21st Conference on Robots and Vision

Rozanski Hall, University of Guelph, 98 Trent Ln
Guelph, Ontario

May 28 - May 31, 2024

Detailed Program

Contents

1 Program at a Glance

Main Room: Rozanski Hall Room 103

Legend	
🟧	CRV Session
🟩	Breaks & Social

	Conference Day 1	Conference Day 2	Conference Day 3	Workshop Day
Time	Tuesday 28-May	Wednesday 29-May	Thursday 30-May	Friday 31-May
8:30	Coffee	Coffee	Coffee	Coffee
8:45				
9:00	Welcome Remarks			Workshop Session - Vision Chair: Renjie Liao
9:15				
9:30	Oral Session 1 Chair: Vincent Sitzmann	Oral Session 3 Chair: Jeong Joon Park	Oral Session 5 Chair: David Lindell	
9:45				
10:00				
10:15				
10:30	Coffee	Coffee	Coffee	Coffee
10:45				
11:00	Keynote Speaker Jitendra Malik	Invited Talks: CIPPRS Dissertation Awards in Computer Vision and Robotics	Keynote Speaker Kirstin Petersen	Workshop Keynote Speaker Igor Gilitschenski
11:15				
11:30				
11:45				
12:00	Lunch University Centre	Lunch University Centre	Lunch University Centre CIPPRS AGM 12:30 to 13:30 (CRV Main Room)	Lunch University Centre
12:15				
12:30				
12:45				
13:00				
13:15				
13:30	Oral Session 2 Chair: Audrey Sedal	Oral Session 4 Chair: Mo Chen	Oral Session 6 Chair: Yue Hu	Workshop Session - Robotics Chair: Yue Hu
13:45				
14:00				
14:15				
14:30	Featured Speaker Qixing Huang	Featured Speaker Bryan Tripp	Featured Speaker Leonid Sigal	
14:45				
15:00				
15:15	Coffee	Coffee	Coffee	Coffee
15:30				
15:45	Campus Tour	Lighting Talks for Posters	Closing Remarks	Workshop Keynote Speaker Francisco Javier Andrade Chavez
16:00				
16:15				
16:30				
16:45				
17:00				
17:15				
17:30		Conference and Workshop Poster Reception (Joint with Canadian AI) Summerlee Science Complex - Waasamowin (formerly Summerlee Science Complex Atrium)		
17:45				
18:00			Banquet and Award Ceremony Summerlee Science Complex - Waasamowin (formerly Summerlee Science Complex Atrium)	
18:15				
18:30				
18:45				
19:00	Welcome Reception Rozanski Lobby			
19:15				
19:30				
19:45				
20:00				
20:15				
20:30				
20:45				

2 Keynote Speakers

In alphabetical order

Jitendra Malik

University of California, Berkeley

Talk Title: Reconstructing and Recognizing Human Actions in Video

Abstract: Humans are social animals. Perhaps this is why we so enjoy watching movies, TV shows and YouTube videos, all of which show people in action. A central problem for artificial intelligence therefore is to develop techniques for analyzing and understanding human behavior from images and video. I will present some recent results from our research group towards this grand challenge. We have developed highly accurate techniques for reconstructing 3D meshes of human bodies from single images using transformer neural networks. Given video input, we link these reconstructions over time by 3D tracking, thus producing "Humans in 4D" (3D in space + 1D in time). As a fun application, we can use this capability to transfer the 3D motion of one person to another e.g. to generate a video of you performing Michael Jackson's moonwalk or Michelle Kwan's skating routine. The ability to do 4D reconstruction of hands is a source of imitation learning for robotics and we show examples of reconstructing human-object interactions. In addition to 4D reconstruction, we are also now able to recognize actions by attaching semantic labels such as "standing", "running", or "jumping". However, long range video understanding, such as the ability to follow characters' activities and understand movie plots over periods of minutes and hours, is still quite a challenge, and even the largest vision-language models struggle on such tasks. There has been substantial progress, but much remains to be done.

Biography: Jitendra Malik is the Arthur J. Chick Professor in the Department of Electrical Engineering and Computer Sciences at UC Berkeley. He is also part-time Research Scientist Director at Meta. Malik's research group has worked on many different topics in computer vision, human visual perception, robotics, machine learning and artificial intelligence, and he has mentored nearly 80 PhD students and postdocs. His honors include the 2013 IEEE PAMI-TC Distinguished Researcher in Computer Vision Award, the 2014 K.S. Fu Prize from the International Association of Pattern Recognition, the 2016 ACM-AAAI Allen Newell Award, the 2018 IJCAI Award for Research Excellence in AI, and the 2019 IEEE Computer Society Computer Pioneer Award. He is a member of the National Academy of Engineering and the National Academy of Sciences, and a fellow of the American Academy of Arts and Sciences.

Kirstin H. Petersen

Cornell University

Talk Title: Robot Superorganisms

Abstract: Natural swarms exhibit sophisticated colony-level behaviors with remarkable scalability and error tolerance. Their evolutionary success stems from more than just intelligent individuals, it hinges on their morphology, their physical interactions, and the way they shape and leverage their environment. Mound-building termites, for instance, are believed to use their own body as a template for construction; the resulting dirt mound serves, among other things, to regulate volatile pheromone cues which in turn guide further construction and colony growth. Throughout this talk I will argue how we can leverage the same principles to achieve greater performance in robot collectives, by paying attention to the interplay between control and hardware, as well as direct- and environmentally-mediated coordination between robots. I will exemplify the strength and challenges of this approach through soft robot collectives, collective robotic construction, and micro-scale robot collectives.

Biography: Kirstin Petersen is an Associate Professor and Aref and Manon Lahham Faculty Fellow in the School of Electrical and Computer Engineering at Cornell University. Her lab, the Collective Embodied Intelligence Lab, is focused on design and coordination of robot collectives able to achieve complex behaviors beyond the reach of an individual, and corresponding studies on how social insects do so in nature. Major research topics include swarm intelligence, embodied intelligence, soft robots, and bio-hybrid systems. Petersen did her postdoc at the Max Planck Institute for Intelligent Systems and her PhD at Harvard University and the Wyss Institute for Biologically Inspired Engineering. Her graduate work was featured in and on the cover of Science, she was elected among the top 25 women to know in robotics by Robohub in 2018, and received the Packard Fellowship in Science and Engineering in 2019 and the NSF CAREER award in 2021.

