19th Conference on Robots and Vision

Toronto, Ontario

May 31 - June 2, 2022

Detailed Program

Contents

1 Program at a Glance

	CRV Session
	CRV & AI Joint Session
	Breaks & Social

		Conference Day 1	Conference Day 2	Conference Day 3
Time (PST)	**Time (EDT)**	**Tuesday 31-May**	**Wednesday 1-June**	**Thursday 2-June**
7:30	10:30	**Coffee**	**Coffee**	**Coffee**
7:45	10:45	**Conference Welcome Steve Waslander/Jonathan Kelly**	**Conference Welcome Steve Waslander/Jonathan Kelly**	**Conference Welcome Steve Waslander/Jonathan Kelly**
7:50	10:50	**Sponsored Talk Ryan Smith Oxbotica**	**Sponsored Talk Bruno Monsarrat NRC**	**Sponsored Talk Apeksha Kumavat Gatik**
8:10	11:10	**Keynote Speaker Lourdes Agapito**	**Keynote Speaker Maurice Fallon**	**Keynote Speaker Peter Corke**
8:30	11:30			
8:45	11:45			
9:00	12:00			
9:15	12:15	**Coffee**	**Coffee**	**Coffee**
9:30	12:30	**Oral Session 1: 3D Vision Chair: Siyu Tang**	**Oral Session 3: Learning-based Planning and Perception Chair: Mo Chen**	**Oral Session 5: Reasoning in 3D Chair: Mengye Ren**
9:45	12:45			
10:00	13:00			
10:15	13:15			
10:30	13:30	**Lunch - CIPPRS AGM**	**Lunch - AI/CRV Steering Committee Meeting**	**Lunch**
10:45	13:45			
11:00	14:00			
11:15	14:15			
11:30	14:30	**Oral Session 2: Planning and Control Chair: Mahdis Bisheban**	**Oral Session 4: Text Recognition Chair: Samira Ebrahimi Kahou**	**Invited Talks: Dissertation Award Chair: Glen Berseth**
11:45	14:45			
12:00	15:00			
12:15	15:15			
12:30	15:30	**Coffee**	**Coffee**	**Closing Remarks**
12:45	15:45	**Poster Session 1**	**Poster Session 2**	
13:00	16:00			
13:15	16:15			
13:30	16:30			
13:45	16:45	**Student Event**	**Sponsored Talk Keith Leung Trimble/Applanix**	
14:00	17:00			
14:15	17:15			
14:30	17:30		**Awards ceremony - online**	
14:45	17:45	**In-person Welcome Event Faculty Club**		
15:00	18:00			
15:15	18:15			
15:30	18:30			
15:45	18:45			
19:00	19:00			
19:15	19:15			
19:30	19:30			

2 Keynote Speakers

Lourdes Agapito

University College London

Talk Title

Learning to Reconstruct the 3D World from Images and Video

Abstract

As humans we take the ability to perceive the dynamic world around us in three dimensions for granted. From an early age we can grasp an object by adapting our fingers to its 3D shape; understand our mother's feelings by interpreting her facial expressions; or effortlessly navigate through a busy street. All these tasks require some internal 3D representation of shape, deformations, and motion. Building algorithms that can emulate this level of human 3D perception, using as input single images or video sequences taken with a consumer camera, has proved to be an extremely hard task. Machine learning solutions have faced the challenge of the scarcity of 3D annotations, encouraging important advances in weak and self-supervision. In this talk I will describe progress from early optimization-based solutions that captured sequence-specific 3D models with primitive representations of deformation, towards recent and more powerful 3D-aware neural representations that can learn the variation of shapes and textures across a category and be trained from 2D image supervision only. There has been very successful recent commercial uptake of this technology and I will show exciting applications to AI-driven video synthesis.

Biography

Lourdes Agapito holds the position of Professor of 3D Vision at the Department of Computer Science, University College London (UCL). Her research in computer vision has consistently focused on the inference of 3D information from single images or videos acquired from a single moving camera. She received her BSc, MSc and PhD degrees from the Universidad Complutense de Madrid (Spain). In 1997 she joined the Robotics Research Group at the University of Oxford as an EU Marie Curie Postdoctoral Fellow. In 2001 she was appointed as Lecturer at the Department of Computer Science at Queen Mary University of London. From 2008 to 2014 she held an ERC Starting Grant funded by the European Research Council to focus on theoretical and practical aspects of deformable 3D reconstruction from monocular sequences. In 2013 she joined the Department of Computer Science at University College London and was promoted to full professor in 2015. She now heads the Vision and Imaging Science Group, is a founding member of the AI centre and co-director of the Centre for Doctoral Training in Foundational AI. Lourdes serves regularly as Area Chair for the top Computer Vision conferences (CVPR, ICCV, ECCV) was Program Chair for CVPR 2016 and will serve again for ICCV 2023. She was keynote speaker at ICRA 2017 and ICLR 2021. In 2017 she co-founded Synthesia, the London based synthetic media startup responsible for the AI technology behind the Malaria no More video campaign that saw David Beckham speak 9 different languages to call on world leaders to take action to defeat Malaria.

Maurice Fallon

University of Oxford

Talk Title

Multi-Sensor Robot Navigation and Subterranean Exploration

Abstract

In this talk I will overview the work of my research group, Dynamic Robot Systems Group. I will focus on multi-sensor state estimation and 3D mapping to enable robots to navigate and explore dirty, dark and dusky environments - with an emphasis on underground exploration

with quadrupeds. This multitude of sensor signals need to be fused efficiently and in real-time to enable autonomy. Much of the work will be presented in the context of the DARPA SubT Challenge (Team Cerberus) and the THING EU project. I will also describe our work on trajectory optimization for dynamic motion planning and the use of learning to bootstrap replanning.

Biography

Maurice Fallon is an Associate Professor and Royal Society University Research Fellow at University of Oxford. His research is focused on probabilistic methods for localization and mapping. He has also made research contributions to state estimation for legged robots and is interested in dynamic motion planning and control. His PhD was from University of Cambridge in the field of sequential Monte Carlo methods. He worked as a PostDoc in Prof. John Leonard's Marine Robotics Group in MIT from 2008 before leading the perception part of MIT's entry in the DARPA Robotics Challenge. He has worked in domains as diverse as marine robots detecting mines, humanoid robotics and mapping radiation in nuclear facilities.

