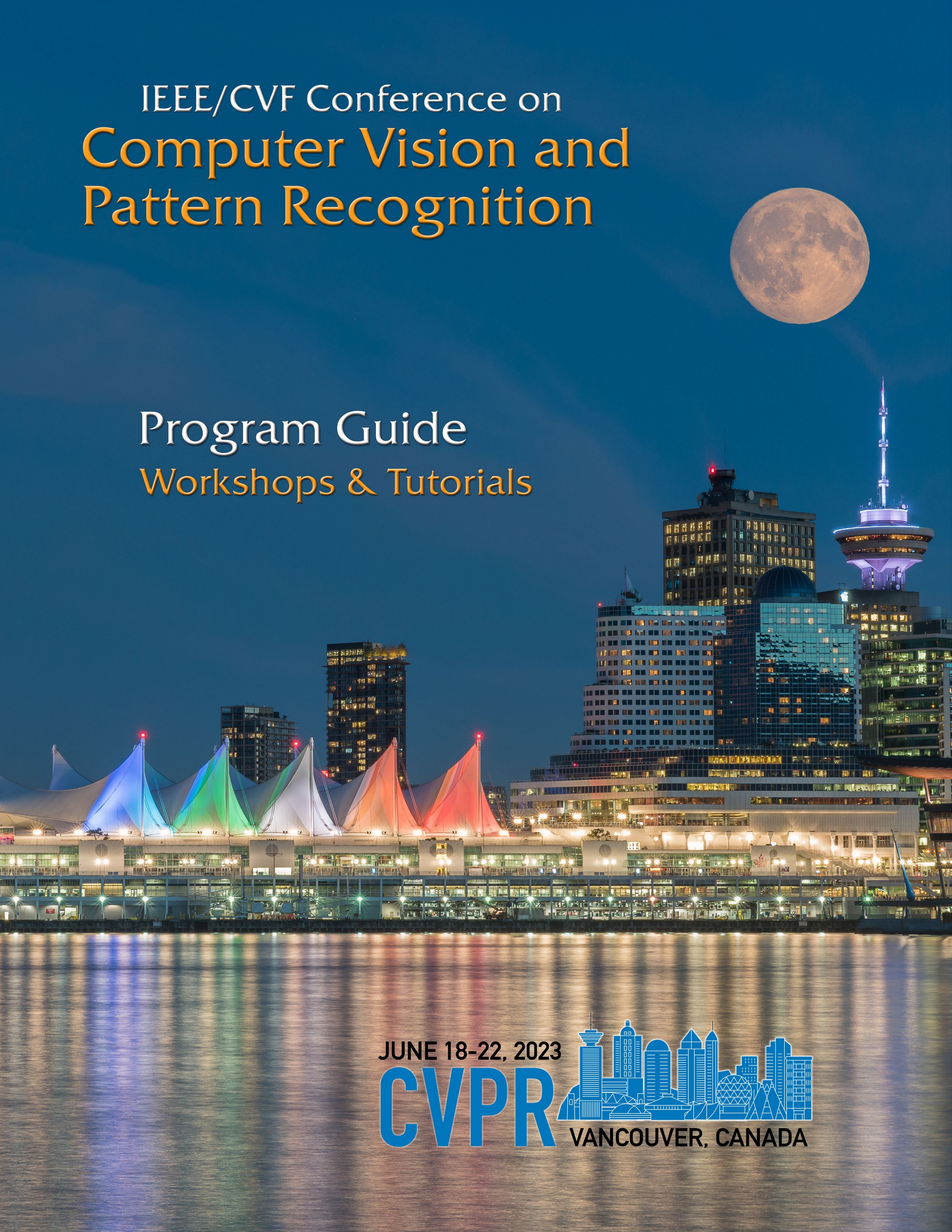
IEEE/CVF Conference on

Computer Vision and Pattern Recognition

Program Guide

Workshops & Tutorials



JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

PLATINUM SPONSORS

amazon | science

ANT RESEARCH

cruise

Google

λ Lambda

Qualcomm

TOYOTA RESEARCH INSTITUTE

GOLD SPONSORS

Alibaba Cloud

Bai du 百度

datagen

FURIOSA

LATITUDE

NEURAL MAGIC

NOVARC TECHNOLOGIES

speechocean

Synthesis AI

TELUS® International

TESLA

TikTok

Weights & Biases

ZOOX

SILVER SPONSORS

Dataminr®

DATATANG

DDD
Changing How
the World Works.

