely appearance-driven and can produce geometric artifacts and depth inconsistencies. We present a semantic-aware reconstruction pipeline that integrates semantic Structure-from-Motion with bootstrapped Gaussian Splatting for class-specific geometric reasoning. By leveraging semantic priors during sparse reconstruction and Gaussian optimization, our method improves geometric fidelity while preserving rendering quality. Experiments on Waymo and nuScenes show that GO-SfM improves downstream rendering by up to 33% in PSNR and 17% in SSIM, with average gains of approximately 6% and 4%, respectively. Qualitative results further show reduced floaters, semantic bleeding, and cross-surface blending compared to standard 3DGS and PGSR. [↗](#)

Robotics – Session 2

AMP2026: A Multi-Platform Marine Robotics Dataset for Tracking and Mapping


Edwin Meriaux, Shuo Wen, David Widhalm, Zhizun Wang, Junming Shi, Mariana Sosa Guzmán, Kalvik Jakkala, Bennett A. Carley, Elias Sokolova, Yogesh Girdhar, Monika Roznere, Jason O’Kane, Junaed Sattar, Gregory Dudek

Marine environments present significant challenges for perception and autonomy due to dynamic surfaces, limited visibility, and complex interactions between aerial, surface, and submerged sensing modalities. This paper introduces the Aerial–Marine Perception Dataset (AMP2026), a multi-platform marine robotics dataset collected across multiple field deployments designed to support

research in two primary areas: multi-view tracking and marine environment mapping. The dataset includes synchronized data from aerial drones, boat-mounted cameras, and submerged robotic platforms, along with associated localization and telemetry information when available. The goal of this work is to provide a publicly available dataset enabling research in marine perception and multi-robot observation scenarios. This paper describes the data collection methodology, sensor configurations, dataset organization, and intended research tasks supported by the dataset. 


Enhancing Visual Odometry with Reliable Pixel Masking

Yash Turkar (University at Buffalo), Timothy Chase (University at Buffalo), Christo Aluckal (University at Buffalo), Karthik K Dantu (University at Buffalo)

Robust feature detection and matching are fundamental for visual odometry and SLAM, yet most methods lack a principled measure of a feature’s reliability prior to downstream use. We present Reliable Pixel Masking (PIXER), a learning-based method that predicts a dense pixel-wise estimated uncertainty map for candidate feature locations. PIXER is a lightweight, single-shot model trained with variance supervision from a Bayesian version of a baseline detector. At inference, a deterministic reliability head predicts this map without Monte Carlo sampling. The map is thresholded into a binary reliability mask, which filters high-uncertainty, low-reliability features prior to matching, improving downstream matching and pose estimation. Integrated into a standard visual odometry pipeline, PIXER reduces average trajectory error by 31% while reducing feature usage by 49% across eight different feature detectors. Our results demonstrate that feature selection based on learned uncertainty estimates enhances the robustness and efficiency of SLAM systems. 

Hybrid Dynamics Modeling for a Flexible 2-DoF Robotic Arm


Maciek Popik (University of Calgary), Daniel Yang (University of Calgary), Mahdis Bisheban (University of Calgary)

This report examines three approaches for modeling the dynamics of a flexible-link 2-DoF robotic arm to address unmodeled dynamics not captured by rigid-body models. Two physics informed models combine rigid-body dynamics (RBD) formulations with a Gaussian Mixture Model (GMM) to capture residual model errors and linkage flexibility. A kinematics-based regression model serves as a purely data-driven baseline. Using an open-source dataset, torque predictions are first estimated using Ridge regression on kinematic features, while the physics-based baseline is constructed from published specifications, and ordinary least-squares regression is subsequently used to estimate the same parameter set directly from data. Results show that the physics-based parameters yield the poorest accuracy, while regularized and least-squares estimators align more closely with measured torques. Residual analysis and error metrics highlight the limitations of purely parametric models for flexible-link systems and underscore the value of regularization and data-driven identification, supporting future developments of semi-parametric residual learning methods. 