Peter Corke

Queensland University of Technology

Talk Title

Hand-eye coordination (and other things)

Abstract

Hand-eye coordination is an under-appreciated human super power. This talk will cover the robot equivalent, robot hand-camera coordination, where computer vision meets robotic manipulation. This robotic skill is needed wherever the robot's workpiece is not precisely located, or is moving, or the robot moving. The talk will motivate the problem, review recent progress in the field, and give an update on some new software tools for robotics research.

Biography

Peter Corke is a robotics researcher and educator. He is the distinguished professor of robotic vision at Queensland University of Technology, and former director of the ARC Centre of Excellence for Robotic Vision. He is a technical advisor to emesent and LYRO Robotics, and Chief Scientist of Dorabot. His research is concerned with enabling robots to see, and the application of robots to mining, agriculture and environmental monitoring. He created widely used open-source software for teaching and research, wrote the best selling textbook "Robotics, Vision, and Control", created several MOOCs and the Robot Academy, and has won national and international recognition for teaching including 2017 Australian University Teacher of the Year. He is a fellow of the IEEE, the Australian Academy of Technology and Engineering, the Australian Academy of Science; former editor-in-chief of the IEEE Robotics Automation magazine; founding editor of the Journal of Field Robotics; founding multimedia editor and executive editorial board member of the International Journal of Robotics Research; member of the editorial advisory board of the Springer Tracts on Advanced Robotics series; recipient of the Qantas/Rolls-Royce and Australian Engineering Excellence awards; and has held visiting positions at Oxford, University of Illinois, Carnegie-Mellon University and University of Pennsylvania. He received his undergraduate and masters degrees in electrical engineering and PhD from the University of Melbourne.

3 Symposium Speakers

Glen Berseth

Université de Montréal

Title

Developing Robots that Autonomously Learn and Plan in the Real World

Abstract

While humans plan and solve tasks with ease, simulated and robotics agents struggle to reproduce the same fidelity, robustness and skill. For example, humans can grow to perform incredible gymnastics, prove that black holes exist, and produce works of art, all starting from the same base learning system. I will present a collection of recent work that indicates we can train agents to learn these skills; however, they need to learn from a large amount of experience. To enable this learning, the agent needs to (1) be able to collect a large amount of experience, (2) train its model to best reuse this experience and (3) optimize general objectives for understanding and controlling its environment.

Biography

Glen Berseth is an assistant professor at the Université de Montréal, a core academic member of Mila, CIFAR AI chair, and co-director of the Robotics and Embodied AI Lab (REAL). He was a Postdoctoral Researcher with Berkeley Artificial Intelligence Research (BAIR) working in the Robotic AI Learning (RAIL) lab with Sergey Levine. He completed his NSERC-supported Ph.D. in Computer Science at the University of British Columbia in 2019, where he worked with Michiel van de Panne. He received his MSc from York University under the supervision of Petros Faloutsos in 2014 and worked at IBM (2012) and with Christopher Pal at ElementAI (2018). His goal is to create systems that can learn and act in the world intelligently by developing deep learning and reinforcement learning methods that solve diverse planning problems from vision.

Mahdis Bisheban

University of Calgary

Title

Control of UAVs in wind

Abstract

In this talk, first, I would like to present my research on the problem of estimation for rigid body dynamics. I will present a method to estimate unknown parameters of a rigid body dynamics model specifically on the Special Euclidian Group. Next, I would like to talk about a geometric adaptive controller for a quadrotor unmanned aerial vehicle with artificial neural networks. The dynamics of a quadrotor can be disturbed by the arbitrary, unstructured forces and moments caused by wind. I will show that if the control system is directly developed on the special Euclidean group and augmented with the multilayer neural networks, and the weights of the neural networks are adjusted online according to an adaptive law, we can mitigate the wind effects.

Biography

Dr. Mahdis Bisheban is an Assistant Professor in the Department of Mechanical and Manufacturing Engineering since July 2021. Previously, she served as a Research Associate at the Aerospace Research Center, National Research Council Canada (NRC), and as a Postdoctoral Fellow in the Department of Mechanical and Materials Engineering, Queen's University. Her current research focuses on two pillars: (i) Control, Estimation and Modelling of Aerial and Underwater Vehicles and (ii) Collaborative robots.

Mo Chen

Simon Fraser University

Title

Control, learning, and multi-agent RL

Abstract

Autonomous mobile robots are becoming pervasive in everyday life, and hybrid approaches that merge traditional control theory and modern data-driven methods are becoming increasingly important. In this talk, we first examine how value functions and control policies obtained from control theory can improve data efficiency and generalization of robotic learning. Then, we discuss recent developments in hierarchical multi-agent reinforcement learning.

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 completed his PhD in the Electrical Engineering and Computer Sciences Department at the University of California, Berkeley with Claire Tomlin in 2017, and received his BASc in Engineering Physics from the University of British Columbia in 2011. From 2017 to 2018, Mo was a postdoctoral researcher in the Aeronautics and Astronautics Department in Stanford University with Marco Pavone. His research interests include multi-agent systems, safety-critical systems, human-robot interactions, control theory, and reinforcement learning. Mo received the 2017 Eli Jury Award for his research and the 2016 Demetri Angelakos Memorial Achievement Award and his mentorship of students.

Samira Ebrahimi Kahou

ÉTS/Mila

Title

Learning Dynamical Representations

Abstract

Capturing aspects of dynamics plays an important role in many decision making tasks. Some of the main challenges in learning dynamics are: uncertainty in future, modeling long-term dependencies, partial observability and learning efficient representations. In this talk, I present three works on learning dynamics, for multi-agent trajectory prediction, learning robust representations in partially observable environments, and a sequential model for designing reward in automatic evaluation of crane operators.

Biography

Samira Ebrahimi Kahou is an Associate Professor at École de technologie supérieure/Vision and RL Lab. She is a Canada CIFAR AI Chair and member of Mila. Before joining ÉTS, she was a postdoctoral fellow working with Doina Precup at McGill/Mila. She received her Ph.D. from Polytechnique Montréal/Mila in 2016 under the supervision of Chris Pal. She also worked as a Researcher at Microsoft Research Montréal. Her research lab focuses on the intersection of computer vision and reinforcement learning with diverse applications.