HYUNDAI

kitware

LG AI Research

Lightning AI

manot

meitu | Lab
Meitu Imaging & Vision Lab

PROPHESEE
METAVISION FOR MACHINES

RIVIAN

scale

CHEN TIANQIAO & CHRISSY INSTITUTE

Visual Layer

WAYMO

Message from the General and Program Chairs

Welcome to the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition in Vancouver, Canada. As in previous years, CVPR is the premier and flagship annual meeting of IEEE/CVF and PAMI-TC, where researchers in our community present their latest advances in computer vision, pattern recognition, machine learning, robotics, and artificial intelligence, both in theory and practice. Our program includes invited keynote talks, award paper presentations, poster presentations, tutorials, workshops, demos, exhibitions, and an amiable social setting, all aimed at providing attendees with an exciting and enriching experience.

This year marks the first time in a while that many pandemic restrictions have been lifted, allowing us to come together in person again to celebrate the latest advances in our field. For those unable to join us physically, we are pleased to offer a virtual component that will provide access to conference papers, posters, videos, and talks. We hope this virtual option will allow everyone to engage with the exciting research being presented.

CVPR 2023 received 9155 submissions, a 12% increase from the 8161 submissions to CVPR 2022. The review process was managed by 400+ area chairs and, new to the process this year, 30 senior area chairs. The senior area chairs helped in a number of respects, most importantly by adjudicating difficult cases, covering emergencies, selecting highlight papers, and selecting the award candidates. During the review phase, each paper received at least 3 reviews from the pool of 6625 reviewers. As in prior years, after receiving these initial reviews, the process continued with an author rebuttal phase, discussion among reviewers and ACs, finalizing of reviews, and ACs working in triplets to make final accept/reject decisions for each paper. At the end of this process, 2359 papers were accepted (25.8% acceptance rate). In keeping with the CVPR tradition, the PCs did not pre-set any acceptance cap. The resulting acceptance rate reflects the community consensus, and is well aligned with past CVPRs.

Of the 2359 accepted papers, 235 (10%) were selected as highlights. In addition, 12 (0.5%) papers have been shortlisted as best paper award candidates. The final best papers and honorable mentions are selected from these 12 papers by an independent award committee appointed by the program chairs, which is composed of experienced researchers from our community. The award committee is led by an award committee chair appointed by the program chairs, who moderates the selection process.

This year, CVPR will be single-track to allow everyone to attend everything. The focus will be on a few plenary talks, keynotes and panels, and plenty of time for poster sessions, networking, and socializing. Every paper will be presented at a poster session. All paper award candidates will have an additional plenary oral presentation. Every attendee will have access to a personalized digital program to easily navigate the ~400 posters in each poster session. The virtual platform will host papers, posters, videos, and a chat room for every paper. All plenary events will be streamed online for all attendees that cannot attend in person.

We would like to thank everyone involved in making CVPR 2023 a success. This includes the organizing committee, area chairs, senior area chairs, reviewers, authors, demo session participants, donors, exhibitors, and everyone else without whom this meeting would not be possible. We also thank Nicole Finn and her C to C Events team for organizing the conference logistics, Lee Campbell and the Event Hosts team for their work on the website and virtual platform, and Mike Weil and Hall Erickson for handling sponsorships and the exhibition. Last but not least, we thank all of you for attending CVPR 2023 and making it one of the top venues for computer vision research in the world. We hope that you also have some time to explore gorgeous Vancouver during the conference. Enjoy CVPR 2023. We look forward to meeting you in person!