3 Featured Speakers

In alphabetical order

Qixing Huang

University of Texas at Austin

Talk Title: Geometric Regularizations for 3D Shape Generation

Abstract: Generative models, which map a latent parameter space to instances in an ambient space, enjoy various applications in 3D Vision and related domains. A standard scheme of these models is probabilistic, which aligns the induced ambient distribution of a generative model from a prior distribution of the latent space with the empirical ambient distribution of training instances. While this paradigm has proven to be quite successful on images, its current applications in 3D generation encounter fundamental challenges in the limited training data and generalization behavior. The key difference between image generation and shape generation is that 3D shapes possess various priors in geometry, topology, and physical properties. Existing probabilistic 3D generative approaches do not preserve these desired properties, resulting in synthesized shapes with various types of distortions. In this talk, I will discuss recent work that seeks to establish a novel geometric framework for learning shape generators. The key idea is to model various geometric, physical, and topological priors of 3D shapes as suitable regularization losses by developing computational tools in differential geometry and computational topology. We will discuss the applications in deformable shape generation, latent space design, joint shape matching, and 3D man-made shape generation.

Biography: Qixing Huang is an associate professor with tenure at the computer science department of the University of Texas at Austin. His research sits at the intersection of graphics, geometry, optimization, vision, and machine learning. He has published more than 100 papers at leading venues across these areas. His recent research is on 3D generation, focusing on integrating domain specific knowledge in geometry, physics, and topology, and learning 3D foundation models. He has won an NSF Career award and multiple best paper awards in graphics and vision.

Leonid Sigal

University of British Columbia

Talk Title: Opportunities and Limitations of Foundational and Vision-Language Models

Abstract: The capabilities and the use of foundational (FM) and vision-language (VLM) models (LLMs) in computer vision have exploded over the past 1-2 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. Specifically, I will discuss three avenues of research. First, I will discuss our semi-recent work that deals with building foundational image representation models by combining two successful strategies of masking (e.g., BERT) and sequential token prediction (e.g., GPT). We find that such a combination results in a better, more efficient and transferable pre-training strategy. Second, I will discuss a series of papers focusing on text-to-image (TTI) generative models, where we introduce a novel autoregressive diffusion-based framework with a visual memory module that implicitly captures the actor and background for multi-frame story visualization. This design is able to maintain consistency and resolve references in longer story text. Third, I will discuss biases in such models and our work on bias quantification and mitigation in the TTI models.

Biography: Prof. Leonid Sigal is a Professor at the University of British Columbia (UBC). He is also currently a part-time Visiting Researcher at Google. 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.

Bryan Tripp

University of Waterloo

Talk Title: The gap between deep networks and the brain

Abstract: Deep networks have roots in early efforts to model brain function, and their internal activations are among the best predictors of brain activity. Biological brains outperform deep networks in their versatility, sample efficiency, power efficiency, and real-world autonomy, suggesting that the brain may be a source of insight into how to further improve deep networks. However, the brain has many complexities, and it is unclear which of them are important. This talk will describe some first steps in developing functional, anatomically and physiologically realistic brain models based on deep networks, to better understand how deep networks should be elaborated to close the gap. This work points away from vision transformers and suggests a new parameter space for convolutional networks.

Biography: Bryan Tripp is an Associate Professor in the Department of Systems Design Engineering and the Centre for Theoretical Neuroscience at the University of Waterloo. His lab studies intelligence from the perspectives of computational neuroscience, applied deep learning, and robotics. Before joining the University of Waterloo, he was a post-doctoral fellow at McGill University, studying visual neuroscience.

4 Symposium Speakers

In alphabetical order

Mo Chen

Simon Fraser University

Talk Title: Control and Learning in Robotic Decision Making and Human Motion Prediction

Abstract: The combination of control theory and machine learning is becoming increasingly important, and being able to get the best of both worlds would unlock many robotic applications. In this talk, we will first discuss connections between control and reinforcement learning, and how they can enable more data-efficient, generalizable, and interpretable robot learning. Afterwards, we will discuss how ideas from control can be incorporated into deep learning methods to guide long-term human motion prediction.

Biography: Mo Chen is an Assistant Professor in the School of Computing Science at Simon Fraser University, Burnaby, BC, Canada, where he directs the Multi-Agent Robotic Systems Lab. He holds a Canada CIFAR AI Chair position and is an Amii Fellow. Dr. Chen completed his PhD in the Electrical Engineering and Computer Sciences Department at the University of California, Berkeley in 2017, and received his BASc in Engineering Physics from the University of British Columbia in 2011. From 2017 to 2018, He was a postdoctoral researcher in the Aeronautics and Astronautics Department in Stanford University. Dr. Chen's research interests include multi-agent systems, safety-critical systems, human-robot interactions, control theory, reinforcement learning, and their intersections.

Yue Hu

University of Waterloo

Talk Title: Engaging with Collaborative Robots: Insights from Human Factors

Abstract: Research in Human-Robot Interaction (HRI) has evolved into two distinct branches: physical HRI (pHRI), focusing on task efficiency and safety, and social HRI (sHRI), which examines human perceptions. To achieve collaboration and coexistence between humans and robots, a new perspective is essential. In this talk, I will explore experimental studies on active physical interactions between humans and collaborative robots. I will discuss critical human factors involved, detailing methodologies to measure and quantify these interactions form a diverse perspective.

Biography: Dr. Yue Hu has been an Assistant Professor at the Department of Mechanical and Mechatronics Engineering at the University of Waterloo since September 2021 where she is the Head of the Active and Interactive Robotics Lab. Yue obtained her doctorate in robotics from Heidelberg University, Germany in 2017. She was a postdoc first at Heidelberg University, then at the Italian Institute of Technology (IIT), in Italy. Between 2018 and 2021 she was first a JSPS (Japan Society for the Promotion of Science) fellow at the National Institute of Advanced Industrial Science and Technology (AIST) in Japan, and then an Assistant Professor at the Department of Mechanical Systems Engineering, Tokyo University of Agriculture and Technology. She is one of the co-chairs of the IEEE-RAS Technical Committee on Model-based Optimization for Robotics. Her research interests include physical human-robot interaction, collaborative robots, humanoid robots, and optimal control. Yue is also on the Advisory Board of the not-for-profit organization Women in AI & Robotics.

David Lindell

University of Toronto

Talk Title: Flying with Photons: Rendering Novel Views of Propagating Light

Abstract: In this talk I discuss an imaging and neural rendering technique that seeks to synthesize videos of light propagating through a scene from novel, moving camera viewpoints. Our approach relies on a new ultrafast imaging setup to capture a first-of-its kind, multi-viewpoint video dataset with picosecond-level temporal resolution. Combined with this dataset, we introduce an efficient neural volume rendering framework based on the transient field. This field is defined as a mapping from a 3D point and 2D direction to a high-dimensional, discrete-time signal that represents time-varying radiance at ultrafast timescales. Rendering with transient fields naturally accounts for effects due to the finite speed of light, including viewpoint-dependent appearance changes caused by light propagation delays to the camera. I will demonstrate time-resolved visualization of complex, captured light transport effects, including scattering, specular reflection, refraction, and diffraction. Finally, I will discuss future directions in propagation-aware inverse rendering.