Mengye Ren

Google Brain

Title

Visual Learning in the Open World

Abstract

Over the past decades, we have seen machine learning making great strides in understanding visual scenes. Yet, most of its success relies on training models on a massive amount of data offline in a closed world and evaluating them in a similar test environment. In this talk, I would like to envision an alternative paradigm that will allow machines to acquire visual knowledge through an online stream of data in an open world, which entails abilities such as learning

visual representations and concepts efficiently with limited and non-iid data. These capabilities will be core to future applications of real-world agents such as robotics and assistive technologies. I will share three recent papers towards this goal, and these works form three levels in our open-world visual recognition pipeline: the concept level, the grouping level, and the representation level. First, on the concept level, I will introduce a new learning paradigm that rapidly learns new concepts in a continual stream with only a few labels. Second, on the grouping level, I will discuss how to learn both representations and concept classes online without any labeled data by grouping similar objects into clusters. Lastly, on the representation level, I will present a new algorithm that learns general visual representations from high resolution raw video. With these different levels combined, I am hopeful that future intelligent agents will be able to learn on-the-fly without manually collecting data beforehand.

Biography

Mengye Ren is a visiting researcher at Google Brain and an incoming assistant professor at New York University. Previously, he obtained his PhD from the University of Toronto. From 2017 to 2021 he was also a research scientist at Uber ATG and Waabi working on self-driving cars. His research focuses on enabling machines to continually learn, adapt, and reason in naturalistic environments, and he has done a series of work on combining few-shot, semi/unsupervised, and continual learning algorithms. He has won several awards including twice the NVIDIA research pioneer award and the Alexander Graham Bell Canada Graduate Fellowship.

Siyu Tang

ETH Zurich

Title

Inhabiting a virtual city

Abstract

In recent years, many high-quality datasets of 3D indoor scenes have emerged, such as Replica and Gibson, which employ 3D scanning and reconstruction technologies to create digital 3D environments. Also, virtual robotic agents exist inside 3D environments, such as the Habitat simulator. These are used to develop scene understanding methods from embodied views, thus providing platforms for indoor robot navigation, AR/VR, and many other applications. Despite this progress, a significant limitation of these environments is that they do not contain people. The reason such worlds contain no people is that there are no fully automated tools to synthesize realistic people interacting with 3D scenes naturally, and manually doing this requires significant artist effort. In this talk, I will present our previous and ongoing research about the capture and synthesis of realistic people interacting realistically with 3D scenes and objects.

Biography

Siyu Tang is an assistant professor at ETH Zürich in the Department of Computer Science since January 2020. She received an early career research grant to start her research group at the Max Planck Institute for Intelligent Systems in November 2017. She was a postdoctoral researcher in the same institute, advised by Dr. Michael Black. She finished her PhD at the Max Planck Institute for Informatics and Saarland University in 2017, under the supervision of Professor Bernt Schiele. Before that, she received her Master's degree in Media Informatics at RWTH Aachen University, advised by Prof. Bastian Leibe, and her Bachelor's degree in Computer Science at Zhejiang University, China. She has received several awards for her research, including the Best Paper Award at BMVC 2012 and 3DV 2020, Best Paper Award Candidates at CVPR 2021, an ELLIS PhD Award, and a DAGM-MVTec Dissertation Award.

4 Oral Session 1: 3D Vision

23 A Simple Method to Boost Human Pose Estimation Accuracy by Correcting the Joint Regressor for the Human3.6m Dataset

Eric W Hedlin (University of British Columbia); Helge Rhodin (University of British Columbia); Kwang Moo Yi (University of British Columbia)*

Abstract: Many human pose estimation methods estimate Skinned Multi-Person Linear (SMPL) models and regress the human joints from these SMPL estimates. In this work, we show that the most widely used SMPL-to-joint linear layer (joint regressor) is inaccurate, which may mislead pose evaluation results. To achieve a more accurate joint regressor, we propose a method to create pseudo-ground-truth SMPL poses, which can then be used to train an improved regressor. Specifically, we optimize SMPL estimates coming from a state-of-the-art method so that its projection matches the silhouettes of humans in the scene, as well as the ground-truth 2D joint locations. While the quality of this pseudo-ground-truth is challenging to assess due to the lack of actual ground-truth SMPL, with the Human 3.6m dataset, we qualitatively show that our joint locations are more accurate and that our regressor leads to improved pose estimations results on the test set without any need for retraining. We release our code and joint regressor at

30 Instance Segmentation of Herring and Salmon Schools in Acoustic Echograms using a Hybrid U-Net

Alex L Slonimer (University of Victoria); Melissa Cote (University of Victoria); Tunai Porto Marques (University of Victoria); Alireza Rezvanifar (University of Victoria); Stan Dosso (University of Victoria); Alexandra Branzan Albu (University of Victoria); Kaan Ersahin (ASL Environmental Sciences); Todd Mudge (ASL Environmental Sciences); Stephane Gauthier (Fisheries and Oceans Canada)*

Abstract: The automated classification of fish, such as herring and salmon, in multi-frequency echograms is important for ecosystems monitoring. This paper implements a novel approach to instance segmentation: a hybrid of deep-learning and heuristic methods. This approach implements semantic segmentation by a U-Net to detect fish, which are converted to instances of fish-schools derived from candidate components within a defined linking distance. In addition to four frequency channels of echogram data (67.5, 125, 200, 455 kHz), two simulated channels (water depth and solar elevation angle) are included to encode spatial and temporal information, which leads to substantial improvement in model performance. The model is shown to out-perform recent experiments that have used a Mask R-CNN architecture. This approach demonstrates the ability to classify sparsely distributed objects in a way that is not possible with state-of-the-art instance segmentation methods.