Fixed-Time Dynamic Landing of Quadrotors using Adaptive Unscented Kalman Filtering and Nonlinear Model Predictive Control


Mohammadreza Izadi (Toronto Metropolitan University), Zeinab Shayan (Toronto Metropolitan University), Steven L. Waslander (University of Toronto), Reza Faieghi (Toronto Metropolitan University)

This paper introduces an estimation and control framework for dynamic landing of multi-rotor uncrewed aerial vehicles on moving platforms. The proposed method integrates nonlinear model predictive control with a real-time minimum-jerk trajectory planner that enforces a prescribed touchdown time, enabling consistent timing during the terminal descent. To enhance robustness

in the presence of time-varying sensing quality, we utilize an adaptive unscented kalman filter that updates the process and measurement noise statistics online. In addition, we provide a reference feasibility analysis showing that minimum-jerk references induce bounded thrust and torque commands under standard tracking hypotheses. The proposed framework is evaluated in simulation and hardware experiments, and it is shown to achieve repeatable landings and improved platform velocity prediction accuracy relative to EKF/UKF-based methods. 

Vision-Based Semantic SLAM for Autonomous Navigation in Mill Yard

Junrui Huang, Inna Sharf, Elie Ayoub, Nicolas Lemieux, Heshan Fernando

Log-loading machines are essential in mill-yard operations for unloading logs from incoming transport trucks onto mill infeed deck, as well as managing log inventory in stockpiles. This paper focuses on the log-loading operation in the vicinity of the infeed deck, with the goal of enabling higher levels of autonomy in this task. Near the infeed deck, the machine must localize reliably relative to the infeed deck and adjacent buffer piles, while also detecting and localizing arriving trucks and trailers; this is a highly dynamic outdoor environment. We present a vision-based semantic SLAM system that uses a stereo camera mounted on the log-loading machine as the sole perception sensor. The proposed pipeline is based on stereo ORB-SLAM2 for real-time pose estimation and mapping. It integrates a parallel semantic thread that converts stereo depth into pseudo-LiDAR point clouds and predicts oriented 3D bounding boxes for key objects, including the infeed deck, log piles, and log trucks. The estimated 3D bounding boxes are used to remove features on potentially dynamic objects during SLAM tracking for improving robustness, and to construct a persistent object-level semantic map by transforming 3D bounding boxes into the global SLAM frame. We evaluated the system in a virtual NVIDIA Isaac Sim infeed-deck environment using synthetic stereo image sequences. The evaluation reports camera trajectory accuracy, semantic object localization accuracy, and runtime performance, and includes ablations to isolate the impact of dynamic-feature removal and object-level semantic mapping. The results indicate that incorporating object-level 3D detections improves the robustness and accuracy of stereo SLAM in dynamic infeed-deck scenes while producing a globally consistent semantic map in practical runtime. 

Vision – Session 2

Benchmarking YOLOv11 Against RF-DETR for Defect Detection in Automotive Heat Shields

Saber Yu (Ascension Automation Solutions Ltd.), Mengchen Liu (McMaster University), Timothy Reinhart (Ascension Automation Solutions Ltd.), Seshasai Srinivasan (McMaster University), Zhen Gao (McMaster University)

Industrial defect detection often faces challenges from limited defective samples and low data variations. Through evaluating different augmentation techniques on YOLOv11-Nano using a custom dataset of automotive heat shields, we developed an optimized strategy that improved the model’s mean recall from 76.43% (without augmentation) to 97.86% (+21.43%). Comparative benchmarking of YOLOv11 on the augmented data against RF-DETR, RT-DETR, and YOLOv12 demonstrated that RF-DETR-Base (100% recall — 98.26% precision — 40 FPS) and YOLOv11-Nano (97.86% recall — 98.04% precision — 103 FPS) were the best performing models at 0.25 IOU and 0.25 confidence thresholds, with the former excelling in detection accuracy and the latter in inference speed. Our findings align with recent COCO benchmarks, where RF-DETR emerges as the latest state-of-the-art architecture, and confirm the model’s capability for industrial applications.