Biography: David Lindell is an Assistant Professor in the Department of Computer Science at the University of Toronto. His research combines optics, emerging sensor platforms, machine learning, and physics-based algorithms to enable new capabilities in visual computing. Prof. Lindell's research has a wide array of applications including autonomous navigation, virtual and augmented reality, and remote sensing. Prior to joining the University of Toronto, he received his Ph.D. from Stanford University. He is a recipient of the 2021 ACM SIGGRAPH Outstanding Dissertation Honorable Mention Award and the 2023 Marr Prize.

Jeong Joon Park

University of Michigan

Talk Title: Towards compositional 3D scene generation

Abstract: Recently, numerous 3D generative model approaches have been proposed to automatically produce highly realistic objects. However, producing larger-scale scenes, rather than objects, still stands as a formidable challenge. In this talk, I'll show my past work on scene-scale generative models. I'll start by discussing a 3D-aware diffusion model technique that auto-regressively accumulates image-aligned 3D features for scene generations. Next, I will discuss other recent works that exploit compositional structures of large scenes to effectively produce scenes, which is difficult for non-compositional approaches.

Biography: Jeong Joon (JJ) Park is an assistant professor at the University of Michigan, Ann Arbor, in the Computer Science and Engineering Department. His research interests lie in the intersection of computer vision and graphics, where he studies realistic reconstruction and generation of 3D scenes using neural and physical representations. Generations of large-scale, dynamic, and interactive 3D scenes are his current primary targets. His group explores 3D vision and graphics, and their applications to robotics, medical imaging, and scientific problems. He is the lead author of DeepSDF, which introduced neural implicit representation to 3D computer vision. Before coming to Michigan, he was a postdoctoral researcher at Stanford University and a Ph.D. student at the University of Washington, supported by Apple AI/ML Fellowship. He did his undergraduate studies in computer science at Caltech.

Audrey A. Sedal

McGill University

Talk Title: Simulation-Driven Soft Robotics

Abstract: Soft-bodied robots present a compelling solution for navigating tight spaces and interacting with unknown obstacles, with potential applications in inspection, medicine, and AR/VR. Yet, even after a decade, soft robots remain largely in the prototype phase without scaling to the tasks where they show the most promise. These systems are difficult to design and control because their morphology is coupled with both their actuation and the environment, creating a large joint space that cannot be exhaustively explored through prototype iteration. Soft roboticists need new tools to repeatably develop systems that leverage deformability and contact. Dr. Sedal will first present recent work on jointly optimizing the design and control of soft robots in simulation. Through a combination of reduced-order finite element simulation and reinforcement learning, this work trained soft-legged, crawling robots, achieving performance that surpassed expert baselines and transferred to real physical results. Second, Dr. Sedal will present work on deformable acoustic tactile sensors and design with auxetic meta-materials. The research presented here will enable engineering tools for development of intelligent structures with high compliance for high-contact settings, taking soft robots out of the laboratory and into the world.

Biography: Dr. Sedal is an Assistant Professor at McGill University in Montreal, Canada where she leads the MACRObotics (Morphology, Actuation and Computation for Robotics) research group. She is also an Associate Member (Academic) of Mila, Quebec AI Institute. Prior, she was a Research Assistant Professor (comparable to endowed postdoc) at TTI-Chicago. She holds a PhD and MSc from the University of Michigan, as well as a BSc from MIT, all in Mechanical Engineering.

Vincent Sitzmann

Massachusetts Institute of Technology

Talk Title: Enabling New Robotic Capabilities with Spatial AI

Abstract: Recent progress in 3D computer vision has enabled a set of previously impossible capabilities. In this talk, I will present a set of results at the cutting edge of 3D computer vision and scene representation, revolving around imbuing 3D representations with a semantic understanding of the underlying 3D scene, neural networks that learn to solve the structure-from-motion problem and are capable of learning to reconstruct interpretable 3D scenes just from unprocessed video. I will relate all of these results to capabilities that they enable in robotics, and finally give an outlook on near-term results at the interface of robotics and vision.

Biography: Vincent Sitzmann is an Assistant Professor at MIT EECS, where he is leading the Scene Representation Group. Previously, he did his Ph.D. at Stanford University as well as a Postdoc at MIT CSAIL. His research interest lies in building models that perceive and model the world the way that humans do. Specifically, Vincent works towards models that can learn to reconstruct a rich state description of their environment, such as reconstructing its 3D structure, materials, semantics, etc. from vision. More importantly, these models should then also be able to model the impact of their own actions on that environment, i.e., learn a "mental simulator" or "world model". Vincent is particularly interested in models that can learn these skills fully self-supervised only from video and by self-directed interaction with the world.

5 Oral Sessions

Oral Session 1 — Day 1

Meta Episodic learning with Dynamic Task Sampling for CLIP-based Point Cloud Classification

Shuvozit Ghose (University of Manitoba) and Yang Wang (Concordia University)

Point cloud classification refers to the process of assigning semantic labels or categories to individual points within a point cloud data structure. Recent works have explored the extension of pre-trained CLIP to 3D recognition. In this direction, CLIP-based point cloud models like Point-CLIP, CLIP2Point have become state-of-the-art methods in the few-shot setup. Although these methods show promising performance for some classes like airplanes, desks, guitars, etc, the performance for some classes like the cup, flower pot, sink, nightstand, etc is still far from satisfactory. This is due to the fact that the adapter of CLIP-based models is trained using randomly sampled N-way K-shot data in the standard supervised learning setup. In this paper, we propose a novel meta-episodic learning framework for CLIP-based point cloud classification, addressing the challenges of limited training examples and sampling unknown classes. Additionally, we introduce dynamic task sampling within the episode based on performance memory. This sampling strategy effectively addresses the challenge of sampling unknown classes, ensuring that the model learns from a diverse range of classes and promotes the exploration of underrepresented categories. By dynamically updating the performance memory, we adaptively prioritize the sampling of classes based on their performance, enhancing the model's ability to handle challenging and real-world scenarios. Experiments show an average performance gain of 3-6% on ModelNet40 and ScanobjectNN datasets in a few-shot setup. ↗

Distribution and Depth-Aware Transformers for 3D Human Mesh Recovery

Jerrin Bright (University of Waterloo), Bavesh Balaji (University of Waterloo), Harish Prakash (University of Waterloo), Yuhao Chen (University of Waterloo), David A. Clausi (University of Waterloo), John S. Zelek (University of Waterloo)

Precise Human Mesh Recovery (HMR) with in-the-wild data is a formidable challenge and is often hindered by depth ambiguities and reduced precision. Existing works resort to either pose priors or multi-modal data such as multi-view or point cloud information, though their methods often overlook the valuable scene-depth information inherently present in a single image. Moreover, achieving robust HMR for out-of-distribution (OOD) data is exceedingly challenging due to inherent variations in pose, shape and depth. Consequently, understanding the underlying distribution becomes a vital subproblem in modeling human forms. Motivated by the need for unambiguous and robust human modeling, we introduce Distribution and depth-aware human mesh recovery (D2AHMR), an end-to-end transformer architecture meticulously designed to minimize the disparity between distributions and incorporate scene-depth leveraging prior depth information. Our approach demonstrates superior performance in handling OOD data in certain scenarios while consistently achieving competitive results against state-of-the-art HMR methods on controlled datasets. ↗