33 Supervised Contrastive Learning for Detecting Anomalous Driving Behaviours from Multimodal Videos

*Shehroz Khan (University Health Network); Ziting Shen (University of Toronto); Haoying Sun (University of Toronto); Ax Patel (University of Toronto); Ali Abedi (University Health Netowrk)**

Abstract: Distracted driving is one of the major reasons for vehicle accidents. Therefore, detecting distracted driving behaviours is of paramount importance to reduce the millions of deaths and injuries occurring worldwide. Distracted or anomalous driving behaviours are deviations from 'normal' driving that need to be identified correctly to alert the driver. However, these driving behaviours do not comprise one specific type of driving style and their distribution can be different during the training and test phases of a classifier. We formulate this problem as a supervised contrastive learning approach to learn a visual representation to detect normal, and seen and unseen anomalous driving behaviours. We made a change to the standard contrastive loss function to adjust the similarity of negative pairs to aid the optimization. Normally, in a (self) supervised contrastive framework, the projection head layers are omitted during the test phase as the encoding layers are considered to contain general

visual representative information. However, we assert that for a video-based supervised contrastive learning task, including a projection head can be beneficial. We showed our results on a driver anomaly detection dataset that contains 783 minutes of video recordings of normal and anomalous driving behaviours of 31 drivers from various top and front cameras (both depth and infrared). We also performed an extra step of fine tuning the labels in this dataset. Out of 9 video modalities combinations, our proposed contrastive approach improved the ROC AUC on 6 in comparison to the baseline models (from 4.23% to 8.91% for different modalities). We performed statistical tests that showed evidence that our proposed method performs better than the baseline contrastive learning setup. Finally, the results showed that the fusion of depth and infrared modalities from top and front view achieved the best AUC ROC of 0.9738 and AUC PR of 0.9772.

5 Oral Session 2: Planning and Control

20 ROS-X-Habitat: Bridging the ROS Ecosystem with Embodied AI

Guanxiong Chen (The University of British Columbia); Haoyu Yang (The University of British Columbia); Ian Mitchell (The University of British Columbia)*

Abstract: We introduce ROS-X-Habitat, a software interface that bridges the AI Habitat platform for embodied learning-based agents with other robotics resources via ROS. This interface not only offers standardized communication protocols between embodied agents and simulators, but also enables physically and photorealistic simulation that benefits the testing of vision-based embodied agents. With this interface, roboticists are able to train their own Habitat RL agents in another simulation environment or to develop their own robotic algorithms inside Habitat Sim v2. Through in silico experiments, we demonstrate that ROS-X-Habitat has minimal impact on the navigation performance and simulation speed of a Habitat RGBD agent; that a standard set of ROS mapping, planning and navigation tools can run in Habitat Sim v2; and that a Habitat agent can run in the standard ROS simulator Gazebo.

50 Temporal Convolutions for Multi-Step Quadrotor Motion Prediction

Steven L Waslander (University of Toronto); Samuel Y Looper (ETH Zurich)*

Abstract: Model-based control methods for robotic systems such as quadrotors, autonomous driving vehicles and flexible manipulators require motion models that generate accurate predictions of complex nonlinear system dynamics over long periods of time. Temporal Convolutional Networks (TCNs) can be adapted to this challenge by formulating multi-step prediction as a sequence-to-sequence modeling problem. We present End2End-TCN: a fully convolutional architecture that integrates future control inputs to compute multi-step motion predictions in one forward pass. We demonstrate the approach with a thorough analysis of TCN performance for the quadrotor modeling task, which includes an investigation of scaling effects and ablation studies. Ultimately, End2End-TCN provides 55% error reduction over the state of the art in multi-step prediction on an aggressive indoor quadrotor flight dataset. The model yields accurate predictions across 90 timestep horizons over a 900 ms interval.

6 Oral Session 3: Learning-based Planning and Perception

24 Inter- & Intra-City Image Geolocalization

Joshua Tanner (Carleton University); Kevin Dick (Carleton University); James R. Green (Carleton University)*

Abstract: Can a photo be accurately geolocated within a city from its pixels alone? While this image geolocation problem has been successfully addressed at the planetary- and nation-levels when framed as a classification problem using convolutional neural networks, no method has yet been able to precisely geolocate images within the city- and/or at the street-level when framed as a latitude/longitude regression-type problem. We leverage the highly densely sampled Streetlearn dataset of imagery from Manhattan and Pittsburgh to first develop a highly accurate inter-city predictor and then experimentally resolve, for the first time, the intra-city

performance limits of framing image geolocation as a regression-type problem. We then reformulate the problem as an extreme-resolution classification task by subdividing the city into hundreds of equirectangular-scaled bins and train our respective intra-city deep convolutional neural network on tens of thousands of images. Our experiments serve as a foundation to develop a scalable inter- and intra-city image geolocation framework that, on average, resolves an image within 250 m2. We demonstrate that our models outperform SIFT-based image retrieval-type models based on differing weather patterns, lighting conditions, location-specific imagery, and are temporally robust when evaluated upon both past and future imagery. Both the practical and ethical ramifications of such a model are also discussed given the threat to individual privacy in a technocentric surveillance capitalist society.

27 A Permutation Model for the Self-Supervised Stereo Matching Problem

Pierre-André Brousseau (Université de Montréal); Sebastien Roy (Universite de Montreal)*

Abstract: This paper proposes a novel permutation formulation to the stereo matching problem. Our proposed approach introduces a permutation volume which provides a natural representation of stereo constraints and disentangles stereo matching from monocular disparity estimation. It also has the benefit of simultaneously computing disparity and a confidence measure which provides explainability and a simple confidence heuristic for occlusions. In the context of self-supervised learning, the stereo performance is validated for standard testing datasets and the confidence maps are validated through stereo-visibility. Results show that the permutation volume increase stereo performance and features good generalization behaviour. We believe that measuring confidence is a key part of explainability which is instrumental to adoption of deep methods in critical stereo applications such as autonomous navigation.

31 Occlusion-Aware Self-Supervised Stereo Matching with Confidence Guided Raw Disparity Fusion

Xiule Fan (University of Waterloo); Soo Jeon (University of Waterloo); Baris Fidan (University of Waterloo)*

Abstract: Commercially available stereo cameras used in robots and other intelligent systems to obtain depth information typically rely on traditional stereo matching algorithms. Although their raw (predicted) disparity maps contain incorrect estimates, these algorithms can still provide useful prior information towards more accurate prediction. We propose a pipeline to incorporate this prior information to produce more accurate disparity maps. The proposed pipeline includes a confidence generation component to identify raw disparity inaccuracies as well as a self-supervised deep neural network (DNN) to predict disparity and compute the corresponding occlusion masks. The proposed DNN consists of a feature extraction module, a confidence guided raw disparity fusion module to generate an initial disparity map, and a hierarchical occlusion-aware disparity refinement module to compute the final estimates. Experimental results on public datasets verify that the proposed pipeline has competitive accuracy with real-time processing rate. We also test the pipeline with images captured by commercial stereo cameras to show its effectiveness in improving their raw disparity estimates.