Few-Shot Adaptation of CLIP for Security X-Ray Image Analysis: An Empirical Study

Shamita Datta (Concordia University), Yang Wang (Concordia University, MILA - Quebec AI Institute), Xinxin Zuo (Concordia University)

X-ray image analysis plays a crucial role in security screening, but progress is constrained due to the lack of large-scale, labeled datasets required for training deep models. Contrastive Language-Image Pretraining (CLIP) offers a promising alternative, as it has shown impressive zero-shot generalization on natural image benchmarks by aligning visual and textual representations. However, its performance degrades on security X-ray imagery, where domain-specific characteristics are absent from pre-training. To address this gap, we explore a few-shot adaptation strategy that enables CLIP to specialize in the X-ray domain under limited supervision. We present a systematic study comparing adapter-based fine-tuning, full-model fine-tuning, and low-rank adaptation for classification and detection tasks. Experiments on the OPIXray dataset show that LoRA and full fine-tuning substantially outperform adapter-based methods in X-ray image classification. On various X-ray datasets, our CLIP-augmented Faster R-CNN method consistently improves over standard few-shot detection baselines. These results demonstrate that both parameter-efficient and full fine-tuning approaches can successfully adapt CLIP to the X-ray domain, advancing few-shot learning in high-stakes security applications.

Semantic Segmentation of Node and Edge Diagrams for Assistive Technology

Michael Cormier (Mount Allison University), Yichun Zhao (University of Victoria), Laura Paul (University of Victoria), Cameron Swift (Mount Allison University), Duc Tri Dang (Mount Allison University), Miguel Nacenta (University of Victoria)

In this paper, we present a novel set of related models for semantic segmentation of node-link diagrams. These diagrams are frequently used to represent mathematical graphs, relationships between concepts, and flowcharts. Such diagrams are difficult to access non-visually; while some assistive interfaces have been designed for node-link diagrams, they rely upon a machine-readable representation of the diagram, whereas such diagrams will generally be made available as bitmap images. Our compact deep learning models show excellent quantitative and qualitative performance on a large synthetic dataset of node-link diagrams, reaching per-pixel accuracy over 93%.

Brain Age Prediction in Neurodegenerative Disorders via Autoencoder Architectures

Justin Rozeboom (University of Alberta), Ethan Tu (Pitzer College), Dana Cobzas (MacEwan University), Aakanksha Singh (University of Alberta), Karteek Popuri (Memorial University of Newfoundland), Mirza Faisal Beg (Simon Fraser University), The Canadian ALS Neuroimaging Consortium (CALSNIC), Nilanjan Ray (University of Alberta), Sanjay Kalra (University of Alberta)

As global populations age and the prevalence of neurodegenerative diseases increases, it is crucial to develop reliable neuroimaging biomarkers that can identify deviations from the normal aging trajectory to aid in the detection of neurodegenerative pathology. Brain age prediction, leading to the Predicted Age Difference (PAD) metric, provides an analytical tool for assessing neuroanatomical health from non-invasive magnetic resonance imaging (MRI) scans. In this study, we propose a dual-task ResNet-style autoencoder architecture designed to learn robust latent representations of brain structure for accurate age estimation. Our brain age prediction framework demonstrates strong generalization performance for age prediction from relatively smaller clinical cohorts, achieving performance competitive with a state-of-the-art benchmark (SFCN-reg) when trained and evaluated on the same data. Using the age prediction model, we evaluated the PAD biomarker on 502 sub-

jects across Alzheimer’s Disease (AD), Amyotrophic Lateral Sclerosis (ALS), and healthy control (HC) groups. Our results show that while AD was linked to accelerated apparent brain age (Mean PAD =6.49 years; Cohen’s $d=1.08$), ALS presented a more subtle and heterogeneous disease profile (Mean PAD =1.56 years; $d=0.19$). PAD strongly correlated with cognitive scores and posterior cortical thinning in AD, and correlated with ventricular expansion and corpus callosum atrophy in ALS with weak clinical correlations. The weaker sensitivity in ALS highlights the challenges of using global age biomarkers for system-specific neurodegeneration. This work presents a high-performing, anatomically grounded framework for brain age prediction that differentiates between two distinct disease trajectories. [↗](#)

6 AI-CRV Nectar Track

Please join us for a series of talks for works accepted at recent conferences and venues.