Program Chairs: Andreas Geiger, Ross Girshick, Judy Hoffman,
Vladlen Koltun, and Svetlana Lazebnik

General Chairs: Michael S. Brown, Fei-Fei Li,
Greg Mori, and Yoichi Sato

CVPR 2023 Organizing Committee

General Chairs:	Michael S. Brown
	Fei-Fei Li
	Greg Mori
	Yoichi Sato
Program Chairs:	Andreas Geiger
	Ross Girshick
	Judy Hoffman
	Vladlen Koltun
	Svetlana Lazebnik
Workshops Chairs:	Olga Russakovsky
	Yu Wu
	Serena Yeung
Tutorials Chairs:	Siyu Tang
	Jianxin Wu
Demonstrations Chairs:	Jon Barron
	Gim Hee Lee
Doctoral Consortium Chairs:	'YZ' Yezhou Yang
	Catherine Qi Zhao
Program Advisory Board:	Michael J. Black
	David Forsyth
	Kristen Grauman
	Jitendra Malik
	Cordelia Schmid
	Richard Szeliski
	Andrew Zisserman
Diversity, Equity, & Inclusion Chairs:	Thibaut Durand
	Fatma Güney
	Kate Saenko
Publicity Chairs:	Kosta Derpanis
	Boqing Gong
	Abby Stylianou
Social Chair:	Yale Song
Social Activities Chairs:	Angel Chang
	Giovanni Farinella
Local Arrangements Chairs:	Kwang Moo Yi
	Leon Sigal
Virtual Platform Chair:	Andreas Geiger
Accessibility Chair:	Danna Gurari
Finance Chair:	Bryan Morse
Technical Chair:	David Hafner
Webmaster:	Lee Campbell
Senior PAMI-TC Ombuds:	David Forsyth
	Linda Shapiro
CVPR 2023 Ombuds:	Kyoung Mu Lee
	Xiaodan Liang
Publications Chair:	Eric Mortensen
Event Producer:	Nicole Bumpus Finn

Saturday, June 17

1100–2000 Registration (West Ballroom Foyer)

Sunday, June 18

NOTE: Tutorial rooms are subject to change. Refer to the online site for up-to-date locations. Use the QR code for each tutorial to see its schedule. Here is the QR code for the CVPR 2023 Tutorials page.

0700–1700 Registration (West Ballroom Foyer)

0700–0900 Breakfast (West Ballrooms A–D)

1000–1045 Morning Break West Ballrooms A–D

1145–1345 Lunch (West Ballrooms A–D)

1500–1545 Afternoon Break (West Ballrooms A–D)

Tutorial: A Comprehensive Tour and Recent Advancements Toward Real-World Visual Geo-Localization

Organizers: Rakesh "Teddy" Kumar
Chen Chen
Mubarak Shah
Han-Pang Chiu
Sijie Zhu

Location: East 6

Time: Full Day (0830-0530)

Summary: Precise geo-location of a ground image within a large-scale environment is crucial to many applications, including autonomous vehicles, robotics, wide area augmented reality and image search. Localizing the ground image by matching to an aerial/ overhead geo-referenced database has gained noticeable momentum in recent years, due to significant growth in the availability of public aerial/ overhead data with multiple modalities (such as aerial images from Google/ Bing maps, and USGS 2D and 3D data, Aerial LiDAR data, Satellite 3D Data etc.). Matching a ground image to aerial/ overhead data, whose acquisition is simpler and faster, also opens more opportunities to industrial and consumer applications. However, cross-view and cross-modal visual geo-localization comes with additional technical challenges due to dramatic changes in appearance between the ground image and aerial/ overhead database, which capture the same scene differently in time, viewpoints or/and sensor modalities. This tutorial will provide a comprehensive review on the research problem of visual geo-localization, including same-view/cross-time, cross-view, cross-modal settings to both new and experienced researchers. It also provides connection opportunities for the researchers of visual geo-localization and other related fields.

Tutorial: Recent Advances in Anomaly Detection

Organizers: Guansong Pang
Joey Tianyi Zhou
Radu Tudor Ionescu
Yu Tian
Kihyuk Sohn

Location: East 18

Time: Full Day (0830-0515)

Summary: The tutorial will present a comprehensive review of recent advances in (deep) anomaly detection on image and video data. Three major AD paradigms will be discussed, including unsupervised/self-supervised approaches (anomaly-free training data), semi-supervised approaches (few-shot training anomaly examples are available), and weakly-supervised approaches (videl-level labels are available for frame-level detection). Additionally, we will also touch on anomaly segementation tasks, focusing on autonomous driving settings. The tutorial will be ended with a panel discussion on AD challenges and opportunities.