7 Oral Session 4: Text Recognition

39 Occluded Text detection and Recognition in the Wild

Zobeir Raisi (University of Waterloo); John Zelek (University of Waterloo)*

Abstract: The performance of existing deep-learning scene text recognition-based methods fails significantly on occluded text instances or even partially occluded characters in a text due to their reliance on the visibility of the target characters in images. This failure is often due to features generated by the current architectures with limited robustness to occlusion, which opens the possibility of improving the feature extractors and/or the learning models to better handle these severe occlusions. In this paper, we first evaluate the performance of the current scene text detection, scene text recognition, and scene text spotting models

using two publicly-available occlusion datasets: Occlusion Scene Text (OST), that is designed explicitly for scene text recognition, and we also prepare an Occluded Character-level using the Total-Text (OCTT) dataset for evaluating the scene text spotting and detection models. Then we utilize a very recent Transformer-based framework in deep learning, namely Masked Auto Encoder (MAE), as a backbone for scene text detection and recognition pipelines to mitigate the occlusion problem. The performance of our scene text recognition and end-to-end scene text spotting models improves by transfer learning on the pre-trained MAE backbone. For example, our recognition model witnessed a 4% word recognition accuracy on the OST dataset. Our end-to-end text spotting model achieved 68.5% F-measure performance outperforming the stat-of-the-art methods when equipped with an MAE backbone compared to a convolutional neural network (CNN) backbone on the OCTT dataset.

49 Classification of handwritten annotations in mixed-media documents

Amanda Dash (University of Victoria); Alexandra Branzan Albu (University of Victoria)*

Abstract: Handwritten annotations in documents contain valuable information, but they are challenging to detect and identify. This paper addresses this challenge. We propose an algorithm for generating a novel mixed-media document dataset, Annotated Docset, that consists of 14 classes of machine-printed and handwritten elements and annotations. We also propose a novel loss function, Dense Loss, which can correctly identify small objects in complex documents when used in fully convolutional networks (e.g. U-NET, DeepLabV3+). Our Dense Loss function is a compound function that uses local region homogeneity to promote contiguous and smooth segmentation predictions while also using an L1-norm loss to reconstruct the dense-labelled ground truth. By using regression instead of a probabilistic approach to pixel classification, we avoid the pitfalls of training on datasets with small or underrepresented objects. We show that our loss function outperforms other semantic segmentation loss functions for imbalanced datasets, containing few elements that occupy small areas. Experimental results show that the proposed method achieved a mean Intersection-over-Union (mIoU) score of 0.7163 for all document classes and 0.6290 for handwritten annotations, thus outperforming state-of-the-art loss functions.

8 Oral Session 5: Reasoning in 3D

40 Learned Intrinsic Auto-Calibration from Fundamental Matrices

Karim Samaha (American University of Beirut); Georges Y Younes (University of Waterloo); Daniel Asmar (American University of Beirut); John Zelek (University of Waterloo)*

Abstract: Auto-calibration that relies on unconstrained image content and epipolar relationships is necessary in online operations, especially when internal calibration parameters such as focal length can vary. In contrast, traditional calibration relies on a checkerboard and other scene information and are typically conducted offline. Unfortunately, auto-calibration may not always converge when solved traditionally in an iterative optimization formalism. We propose to solve for the intrinsic calibration parameters using a neural network that is trained on a synthetic Unity dataset that we created. We demonstrate our results on both synthetic and real data to validate the generalizability of our neural network model, which outperforms traditional methods by 2% to 30, and outperforms recent deep learning approaches by a factor of 2 to 4 times.

44 3DVQA: Visual Question Answering for 3D Environments

Yasaman Etesam (Simon Fraser University); Leon Kochiev (Simon Fraser University); Angel X Chang (Simon Fraser University)*

Abstract: Visual Question Answering (VQA) is a widely studied problem in computer vision and natural language processing. However, current approaches to VQA have been investigated primarily in the 2D image domain. We study VQA in the 3D domain, with our input being point clouds of real-world 3D scenes, instead of 2D images. We believe that this 3D data modality provide richer spatial relation information that is of interest in the VQA task. In this paper,

we introduce the 3DVQA-ScanNet dataset, the first VQA dataset in 3D, and we investigate the performance of a spectrum of baseline approaches on the 3D VQA task.

9 Poster Session - 1

2 Multiple Classifiers Based Adversarial Training for Unsupervised Domain Adaptation

*Yiju Yang (The University of Kansas); Taejoon Kim (University of Kansas); Guanghui Wang (Ryerson University)**

Abstract: Adversarial training based on the maximum classifier discrepancy between two classifier structures has achieved great success in unsupervised domain adaptation tasks for image classification. The approach adopts the structure of two classifiers, though simple and intuitive, the learned classification boundary may not well represent the data property in the new domain. In this paper, we propose to extend the structure to multiple classifiers to further boost its performance. To this end, we develop a very straightforward approach to adding more classifiers. We employ the principle that the classifiers are different from each other to construct a discrepancy loss function for multiple classifiers. The proposed construction method of loss function makes it possible to add any number of classifiers to the original framework. The proposed approach is validated through extensive experimental evaluations. We demonstrate that, on average, adopting the structure of three classifiers normally yields the best performance as a trade-off between accuracy and efficiency. With minimum extra computational costs, the proposed approach can significantly improve the performance of the original algorithm. The source code of the proposed approach can be downloaded from https://github.com/rucv/MMCD_DA.