Oral Session 2 — Day 1

BACS: Background Aware Continual Semantic Segmentation

Mostafa ElAraby (Université de Montréal), Ali Harakeh (Université de Montréal), and Liam Paull (Université de Montréal)

Semantic segmentation plays a crucial role in enabling comprehensive scene understanding for robotic systems. However, generating annotations is challenging, requiring labels for every pixel in an image. In scenarios like autonomous driving, there's a need to progressively incorporate new classes as the operating environment of the deployed agent becomes more complex. For enhanced annotation efficiency, ideally, only pixels belonging to new classes would be annotated. This approach is known as Continual Semantic Segmentation (CSS). Besides the common problem of classical catastrophic forgetting in the continual learning setting, CSS suffers from the inherent ambiguity of the background, a phenomenon we refer to as the "background shift", since pixels labeled as background could correspond to future classes (forward background shift) or previous classes (backward background shift). As a result, continual learning approaches tend to fail. This paper proposes a Backward Background Shift Detector (BACS) to detect previously observed classes based on their distance in the latent space from the foreground centroids of previous steps. Moreover, we propose a modified version of the cross-entropy loss function, incorporating the BACS detector to down-weight background pixels associated with formerly observed classes. To combat catastrophic forgetting, we employ masked feature distillation alongside dark experience replay. Additionally, our approach includes a transformer decoder capable of adjusting to new classes without necessitating an additional classification head. We validate BACS's superior performance over existing state-of-the-art methods on standard CSS benchmarks. ↗

Change of Scenery: Unsupervised LiDAR Change Detection for Mobile Robots

Alexander D. Krawciw (University of Toronto), Jordy Sehn (University of Toronto), and Timothy D. Barfoot (University of Toronto)

This paper presents a fully unsupervised deep change detection approach for mobile robots with 3D LiDAR. In unstructured environments, it is infeasible to define a closed set of semantic classes. Instead, semantic segmentation is reformulated as binary change detection. We develop a neural network, RangeNetCD, that uses an existing point-cloud map and a live LiDAR scan to detect scene changes with respect to the map. Using a novel loss function, existing point-cloud semantic segmentation networks can be trained to perform change detection without any labels or assumptions about local semantics. The mean intersection over union (mIoU) score is used for quantitative comparison. RangeNetCD outperforms the baseline by 3.8% to 7.7% depending on the amount of environmental structure. The neural network operates at 67.1 Hz and is integrated into a robot's autonomy stack to allow safe navigation around obstacles that intersect the planned path. In addition, a novel method for the rapid automated acquisition of per-point ground-truth labels is described. Covering changed parts of the scene with retroreflective materials and applying a threshold filter to the intensity channel of the LiDAR allows for quantitative evaluation of the change detector. ↗

Oral Session 3 — Day 2

Domain-guided Masked Autoencoders for Unique Player Identification

Bavesh Balaji (University of Waterloo), Jerrin Bright (University of Waterloo), Sirisha Rambhatla (University of Waterloo), Yuhao Chen (University of Waterloo), Alexander Wong (University of Waterloo), John S. Zelek (University of Waterloo), David A. Clausi (University of Waterloo)

Unique player identification is a fundamental module in vision-driven sports analytics. Identifying players from broadcast videos can aid with various downstream tasks such as player assessment, in-game analysis, and broadcast production. However, automatic detection of jersey numbers using deep features is challenging primarily due to: a) motion blur, b) low resolution video feed, and c) occlusions. With their recent success in various vision tasks, masked autoencoders (MAEs) have emerged as a superior alternative to conventional feature extractors. However, most MAEs simply zero-out image patches either randomly or focus on where to mask rather than how to mask. Motivated by human vision, we devise a novel domain-guided masking policy for MAEs termed d-MAE to facilitate robust feature extraction in the presence of motion blur for player identification. We further introduce a new spatiotemporal network leveraging our novel d-MAE for unique player identification. We conduct experiments on three large-scale sports datasets, including a curated Baseball dataset, the SoccerNet dataset, and an in-house Ice Hockey dataset. We preprocess the datasets using an upgraded Keyframe Identification (KfID) module by focusing on frames containing jersey numbers. Additionally, we propose a keyframe-fusion technique to augment keyframes, preserving spatial and temporal context. Our spatiotemporal network showcases significant improvements, surpassing the current state-of-the-art by 8.58%, 4.29%, and 1.20% in the test set accuracies, respectively. Rigorous ablations highlight the effectiveness of our domain-guided masking approach and the refined KfID module, resulting in performance enhancements of 1.48% and 1.84% respectively, compared to original architectures.

Detection of Micromobility Vehicles in Urban Traffic Videos

Khalil Sabri (Polytechnique Montréal), Célia Djilali (Polytechnique Montréal), Guillaume-Alexandre Bilodeau (Polytechnique Montréal), Nicolas Saunier (Polytechnique Montréal), Wassim Bouachir (University of Québec)

Urban traffic environments present unique challenges for object detection, particularly with the increasing presence of micromobility vehicles like e-scooters and bikes. To address this object detection problem, this work introduces an adapted detection model that combines the accuracy and speed of single-frame object detection with the richer features offered by video object detection frameworks. This is done by applying aggregated feature maps from consecutive frames processed through motion flow to the YOLOX architecture. This fusion brings a temporal perspective to YOLOX detection abilities, allowing for a better understanding of urban mobility patterns and substantially improving detection reliability. Tested on a custom dataset curated for urban micromobility scenarios, our model showcases substantial improvement over existing state-of-the-art methods, demonstrating the need to consider spatiotemporal information for detecting such small and thin objects. Our approach enhances detection in challenging conditions, including occlusions, ensuring temporal consistency, and effectively mitigating motion blur.