Chairs: Evan Shelhamer and Junhui Li

1. Space-Time Graphs of Convex Sets for Multi-Robot Motion Planning

Area: Robotics

Authors: Jingtao Tang, Zining Mao, Lufan Yang, Hang Ma

Presenter: Jingtao Tang

Affiliation: Simon Fraser University

2. MotionScript: Natural Language Descriptions for Expressive 3D Human Motions

Area: Machine Learning

Authors: Payam Jome Yazdian, Rachel Lagasse, Hamid Mohammadi, Eric Liu, Li Cheng, Angelica Lim

Presenter: Payam Jome Yazdian

Affiliation: Simon Fraser University

3. GHOST: Solving the Traveling Salesman Problem on Graphs of Convex Sets

Area: Robotics

Authors: Jingtao Tang, Hang Ma

Presenter: Jingtao Tang

Affiliation: Simon Fraser University

4. FACT-GS: Frequency-Aligned Complexity-Aware Texture Reparameterization for 2D Gaussian Splatting

Area: Vision

Authors: Tianhao Xie, Linlian Jiang, Xinxin Zuo, Yang Wang, Tiberiu Popa

Presenter: Tianhao Xie

Affiliation: Concordia University

5. Semantic Mapping in Indoor Embodied AI – A Survey on Advances, Challenges, and Future Directions

Area: Robotics

Authors: Sonia Raychaudhuri, Angel X. Chang

Presenter: Sonia Raychaudhuri

Affiliation: Simon Fraser University

6. Synthetic Geology: Structural Geology Meets Deep Learning

Area: Machine Learning

Authors: Simon Ghyselincks, Valeriia Okhmak, Stefano Zampini, George Turkiyyah, David Keyes, Eldad Haber

Presenter: Simon Ghyselincks

Affiliation: The University of British Columbia, Department of Computer Science

7. Threshold Strategy for Leaking Corner-Free Hamilton-Jacobi Reachability with Decomposed Computations

Area: Robotics

Authors: Chong He, Mugilan Mariappan, Keval Vora, Mo Chen

Presenter: Chong He

Affiliation: Simon Fraser University

8. PointMAC: Meta-learned Adaptation for Robust Test-time Point Cloud Completion

Area: Vision

Authors: Linlian Jiang, Rui Ma, Li Gu, Ziqiang Wang, Xinxin Zuo, Yang Wang

Presenter: Linlian Jiang

Affiliation: Concordia University, MILA, Jilin University

9. Reach–Avoid Differential Game with Reachability Analysis for Unmanned Aerial Vehicles: A Decomposition Approach

Area: Robotics

Authors: Minh Bui, Simon Monckton, Mo Chen

Presenter: Minh Bui

Affiliation: Simon Fraser University

10. Asymmetric Duos: Sidekicks Improve Uncertainty

Area: Machine Learning

Authors: Tim G. Zhou, Evan Shelhamer, Geoff Pleiss

Presenter: Tim Zhou

Affiliation: UBC, Vector

11. Implicit Maximum Likelihood Estimation for Real-time Generative Model Predictive Control

Area: Robotics

Authors: Grayson Lee, Minh Bui, Shuzi Zhou, Yankai Li, Mo Chen, Ke Li

Presenter: Shuzi Zhou

Affiliation: Simon Fraser University

12. UCAN: Unified Convolutional Attention Network for Expansive Receptive Fields in Lightweight Super-Resolution

Area: Vision

Authors: Cao Thien Tan, Trang Phan Thi Thu, Duc Nghiem Do, Ho Ngoc Anh, Hanyang Zhuang, Duc Dung Nguyen

Presenter: Duc Nghiem Do

Affiliation: University of Manitoba