3 Semi-supervised Grounding Alignment for Multi-modal Feature Learning

Shih-Han Chou (University of British Columbia); Zicong Fan (ETH Zurich); Jim Little (University of British Columbia, Canada); Leonid Sigal (University of British Columbia)*

Abstract: Self-supervised transformer-based architectures, such as ViLBERT [22] and others, have recently emerged as dominant paradigms for multi-modal feature learning. Such architectures leverage large-scale datasets (e.g., Conceptual Captions [27]) and, typically, image-sentence pairings, for self-supervision. However, conventional multi-modal feature learning requires huge datasets and computing for both pre-training and fine-tuning to the target task. In this paper, we illustrate that more granular semi-supervised alignment at a region-phrase level is an additional useful cue and can further improve the performance of such representations. To this end, we propose a novel semi-supervised grounding alignment loss, which leverages an off-the-shelf pre-trained phrase grounding model for pseudo-supervision (by producing region-phrase alignments). This semi-supervised formulation enables better feature learning in the absence of any additional human annotations on the large-scale (Conceptual Captions) dataset. Further, it shows an even larger margin of improvement on smaller data splits, leading to effective data-efficient feature learning. We illustrate the superiority of the learned features by fine-tuning the resulting models to multiple vision-language downstream tasks: visual question answering (VQA), visual common-sense reasoning (VCR), and visual grounding. Experiments on the VQA, VCR, and grounding benchmarks demonstrate the improvement of up to 1.3% in accuracy (in visual grounding) with large-scale training; up to 5.9% (in VQA) with 1/8 of the data for pre-training and fine-tuning.

6 The GIST and RIST of Iterative Self-Training for Semi-Supervised Segmentation

Eu Wern Teh (University of Guelph); Terrance DeVries (University of Guelph); Brendan Duke (ModiFace Inc); Ruowei Jiang (ModiFace Inc.); Parham Aarabi (ModiFace Inc.); Graham Taylor (University of Guelph)*

Abstract: We consider the task of semi-supervised semantic segmentation, where we aim to produce pixel-wise semantic object masks given only a small number of human-labeled training examples. We focus on iterative self-training methods in which we explore the behavior of self-training over multiple refinement stages. We show that iterative self-training leads to performance degradation if done naively with a fixed ratio of human-labeled to pseudo-labeled training examples. We propose Greedy Iterative Self-Training (GIST) and Random Iterative Self-Training (RIST) strategies that alternate between training on either human-labeled data or pseudo-labeled data at each refinement stage, resulting in a performance boost rather than degradation. We further show that GIST and RIST can be combined with existing semi-supervised learning methods to boost performance.

7 A View Invariant Human Action Recognition System for Noisy Inputs

*Joo W Kim (Escuela Superior Politecnica del Litoral); Jefferson E Hernandez (Escuela Superior Politecnica del Litoral); Richard G Cobos (Escuela Superior Politecnica del Litoral); Ricardo J Palacios (Escuela Superior Politecnica del Litoral); Andres G. Abad (Escuela Superior Politecnica del Litoral)**

Abstract: We propose a skeleton-based Human Action Recognition (HAR) system, robust to both noisy inputs and perspective variation. This system receives RGB videos as input and consists of three modules: (M1) 2D Key-Points Estimation module, (M2) Robustness module, and (M3) Action Classification module; of which M2 is our main contribution. This module uses pre-trained 3D pose estimator and pose refinement networks to handle noisy information including missing points and uses rotations of the 3D poses to add robustness to camera viewpoint variation. To evaluate our approach, we carried out comparison experiments between models trained with M2 and without it. These experiments were conducted on the UESTC view-varying dataset, on the i3DPost multi-view human action dataset and on a Boxing Actions dataset, created by us. Our system achieved positive results, improving the accuracy by 24%, 3% and 11% on each dataset, respectively. On the UESTC dataset, our method achieves the new state of the art for the cross-view evaluation protocols.

11 Adaptive Memory Management for Video Object Segmentation

Ali Pourganjalikhan (Concordia University); Charalambos Poullis (Concordia University)*

Abstract: Matching-based networks have achieved state-of-the-art performance for video object segmentation (VOS) tasks by storing every-k frames in an external memory bank for future inference. Storing the intermediate frames' predictions provides the network with richer cues for segmenting an object in the current frame. However, the size of the memory bank gradually increases with the length of the video, which slows down inference speed and makes it impractical to handle arbitrary-length videos. This paper proposes an adaptive memory bank strategy for matching-based networks for semi-supervised video object segmentation (VOS) that can handle videos of arbitrary length by discarding obsolete features. Features are indexed based on their importance in the segmentation of the objects in previous frames. Based on the index, we discard unimportant features to accommodate new features. We present our experiments on DAVIS 2016, DAVIS 2017, and Youtube-VOS that demonstrate that our method outperforms state-of-the-art that employ first-and-latest strategy with fixed-sized memory banks and achieves comparable performance to the every-k strategy with increasing-sized memory banks. Furthermore, experiments show that our method increases inference speed by up to 80% over the every-k and 35% over first-and-latest strategies.

12 M2A: Motion Aware Attention for Accurate Video Action Recognition

Brennan Gebotys (University of Waterloo); Alexander Wong (University of Waterloo); David A Clausi (University of Waterloo)*

Abstract: Advancements in attention mechanisms have led to significant performance improvements in a variety of areas in machine learning due to its ability to enable the dynamic modeling of temporal sequences. A particular area in computer vision that is likely to benefit greatly from the incorporation of attention mechanisms in video action recognition. However,

much of the current research's focus on attention mechanisms have been on spatial and temporal attention, which are unable to take advantage of the inherent motion found in videos. Motivated by this, we develop a new attention mechanism called Motion Aware Attention (M2A) that explicitly incorporates motion characteristics. More specifically, M2A extracts motion information between consecutive frames and utilizes attention to focus on the motion patterns found across frames to accurately recognize actions in videos. The proposed M2A mechanism is simple to implement and can be easily incorporated into any neural network backbone architecture. We show that incorporating motion mechanisms with attention mechanisms using the proposed M2A mechanism can lead to a +15% to +26% improvement in top-1 accuracy across different backbone architectures, with only a small increase in computational complexity. We further compared the performance of M2A with other state-of-the-art motion and attention mechanisms on the Something-Something V1 video action recognition benchmark. Experimental results showed that M2A can lead to further improvements when combined with other temporal mechanisms and that it outperforms other motion-only or attention-only mechanisms by as much as +60% in top-1 accuracy for specific classes in the benchmark. We make our code available at: https://github.com/gebob19/M2A.