Oral Session 4 — Day 2

Prediction of SLAM ATE Using an Ensemble Learning Regression Model and 1-D Global Pooling of Data Characterization

Islam Ali (University of Alberta), Bingqing (Selina) Wan (University of Toronto), and Hong Zong (University of Alberta)

Robustness and resilience in simultaneous localization and mapping (SLAM) are critical requirements for modern autonomous robotic systems. One of the essential steps to achieving robustness and resilience is the ability of SLAM to have an integrity measure for its estimates, thus having internal fault tolerance mechanisms to deal with performance degradation. In this work, we introduce a novel method for predicting SLAM localization error based on the characterization of raw sensor inputs. The proposed method relies on using a random forest regression model trained on 1-D global pooled features generated from characterized raw sensor data. The model is validated by using it to predict the performance of ORB-SLAM3 on three different datasets running in four different operating modes, resulting in an average prediction accuracy of up to 93.1% and 80.45% for ATE and APE, respectively. Then, the paper studies the quality of prediction with limited training data and proves that we can maintain proper ATE and APE prediction quality when training on only 20% and 40% of the data, respectively. Finally, the paper discusses the impact of out-of-distribution predictions on prediction accuracy. ↗

Towards Optimal Beacon Placement for Range-Aided Localization

Ethan Sequeira (McMaster University), Hussein Saad (McMaster University), Stephen Kelly (McMaster University), and Matthew Giamou (McMaster University)

Range-based localization is ubiquitous: global navigation satellite systems (GNSS) power mobile phone-based navigation, and autonomous mobile robots can use range measurements from a variety of modalities including sonar, radar, and even WiFi signals. Many of these localization systems rely on fixed anchors or beacons with known positions acting as transmitters or receivers. In this work, we answer a fundamental question: given a set of positions we would like to localize, how should beacons be placed so as to minimize localization error? Specifically, we present an information-theoretic method for optimally selecting an arrangement consisting of a few beacons from a large set of candidate positions. By formulating localization as maximum a posteriori (MAP) estimation, we can cast beacon arrangement as a submodular set function maximization problem. This approach is probabilistically rigorous, simple to implement, and extremely flexible. Furthermore, we prove that the submodular structure of our problem formulation ensures that a greedy algorithm for beacon arrangement has suboptimality guarantees. We compare our method with a number of benchmarks on simulated data and release an open source Python implementation of our algorithm and experiments. ↗

Oral Session 5 — Day 3

STF: Spatio-Temporal Fusion Module for Improving Video Object Detection

Noreen Anwar (Polytechnique Montréal), Guillaume-Alexandre Bilodeau (Polytechnique Montréal), Wassim Bouachir (University of Quebec)

Consecutive frames in a video contain redundancy, but they may also contain relevant complementary information for the detection task. The objective of our work is to leverage this complementary information to improve detection. Therefore, we propose a spatio-temporal fusion framework (STF). We first introduce multi-frame and single-frame attention modules that allow a neural network to share feature maps between nearby frames to obtain more robust object representations. Second, we introduce a dual-frame fusion module that merges feature maps in a learnable manner to improve them. Our evaluation is conducted on three different benchmarks including video sequences of moving road users. The performed experiments demonstrate that the proposed spatio-temporal fusion module leads to improved detection performance compared to baseline object detectors. Code is available at `https://github.com/noreenanwar/STF-module`. ↗

Leveraging Prompt-Based Segmentation Models and Large Dataset to Improve Detection of Trees

Vincent Grondin (Université Laval), Philippe Massicotte (Hydro-Québec), Mohamed Gaha (Hydro-Québec), François Pomerleau (Université Laval), Philippe Giguère (Université Laval)

The abundance of unlabeled forest images on the web is a powerful yet untapped resource to train forestry vision models. Two key challenges limiting the use of these unlabeled images are i) collecting the images and ii) obtaining the labels, as supervised learning remains the prevailing approach for model training. In this work, we address the first issue by providing a dataset of 110 k forest images sourced from a repository of pictures taken by amateur photographers worldwide. To generate supplementary labels for supervised training, we propose a two-step approach. First, we train a network on a small labelled dataset, to generate pseudo-labels on the much larger, unlabeled one. Then, we leverage the zero-shot segmentation capability of the Segment Anything Model to improve the quality of these pseudo-labels. Our experiments demonstrate that both the proposed dataset and the pseudo-labeling method increase performance of a tree detector at no additional labeling cost. This performance increase is particularly significant in challenging scenarios, showing that training the model with better segmentation masks notably helps disentangle overlapping trees and detect odd-shaped ones, gaining between 3.3 APbb, 7.7 APseg or 1.6 APbb, 3.5 APseg percentage points depending on the burn-in model. Code and dataset links are available at `https://github.com/norlab-ulaval/PercepTreeV1`. ↗

Oral Session 6 — Day 3

Cross-Graph Domain Adaptation for Skeleton-based Human Action Recognition

Haitao Tian (University of Ottawa), James Dickens (University of Ottawa), and Pierre Payeur (University of Ottawa)

Recent research on human action recognition is largely facilitated by skeletal data, a compact graph representation composed of key joints of the human skeleton that is efficiently extracted by body tracking systems and that offers the merit of being robust to environmental variations. However, the skeleton resolution and joint connectivity of the extracted skeletons may vary with sensor devices, which results in different skeleton graph representations on collected data. This paper investigates a cross skeleton graph domain adaptation approach where a skeleton action recognition model is trained upon a source skeletal data domain but is expected to adapt onto a target domain configured with a different skeleton graph. It proposes an adversarial learning framework where a generation space is developed on which the model learns valid skeletal action knowledge from the source graph domain. Interaction with an embedded discrimination space is employed to extract heterogenous graph features from the target domain. Optimization of the generation space and the discrimination space is realized alternatively under adversarial learning which guarantees action-aware and domain-agnostic skeletal knowledge, thus forming a joint human action recognition model effectively functioning on both graph domains. In experiments, the paper evaluates the proposed method by incorporating graph convolutional networks into two skeleton action recognition benchmarks, NTU-RGB+D and Northwestern-UCLA, where comparisons are conducted to demonstrate the effectiveness of the proposed approach. Code will be available at `https://github.com/tht106/CrossGraphDA`.

AugTrEP: Scene and Occlusion-Aware Pedestrian Crossing Intention Prediction

Aditya Bhattacharjee (University of Toronto Institute for Aerospace Studies) and Steven L. Waslander (University of Toronto Institute for Aerospace Studies)

Accurately predicting the crossing behaviour of pedestrians remains a significant challenge due to their complex behavioural dynamics. Although modern transformer-based models have shown promise in being able to accurately capture these dynamics, the crucial role of contextual information, especially under occluded scenarios, has been underexplored. In this work, we demonstrate that additional contextual features, such as crosswalk visibility and traffic light status, can assist in improving prediction performance under degraded conditions, where accurate pedestrian information is not readily available. We propose AugTrEP, inspired by the existing Transformer-based Evidential Prediction (TrEP) network, which uses two transformer encoders with cross-attention to learn pedestrian behaviour by incorporating global traffic context. We evaluate our models against the PIE benchmark and curated test sets simulating the behaviour of real-world perception systems under varying degrees of occlusion. Our analysis reveals a significant improvement in the accuracy, AUC, F1 score, and precision compared to the baseline under degraded input conditions. These findings highlight AugTrEP's resiliency to disturbances caused by occlusions and emphasize the importance of scene context in accurate behaviour prediction for real-world applicability.