Tutorial: ML Systems for Large Models and Federated Learning

Organizers: Qirong Ho
Samuel Horvath
Hongyi Wang

Location: East 5

Time: Half Day - Morning (0830-1145)

Summary: This tutorial will teach attendees how to overcome performance, cost, privacy, and robustness challenges when using distributed and federated software systems for learning and deploying Computer Vision and ML applications across various hardware settings (networked machines, GPUs, embedded, mobile systems). The audience will learn about theory, implementation, and practice of these topics: state-of-the-art approaches and system architectures, forms of distributed parallelism, pitfalls in the measurement of parallel application performance, parallel ML compilers, computation-communication-memory efficiency in federated learning (FL), trustworthy FL, tackling device heterogeneity in FL, and on-device FL systems.

Tutorial: Efficient Neural Networks: From Algorithm Design to Practical Mobile Deployment

Organizers: Jian Ren
Sergey Tulyakov
Ju Hu

Location: West 212

Time: Half Day - Morning (0830-1200)

Summary: This tutorial will introduce effective methodologies for re-designing algorithms for efficient content understanding, image generation, and neural rendering. Most importantly, we show how the algorithms can be efficiently deployed on mobile devices, eventually achieving real-time interaction between users and mobile devices.

Tutorial: Skull Restoration, Facial Reconstruction and Expression

Organizers: Xin Li
Lan Xu
Yu Ding

Location: Virtual

Time: Half Day - Morning (0930-1145)

Summary: This tutorial focuses on the challenges of reconstructing a 3D model of a human face followed by generating facial expressions. It comprises three parts, covering facial reconstruction from skeletal remains, 4D dynamic facial performance, and audio-driven talking face generation. First, face modeling is a fundamental technique and has broad applications in animation, vision, games, and VR. Facial geometries are fundamentally governed by their underlying skull and tissue structures. This session covers a forensic task of facial reconstruction from skeletal remains, in which we will discuss how to restore fragmented skulls, model anthropological features, and reconstruct human faces upon skulls. Then, we will detail how to capture 4D facial performance, which is the foundation for face modeling and rendering. We will consider the hardware designs for cameras, sensors, lighting, and the steps to obtain dynamic facial geometry along with physically-based textures (pore-level diffuse albedo, specular intensity, normal, etc.,). We will discuss the two complementary workhorses: multi-view stereo and photometric stereo, and the combination with neural rendering advances and medical imaging. Finally, talking face generation will be discussed including 3D animation parameters and 2D photo-realistic video, as well as their applications. It aims to create a talking video of a speaker with authentic facial expressions from an input of simultaneous speech. The face identity may be from a predefined 3D virtual character, a single image, or a few minutes of a specific speaker.

Tutorial: Denoising Diffusion Models: A Generative Learning Big Bang

Organizers: Jiaming Song
Chenlin Meng
Arash Vahdat

Location: West 202-204

Time: Half Day - Morning (0900-1230)

Summary: Diffusion models have been widely adopted in various computer vision applications and are becoming a dominating class of generative models. In the year 2022 alone, diffusion models have been applied to many large-scale text-to-image foundation models, such as DALL-E 2, Imagen, Stable Diffusion and eDiff-I. These developments have also driven novel computer vision applications, such as solving inverse problems, semantic image editing, few-shot textual inversion, prompt-to-prompt editing, and lifting 2d models for 3d generation. This popularity is also reflected in the diffusion models tutorial in CVPR 2022, which has accumulated nearly 60,000 views on YouTube over 8 months. The primary goal of the CVPR 2023 tutorial on diffusion models is to make diffusion models more accessible to a wider computer vision audience and introduce recent developments in diffusion models. We will present successful practices on training and sampling from diffusion models and discuss novel applications that are enabled by diffusion models in the computer vision domain. These discussions will also heavily lean on recent research developments that are released in 2022 and 2023. We hope that this year's tutorial on diffusion models will attract more computer vision practitioners interested in this topic to make further progress in this exciting area.

Tutorial: Boosting Computer Vision Research With OpenMMLab and OpenDataLab

Organizers: Kai Chen
Conghui He
Yanhong Zeng
Songyang Zhang
Wenwei Zhang

Location: East 12

Time: Half Day - Morning (0900-1200)

Summary: This tutorial will introduce two open platforms which can significantly accelerate the research in computer vision — OpenMMLab and OpenDataLab.