13 Object Class Aware Video Anomaly Detection through Image Translation

Mohammad Baradaran (Laval University); Robert Bergevin (Laval University)*

Abstract: Semi-supervised video anomaly detection (VAD) methods formulate the task of anomaly detection as detection of deviations from the learned normal patterns. Previous works in the field (reconstruction or prediction-based methods) suffer from two drawbacks: 1) They focus on low-level features, and they (especially holistic approaches) do not effectively consider the object classes. 2) Object-centric approaches neglect some of the context information (such as location). To tackle these challenges, this paper proposes a novel two-stream object-aware VAD method that learns the normal appearance and motion patterns through image translation tasks. The appearance branch translates the input image to the target semantic segmentation map produced by Mask-RCNN, and the motion branch associates each frame with its expected optical flow magnitude. Any deviation from the expected appearance or motion in the inference stage shows the degree of potential abnormality. We evaluated our proposed method on the ShanghaiTech, UCSD-Ped1, and UCSD-Ped2 datasets and the results show competitive performance compared with state-of-the-art works. Most importantly, the results show that, as significant improvements to previous methods, detections by our method are completely explainable and anomalies are localized accurately in the frames.

17 Integrating High-Resolution Tactile Sensing into Grasp Stability Prediction

Lachlan J Chumbley (Monash University); Morris Gu (Monash University); Rhys Newbury (Monash University); Akansel Cosgun (Monash University); Jurgen Leitner (LYRO Robotics)*

Abstract: We investigate how high-resolution tactile sensors can be utilized in combination with vision and depth sensing, to improve grasp stability prediction. Recent advances in simulating high-resolution tactile sensing, in particular the TACTO simulator, enabled us to evaluate how neural networks can be trained with a combination of sensing modalities. With the large amounts of data needed to train large neural networks, robotic simulators provide a fast way to automate the data collection process. We expand on the existing work through an ablation study and an increased set of objects taken from the YCB benchmark set. Our results indicate that while the combination of vision, depth, and tactile sensing provides the best prediction results on known objects, the network fails to generalize to unknown objects. Our work also addresses existing issues with robotic grasping in tactile simulation and how to overcome them.

21 The Lasso Method for Multi-Robot Foraging

*Andrew Vardy (Memorial University of Newfoundland)**

Abstract: We propose a novel approach to multi-robot foraging. This approach makes use of a scalar field to guide robots throughout an environment while gathering objects towards the goal. The environment must be planar with a closed, contiguous boundary. However, the boundary's shape can be arbitrary. Conventional robot foraging methods assume an open environment or a simple boundary that never impedes the robots—a limitation which our method overcomes. Our distributed control algorithm causes the robots to circumnavigate the environment and nudge objects inwards towards the goal. We demonstrate the performance of our approach using real-world and simulated experiments and study the impact of the number of robots, the complexity of the boundary, and limitations on the sensing range.

10 Poster Session - 2

10 Understanding the impact of image and input resolution on deep digital pathology patch classifiers

Eu Wern Teh (University of Guelph); Graham Taylor (University of Guelph)*

Abstract: We consider annotation efficient learning in Digital Pathology (DP), where expert annotations are expensive and thus scarce. We explore the impact of image and input resolution on DP patch classification performance. We use two cancer patch classification datasets PCam and CRC, to validate the results of our study. Our experiments show that patch classification performance can be improved by manipulating both the image and input resolution in annotation-scarce and annotation-rich environments. We show a positive correlation between the image and input resolution and the patch classification accuracy on both datasets. By exploiting the image and input resolution, our final model trained on < 1% of data performs equally well compared to the model trained on 100% of data in the original image resolution on the PCam dataset.

26 Attention based Occlusion Removal for Hybrid Telepresence Systems

Surabhi Gupta (IIIT Hyderabad); Ashwath Shetty (IIIT Hyderabad); Avinash Sharma (IIIT Hyderabad)*

Abstract: Traditionally, video conferencing is a widely adopted solution for remote communication, but a lack of immersiveness comes inherently due to the 2D nature of facial representation. The integration of Virtual Reality (VR) in a communication/telepresence system through Head Mounted Displays (HMDs) promises to provide users with a much better immersive experience. However, HMDs cause hindrance by blocking the facial appearance and expressions of the user. We propose a novel attention-enabled encoder-decoder architecture for HMD deocclusion to overcome these issues. We also propose to train our person-specific model using short videos of the user, captured in varying appearances, and demonstrated generalization to unseen poses and appearances of the user. We report superior qualitative and quantitative results over state-of-the-art methods. We also present applications of this approach to hybrid video teleconferencing using existing animation and 3D face reconstruction pipelines.

32 Improving tracking with a tracklet associator

Remi Nahon (Polytechnique Montréal); Guillaume-Alexandre Bilodeau (Polytechnique Montréal); Gilles Pesant (Polytechnique Montréal)*

Abstract: Multiple object tracking (MOT) is a task in computer vision that aims to detect the position of various objects in videos and to associate them to a unique identity. While recent advances in machine learning have led to a huge performance gain for the detection phase in MOT, the association phase remains a challenge, especially because of its combinatorial complexity. We propose an approach based on Constraint Programming (CP) whose goal is to be grafted to any existing tracker in order to improve its object association results. We developed a modular algorithm divided into three independent phases. The first phase consists in recovering the tracklets provided by a base tracker and to cut them at the places where uncertain associations are spotted, for example, when tracklets overlap, which may

cause identity switches. In the second phase, we associate the previously constructed tracklets using a Belief Propagation Constraint Programming algorithm, where we propose various constraints that assign scores to each of the tracklets based on multiple characteristics, such as their dynamics or the distance between them in time and space. Finally, the third phase is a rudimentary interpolation model to fill in the holes remaining in the trajectories we built. Experiments show that our model leads to improvements in the results for all three of the state-of-the-art trackers on which we tested it (3 to 4 points gained on HOTA and IDF1).

37 Anomaly Detection with Adversarially Learned Perturbations of Latent Space

*Vahid Reza Khazaie (University of Western Ontario); Anthony Wong (University of Western Ontario); John Jewell (University of Western Ontario); Yalda Mohsenzadeh (University of Western Ontario)**

Abstract: Anomaly detection is to identify samples that do not conform to the distribution of the normal data. Due to the unavailability of anomalous data, training a supervised deep neural network is a cumbersome task. As such, unsupervised methods are preferred as a common approach to solve this task. Deep autoencoders have been broadly adopted as a base of many unsupervised anomaly detection methods. However, a notable shortcoming of deep autoencoders is that they provide insufficient representations for anomaly detection by generalizing to reconstruct outliers. In this work, we have designed an adversarial framework consisting of two competing components, an Adversarial Distorter, and an Autoencoder. The Adversarial Distorter is a convolutional encoder that learns to produce effective perturbations and the autoencoder is a deep convolutional neural network that aims to reconstruct the images from the perturbed latent feature space. The networks are trained with opposing goals in which the Adversarial Distorter produces perturbations that are applied to the encoder's latent feature space to maximize the reconstruction error and the autoencoder tries to neutralize the effect of these perturbations to minimize it. When applied to anomaly detection, the proposed method learns semantically richer representations due to applying perturbations to the feature space. The proposed method outperforms the existing state-of-the-art methods in anomaly detection on image and video datasets.