6 Poster Session

Robotics

Feature Density Estimation for Out-of-Distribution Detection via Normalizing Flows

Evan D. Cook (University of Toronto Institute for Aerospace Studies), Marc-Antoine Lavoie (University of Toronto Institute for Aerospace Studies), and Steven L. Waslander (University of Toronto Institute for Aerospace Studies)

Out-of-distribution (OOD) detection is a critical task for safe deployment of learning systems in the open world setting. In this work, we investigate the use of feature density estimation via normalizing flows for OOD detection and present a fully unsupervised approach which requires no exposure to OOD data, avoiding researcher bias in OOD sample selection. This is a post-hoc method which can be applied to any pretrained model, and involves training a lightweight auxiliary normalizing flow model to perform the out-of-distribution detection via density thresholding. Experiments on OOD detection in image classification show strong results for far-OOD data detection with only a single epoch of flow training, including 98.2% AUROC for ImageNet-1k vs. Textures, which exceeds the state of the art by 7.8%. We additionally explore the connection between the feature space distribution of the pretrained model and the performance of our method. Finally, we provide insights into training pitfalls that have plagued normalizing flows for use in OOD detection. ↗

DTM: Difference-Based Temporal Module for Monocular Category-Level 6 DoF Object Pose Tracking

Zishen Chen (University of Ottawa) and Jochen Lang (University of Ottawa)

We propose DTM, a novel difference-based temporal module to leverage historical information in category-level 6DoF pose tracking tasks. It can be easily integrated with various category-level 6DoF pose tracking models which use RGBD data as input. This module extracts temporal features through a KNN-based difference calculation strategy from both, pixels and 3D points. We evaluate this module on two pose estimation datasets, NOCS-REAL275 and MoVi-E by integrating our module with two state-of-the-art 6D pose tracking models. The result shows that DTM can significantly increase the accuracy and robustness of category-level 6DoF trackers. ↗

LaserSAM: Zero-Shot Change Detection Using Visual Segmentation of Spinning LiDAR

Alexander Krawciw (University of Toronto Robotics Institute), Sven Lilge (University of Toronto Robotics Institute), and Timothy D. Barfoot (University of Toronto Robotics Institute)

This paper presents an approach for applying camera perception techniques to spinning LiDAR data. To improve the robustness of long-term change detection from a 3D LiDAR, range and intensity information are rendered into virtual perspectives using a pinhole camera model. Hue-saturation-value image encoding is used to colourize the images by range and near-IR intensity. The LiDAR's active scene illumination makes it invariant to ambient brightness, which enables night-to-day change detection without additional processing. Using the range-colourized, perspective image allows existing foundation models to detect semantic regions. Specifically, the Segment Anything Model detects semantically similar regions in both a previously acquired map and live view from a path-repeating robot. By comparing the masks in both views, changes in the live scan are detected. Results indicate that the Segment Anything Model accurately captures the shape of arbitrary changes introduced into scenes. The proposed method achieves a segmentation intersection over union of 73.3% when evaluated in unstructured environments and 80.4% when evaluated within the planning corridor. Changes can be detected reliably through day-to-night illumination variations. After pixel-level masks are generated, the one-to-one correspondence with 3D points means that the 2D masks can be used directly to recover the 3D location of the changes. The detected 3D changes are avoided in a closed loop by treating them as obstacles in a local motion planner. Experiments on an unmanned ground vehicle demonstrate the performance of the method.

Critical Infrastructure Asset Imaging Pipeline

Jonathan Dupuis (Carleton University) and James R. Green (Carleton University)

The ability to retrieve and analyze recent images of critical infrastructure assets is beneficial for regular monitoring, post-disaster assessment, or preparing for a service call. Given a high-quality image of an asset, several recently developed deep learning models can automatically assess the state of the infrastructure. However, obtaining such an image automatically remains an open question. Enterprise imaging initiatives, such as Google Street View, permit the viewing of road-adjacent images, given geographic coordinates. The spatial resolution of such systems is excellent, although the temporal resolution varies from months to years. We have recently forecast the emergence of on-demand imaging using instrumented vehicles that would permit more recent or frequent imaging of locations of interest. However, the challenge remains to retrieve a high-quality image of an asset of interest, free from obstructions and imaging artifacts. We here propose a pipeline to retrieve recent images of an asset given an imaging source, GPS coordinates, and an asset class. Object detection is used to automatically identify the asset of interest and to detect obstructions or imaging artifacts. If necessary, additional images are requested for surrounding locations to provide multiple views of the asset of interest culminating in an image free from artifacts. The pipeline is demonstrated using two critical infrastructure asset classes (utility poles and street lights) and two image sources (Streetview and a repository of dashcam video). Robust performance is observed, resulting in correct asset identification and imaging in 76.5% of cases (up from 54.5%), while requiring an average of 1.47 images per asset to achieve a high-quality image free from obstructions and artifacts. The proposed pipeline will be of interest to disaster response teams, utilities, and other critical infrastructure asset managers.

Image-to-Joint Inverse Kinematic of a Supportive Continuum Arm Using Deep Learning

Shayan Sepahvand (Toronto Metropolitan University), Guanghui Wang (Toronto Metropolitan University), and Farrokh Janabi-Sharifi (Toronto Metropolitan University)

In this work, a deep learning-based technique is used to study the image-to-joint inverse kinematics of a tendon-driven supportive continuum arm. An eye-off-hand configuration is considered by mounting a camera at a fixed pose with respect to the inertial frame attached at the arm base. This camera captures an image for each distinct joint variable at each sampling time to construct the training dataset. This dataset is then employed to adapt a feed-forward deep convolutional neural network, namely the modified VGG-16 model, to estimate the joint variable. One thousand images are recorded to train the deep network, and transfer learning and fine-tuning techniques are applied to the modified VGG-16 to further improve the training. Finally, training is also completed with a larger dataset of images that are affected by various types of noises, changes in illumination, and partial occlusion. The main contribution of this research is the development of an image-to-joint network that can estimate the joint variable given an image of the arm, even if the image is not captured in an ideal condition. The key benefits of this research are twofold: 1) image-to-joint mapping can offer a real-time alternative to computationally complex inverse kinematic mapping through analytical models; and 2) the proposed technique can provide robustness against noise, occlusion, and changes in illumination. The dataset is publicly available on Kaggle.

Associating Landmarks from SLAM's Visual Structure

Matthew Bradley (University of Waterloo) and John S. Zelek (University of Waterloo)

Place recognition is the online task of detecting revisits to previously seen locations and is a key to many navigational systems. In Simultaneous Localization and Mapping, recovering the relative camera pose between recognized visit and revisit (e.g. using bundle adjustment) allows for global map optimization, improving localization accuracy. Visual SLAM recovers structure to estimate camera movement but it is typically not used for visual place recognition. Limited past work which adapted LiDAR place recognition descriptors to SLAM-recovered physical structure found superior robustness to visual effects vs appearance-based VPR, but overall had poorer recall. It was found that LiDAR descriptors' whole-scan matching assumes excellent 360 degree pointcloud coverage while cameras have limited FoV. We observe that SLAM-tracked points congregate on objects and distinct elements, resulting in sparsity that impacts whole-scan matching. To us this also suggests use of clustering to extract these aggregate congregations as landmarks whose configuration can be matched. Exploring this approach we found that the landmarks generated still vary in detected position, but a far more significant hurdle is that the same landmarks may not be repeatedly clustered each time a scene is visited. This is due to large-scale clustering still being sensitive to instability in the individual SLAM points. This was improved significantly but not sufficiently through visual semantic labeling of the initial 3D points, helping to provide more stable, guided clustering solutions. Still, single missing or "outlier" landmarks are detrimental to successful association between landmark sets. To address this instability in future work we recommend careful selection of salient points from those collected by SLAM, for those which can be expected to be the most stable and repeatably detected. This is expected to provide more stable landmarks than large-scale clustering of detected points which relies on a center-of-mass approach.

Vision

SealD-NeRF: Interactive Pixel-Level Editing for Dynamic Scenes by Neural Radiance Fields

Zhentao Huang, Yukun Shi (University of Guelph), Neil Bruce (University of Guelph), and Minglun Gong (University of Guelph)

The widespread adoption of implicit neural representations, especially Neural Radiance Fields (NeRF) as detailed by [1], highlights a growing need for editing capabilities in implicit 3D models, essential for tasks like scene post-processing and 3D content creation. Despite previous efforts in NeRF editing, challenges remain due to limitations in editing flexibility and quality. The key issue is developing a neural representation that supports local edits for real-time updates. Current NeRF editing methods, offering pixel-level adjustments or detailed geometry and color modifications, are mostly limited to static scenes. This paper introduces SealD-NeRF, an extension of Seal-3D for pixel-level editing in dynamic settings, specifically targeting the D-NeRF network [2]. It allows for consistent edits across sequences by mapping editing actions to a specific time frame, freezing the deformation network responsible for dynamic scene representation, and using a teacher-student approach to integrate changes. The code and the supplementary video link are available at `https://github.com/ZhentaoHuang/SealD-NeRF`.

SLVVA: Scalable Land Viability via Vision-Language Architecture

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

Digital soil mapping is a process of creating maps of soil properties and their spatial distribution. It plays a vital role in monitoring soil health and promoting sustainable and efficient land use. In the past, environmental data was used to guide the creation of soil property maps. However, the failure to consider the accessibility of locations has led to a bias in the mapping process. In our research, we utilize satellite imagery to evaluate location accessibility, leading to more balanced soil property mapping. We formulate land viability detection and introduce a scalable two-step framework for its detection. Initially, we classify land viability, followed by its segmentation. We leverage Convolutional Neural Networks (CNNs) for classification and a resilient and generalizable vision-language architecture for segmentation. Our most notable results stem from fine-tuning a pre-existing VGGNet for classification and employing a CLIP-based Segmentation method (CLIPSeg) for segmentation. We demonstrate the effectiveness of our approach through extensive experimentation on EuroSAT and OpenEarthMap datasets. Our work is the first to address the challenge of biased sampling in digital soil mapping by incorporating satellite images to assess the accessibility of locations, ensuring a more representative soil property mapping.

Trini: An Efficient Representation of Dynamic Scenes for Sparse-View Camera Settings

Rishav Bhardwaj (University of Waterloo), John S. Zelek (University of Waterloo), and Vasudevan Lakshminarayanan (University of Waterloo)

3D reconstruction of a dynamic scene has been a challenging task in vision. Several strategies have been developed to enhance the reconstruction of dynamic scenes, with some employing tri-projection decomposition techniques that surpass D-NeRF in terms of speed and effectiveness. This paper introduces Trini, which decomposes a dynamic 3D scene into three volumes dealing with the 3D coordinates influenced by time. Each volume is further structured with four marginalized planes. These planes are then integrated with a compact MLP for rendering superior results in a seamless manner. Additionally, we incorporate a technique to efficiently determine coordinates in a set of distinct images for enhancing the reconstruction process for cases involving sparse-view camera images. The efficacy of our method outperforms other state-of-the-art techniques and is particularly evident in capturing the dynamic elements and edges present in the scene.

QWID: Quantized Weed Identification Deep Neural Network

Parikshit Singh Rathore (Maharana Pratap University of Agriculture and Technology)

In this paper, we present an efficient solution for weed classification in agriculture. We focus on optimizing model performance at inference while respecting the constraints of the agricultural domain. We propose a Quantized Deep Neural Network model that classifies a dataset of 9 weed classes using 8-bit integer (int8) quantization, a departure from standard 32-bit floating point (fp32) models. Recognizing the hardware resource limitations in agriculture, our model balances model size, inference time, and accuracy, aligning with practical requirements. We evaluate the approach on ResNet-50 and InceptionV3 architectures, comparing their performance against their int8 quantized versions. Transfer learning and fine-tuning are applied using the DeepWeeds dataset. The results show staggering model size and inference time reductions while maintaining accuracy in real-world production scenarios like Desktop, Mobile and Raspberry Pi. Our work sheds light on a promising direction for efficient AI in agriculture, holding potential for broader applications.

POPCat: Propagation of Particles for Complex Annotation Tasks

Adam Srebrnjak Yang (University of Waterloo), Dheeraj Khanna (University of Waterloo), and John S. Zelek (University of Waterloo)

Novel dataset creation for all multi-object tracking, crowd-counting, and industrial-based videos is arduous and time-consuming when faced with a unique class that densely populates a video sequence. We propose a time efficient method called POPCat that exploits the multi-target and temporal features of video data to produce a semi-supervised pipeline for segmentation or box-based video annotation. The method retains the accuracy level associated with human level annotation while generating a large volume of semi-supervised annotations for greater generalization. The method capitalizes on temporal features through the use of a particle tracker to expand the domain of human-provided target points. This is done through the use of a particle tracker to reassociate the initial points to a set of images that follow the labeled frame. A YOLO model is then trained with this generated data, and then rapidly infers on the target video. Evaluations are conducted on GMOT-40, AnimalTrack, and Visdrone-2019 benchmarks. These multi-target video tracking/detection sets contain multiple similar-looking targets, camera movements, and other features that would commonly be seen in "wild" situations. We specifically choose these difficult datasets to demonstrate the efficacy of the pipeline and for comparison purposes. The method applied on GMOT-40, AnimalTrack, and Visdrone shows a margin of improvement on recall/mAP50/mAP over the best results by a value of 24.5%/9.6%/4.8%, -/43.1%/27.8%, and 7.5%/9.4%/7.5% where metrics were collected.