38 An Exact Fast Fourier Method for Morphological Dilation and Erosion Using the Umbra Technique

Vivek Sridhar (Brandenburg Technical University); Michael Breuß (Brandenburg Technical University)*

Abstract: In this paper we consider the fundamental operations dilation and erosion of mathematical morphology. It is well known that many powerful image filtering operations can be constructed by their combinations. We propose a fast and novel algorithm based on the Fast Fourier Transform to compute greyvalue morphological operations on an image. The novel method may deal with non-flat filters and incorporates no restrictions on shape and size of the filtering window, in contrast to many other fast methods in the field. Unlike fast Fourier techniques from previous works, the novel method gives exact results and is not an approximation. The key aspect which allows to achieve this is to explore here for the first time in this context the umbra formulation of images and filters. We show that the new method is in practice particularly suitable for filtering images with smalltonal range or when employing large filter sizes.

46 Monocular Robot Navigation with Self-Supervised Pretrained Vision Transformers

Miguel Angel Saavedra-Ruiz (Université de Montréal); Sacha Morin (Université de Montréal); Liam Paull (Université de Montréal)*

Abstract: In this work, we consider the problem of learning a perception model for monocular robot navigation using few annotated images. Using a Vision Transformer (ViT) pretrained with a label-free self-supervised method, we successfully train a coarse image segmentation model

for the Duckietown environment using 70 training images. Our model performs coarse image segmentation at the 8x8 patch level, and the inference resolution can be adjusted to balance prediction granularity and real-time perception constraints. We study how best to adapt a ViT to our task and environment, and find that some lightweight architectures can yield good single-image segmentation at a usable frame rate, even on CPU. The resulting perception model is used as the backbone for a simple yet robust visual servoing agent, which we deploy on a differential drive mobile robot to perform two tasks: lane following and obstacle avoidance.

47 TemporalNet: Real-time 2D-3D Video Object Detection

*Meihong Chen (University of Ottawa); Jochen Lang (University of Ottawa)**

Abstract: Designing a video detection network based on state-of-the-art single-image object detectors may seem like an obvious choice. However, video object detection has extra challenges due to the lower quality of individual frames in a video, and hence the need to include temporal information for high-quality detection results. We design a novel interleaved architecture combining a 2D convolutional network and a 3D temporal network. To explore inter-frame information, we propose feature aggregation based on a temporal network. Our TemporalNet utilizes Appearance preserving 3D convolution (AP3D) for extracting aligned features in the temporal dimension. Our temporal network functions at multiple scales for better performance, which allows communication between 2D and 3D blocks at each scale and also across scales. Our TemporalNet is a plug-and-play block that can be added to a multi-scale single-image detection network without any adjustments in the network architecture. When TemporalNet is applied to Yolov3 it is real-time with a running time of 35ms/frame on a low-end GPU. Our real-time approach achieves 77.1% mAP (mean Average Precision) on ImageNet VID 2017 dataset with TemporalNet-4, where TemporalNet-16 achieves 80.9% mAP which is a competitive result.

48 Safe Landing Zones Detection for UAVs Using Deep Regression

Sakineh Abdollahzadeh (Université du Québec en Outaouais); Pier-Luc Proulx (Université du Québec en Outaouais); Mohand Said Allili (Universite du Quebec en Outaoauis); Jean-François Lapointe (National Research Council Canada (NRC))*

Abstract: Finding safe landing zones (SLZ) in urban areas and natural scenes is one of the many challenges that must be overcome in automating Unmanned Aerial Vehicles (UAV) navigation. Using passive vision sensors to achieve this objective is a very promising avenue due to their low cost and the potential they provide for performing simultaneous terrain analysis and 3D reconstruction. In this paper, we propose using a deep learning approach on UAV imagery to assess the SLZ. The model is built on a semantic segmentation architecture whereby thematic classes of the terrain are mapped into safety scores for UAV landing. Contrary to past methods, which use hard classification into safe/unsafe landing zones, our approach provides a continuous safety map that is more practical for an emergency landing. Experiments on public datasets have shown promising results.

53 CellDefectNet: A Machine-designed Attention Condenser Network for Electroluminescence-based Photovoltaic Cell Defect Inspection

Carol Xu (DarwinAI); Mahmoud Famouri (Darwin AI); Gautam Bathla (DarwinAI); Saeejith Nair (University of Waterloo); Mohammad Javad Shafiee (University of Waterloo); Alexander Wong (University of Waterloo)*

Abstract: Photovoltaic cells are electronic devices that convert light energy to electricity, forming the backbone of solar energy harvesting systems. An essential step in the manufacturing process for photovoltaic cells is visual quality inspection using electroluminescence imaging to identify defects such as cracks, finger interruptions, and broken cells. A big challenge faced by industry in photovoltaic cell visual inspection is the fact that it is currently done manually by human inspectors, which is extremely time consuming, laborious, and prone to human error. While deep learning approaches holds great potential to automating this inspection,

the hardware resource-constrained manufacturing scenario makes it challenging for deploying complex deep neural network architectures. In this work, we introduce CellDefectNet, a highly efficient attention condenser network designed via machine-driven design exploration specifically for electroluminesence-based photovoltaic cell defect detection on the edge. We demonstrate the efficacy of CellDefectNet on a benchmark dataset comprising of a diversity of photovoltaic cells captured using electroluminescence imagery, achieving an accuracy of 86.3 while possessing just 410K parameters (13x lower than EfficientNet-B0, respectively) and 115M FLOPs (12x lower than EfficientNet-B0) and 13x faster on an ARM Cortex A-72 embedded processor when compared to EfficientNet-B0.