SatStreaks: Towards Supervised Learning for Delineating Satellite Streaks from Astronomical Images

Susrita Chatterjee (Saint Mary's University), Prachi Kudeshia (Saint Mary's University), Nikolaus Kollo (Saint Mary's University), Muhammad Altaf Agowun (Saint Mary's University), Jiju Peethambaran (Saint Mary's University), and Yasushi Akiyama (Saint Mary's University)

Delineation of satellite streaks in astronomical images is an important aspect of ground based space studies. While deep learning algorithms show promise, training and validation of deep learning models for satellite streak segmentation is challenging due to the limited availability of large-scale, annotated datasets. We introduce SatStreaks, a dataset comprising of 3,130 densely annotated, real images of satellite streaks captured through ongoing citizen science projects. We utilize SatStreaks to develop a U-Net based model for the streak segmentation and conduct an experimental evaluation of data-driven image segmentation algorithms. The satellite streak segmentation codebase consisting of various deep learning models, and the SatStreak dataset has been made publicly available (`https://github.com/jijup/SatStreaks`) to facilitate the advancement of computer vision algorithms for space studies.

Dense Monocular Motion Segmentation Using Optical Flow and Pseudo Depth Map: A Zero-Shot Approach

Yuxiang Huang (University of Waterloo), Yuhao Chen (University of Waterloo), and John S. Zelek (University of Waterloo)

Motion segmentation is the task of detecting and segmenting moving objects from a moving monocular camera. It is a fundamental but challenging problem in computer vision due to unknown camera motion, unknown scene structure and complex object appearance. Deep learning methods currently achieve the best results in dense monocular motion segmentation, but both supervised and unsupervised methods require training on massive datasets, and supervised methods also require a significant amount of annotation. In contrast, traditional methods, which usually rely on optical flow as the only motion cue, do not need training or supervision, but they fail to capture object-level information, leading to over-segmentation or undersegmentation. In addition, they also struggle in complex scenes with substantial depth variations and non-rigid motion, due to the overreliance of optical flow. To overcome these limitations, we propose a novel approach that combines the advantages of both deep learning methods and tradition methods to perform dense motion segmentation without requiring any training. Our approach first generates an object proposal for each video frame using computer vision foundation models, then clusters these proposed objects into different motion groups by using both optical flow and relative depth map as motion cues. the depth map is obtained from an off-the-shelf monocular depth estimation model – it compliments the optical flow to generate a more robust motion cue against motion parallax. Experiments show that our method achieves promising results on the DAVIS-Moving and YTVOS-Moving datasets, outperforming the best unsupervised method and closely matches with the state-of-the-art supervised methods.

Multi Player Tracking in Ice Hockey with Homographic Projections

Harish Prakash (University of Waterloo), Jia Cheng Shang(University of Waterloo), Ken M. Nsiempba(University of Waterloo), Yuhao Chen(University of Waterloo), David A. Clausi(University of Waterloo), and John S. Zelek(University of Waterloo)

Multi Object Tracking (MOT) in ice hockey pursues the combined task of localizing and associating players across a given sequence to maintain their identities. Tracking players from monocular broadcast feeds is an important computer vision problem offering various downstream analytics and enhanced viewership experience. However, existing trackers encounter significant difficulties in dealing with occlusions, blurs, and agile player movements prevalent in telecast feeds. In this work, we propose a novel tracking approach by formulating MOT as a bipartite graph matching problem infused with homography. We disentangle the positional representations of occluded and overlapping players in broadcast view, by mapping their foot keypoints to an overhead rink template, and encode these projected positions into the graph network. This ensures reliable spatial context for consistent player tracking and unfragmented tracklet prediction. Our results show considerable improvements in both the *IDsw* and *IDF1* metrics on the two available broadcast ice hockey datasets. ↗

7 Workshop Speakers

In alphabetical order

Francisco Javier Andrade Chavez

University of Waterloo

Talk Title: Human to robot skill transfer: Using Bio-inspired force distribution in double support for humanoid locomotion

Abstract: In this talk we will discuss how to transfer some aspects of human locomotion into a humanoid robot. Humanoids are made to resemble humans. One of the advantages of such a form factor is the potential to transfer human based skills into the humanoid robot platform. In locomotion, the likelihood of slipping or maintaining contact is determined by the forces applied on the environment. Therefore, it is crucial to find methods for maintaining forces within friction constraints. In single support, the relationship between center of mass acceleration and forces is unique. However, in double support, it becomes a non-deterministic problem. It is often assumed that forces are distributed to minimize a certain effort criterion. An interesting alternative is to distribute the forces in a manner similar to how a human would, which could result in a more human-like gait for humanoid robots.

Biography: Francisco Javier Andrade Chavez has been Lab Manager and Postdoc for the Human-Centred Robotics and Machine Intelligence lab at the Department of Systems and Design at the University of Waterloo since July 2020. He has also been Humanoid Specialist for the Robohub at the Facutly of Engineering of the University of Waterloo since November 2023. Francisco obtained his doctorate in Bioengineering and Robotics from the Universita degli Studi di Genova in collaboration with the Istituto Italiano di Tecnologia (IIT) in 2019. He then stayed as Postdoc and Scrum Master at the Dynamic Interactiona and Control group, now known as Artificial and Machine Intelligence, leading the telexistence research team. His research interest lies in endowing robots with the ability to exploit robot dynamics to adapt in rapidly changing scenarios while seamlessly interacting with humans. This has led to research in the areas of telexistence, balancing, loco-manipulation control, socio-physical human-robot interaction, wearable sensors, human-robot skill transfer and estimation applied to humanoid robots.

Igor Gilitschenski

University of Toronto

Talk Title: Do Androids Dream of Electric Sheep? A Generative Paradigm for Dataset Design

Abstract: Traditional approaches for autonomy and AI robotics typically focus either on large scale data collection or on improving simulation. Although most practitioners rely on both approaches, they are largely still applied in separate workflows and seen as conceptually unrelated. In this talk, I will argue that this is a false dichotomy. Recent advances in generative models enable the unification of these seemingly different methodologies. Using real-world data for building data generation systems has led to numerous advances with impact in robotics and autonomy: going beyond pure distillation approaches, unifying creation and curation enables sophisticated automatic labeling pipelines and data-driven simulators. I will present some of our work following this paradigm and outline several basic research challenges and limitations associated with building systems that learn with generated data.

Biography: Igor Gilitschenski is an Assistant Professor of Computer Science at the University of Toronto where he leads the Toronto Intelligent Systems Lab. Previously, he was a (visiting) Research Scientist at the Toyota Research Institute. Dr. Gilitschenski was a Research Scientist at MIT's Computer Science and Artificial

Intelligence Lab and the Distributed Robotics Lab (DRL). There he was the technical lead of DRL's autonomous driving research team. He joined MIT from the Autonomous Systems Lab of ETH Zurich where he worked on robotic perception, particularly localization and mapping. He obtained his doctorate in Computer Science from the Karlsruhe Institute of Technology and a Diploma in Mathematics from the University of Stuttgart. His research interests involve developing novel robotic perception and decision-making methods for challenging dynamic environments. His work has received multiple awards including best paper awards at the American Control Conference, the International Conference of Information Fusion, and the Robotics and Automation Letters.