OpenMMLab is an open-source algorithm platform for computer vision. It aims to provide a solid benchmark and promote reproducibility for academic research. We have released more than 30 high-quality projects and toolboxes in various research areas such as image classification, object detection, semantic segmentation, action recognition, etc. OpenMMLab has made public more than 300 algorithms and 2,400 checkpoints. Over the past years, OpenMMLab has gained popularity in both academia and industry. It receives over 78,000 stars on GitHub and involves more than 1,700 contributors in the community.

OpenDataLab, which was initially released in March, 2022, is an open data platform for artificial intelligence, especially including a large number of datasets for computer vision.

Tutorial: Trustworthy AI in the Era of Foundation Models

Organizers: Pin-Yu Chen
Chaowei Xiao

Location: East 14

Time: Half Day - Morning (0830-1145)

Summary: While machine learning (ML) models have achieved great success in many perception applications, concerns have risen about their potential security, robustness, privacy, and transparency issues when applied to real-world applications. Irresponsibly applying a foundation model to mission-critical and human-centric domains can lead to serious misuse, inequity issues, negative economic and environmental impacts, and/or legal and ethical concerns. For example, ML models are often regarded as "black boxes" and can produce unreliable, unpredictable, and unexplainable outcomes, especially under domain shifts or maliciously crafted attacks, challenging the reliability of safety-critical applications; Stable Diffusion may generate NSFW content and privacy violated-content.

The goals of this tutorial are to:

- Provide a holistic and complementary overview of trustworthiness issues, including security, robustness, privacy, and societal issues to allow a fresh perspective and some reflection on the induced impacts and responsibility as well as introduce the potential solutions.
- Promote awareness of the misuse and potential risks in existing AI techniques and, more importantly, to motivate rethinking of trustworthiness in research.
- Present case studies from computer vision-based applications.

This tutorial will provide sufficient background for participants to understand the motivation, research progress, known issues, and ongoing challenges in trustworthy perception systems, in addition to pointers to open-source libraries and surveys.

Tutorial: All Things ViTs: Understanding and Interpreting Attention in Vision

Organizers: Hila Chefer
Sayak Paul
Location: West 211
Time: Half Day - Morning (0900-1200)

Summary: The attention mechanism has revolutionized deep learning research across many disciplines starting from NLP and expanding to vision, speech, and more. Different from other mechanisms, the elegant and general attention mechanism is easily adaptable and eliminates modality-specific inductive biases. As attention becomes increasingly popular, it is crucial to develop tools to allow researchers to understand and explain the inner workings of the mechanism to facilitate better and more responsible use of it. This tutorial focuses on understanding and interpreting attention in the vision and the multi-modal setting. We present state-of-the-art research on representation probing, interpretability, and attention-based semantic guidance, alongside hands-on demos to facilitate interactivity. Additionally, we discuss open questions arising from recent works and future research directions.

Tutorial: Vision Transformer: More Is Different

Organizers: Dacheng Tao
Qiming Zhang
Yufei Xu
Jing Zhang
Location: Virtual
Time: Half Day - Morning (0830-1145)

Summary: Big data contains a tremendous amount of dark knowledge. The community has realized that effectively exploring and using such knowledge is essential to achieving superior intelligence. How can we effectively distill the dark knowledge from ultra-large-scale data? One possible answer is: “through Transformers”. Transformers have proven their prowess at extracting and harnessing dark knowledge from data. This is because more is truly different when it comes to Transformers. This tutorial will introduce the structural design, training methods, and applications of Vision Transformers. We will start with the development of neural networks and introduce their theoretical foundations through CNNs to visual transformers. Then, we will discuss the structural design of Vision Transformers, including the plain Vision Transformer and hierarchical Vision Transformers, followed by a discussion of how to train these models in a supervised, self-supervised, and multi-modality way. Next, we will present the applications of Vision Transformers to both low-level tasks and high-level tasks, which have redefined the art of computer vision. Finally, we discuss the open challenges of current Vision Transformers and give future expectations for Vision Transformer developments.

Graduate students, engineers, and researchers interested in or working on image processing, computer vision, deep learning, etc., are highly encouraged to attend the talk.

Tutorial: Recent Advances in Visual Domain Adaptation and Generalization

Organizers: Ronghang Zhu
Xiang Yu
Sheng Li
Location: West 215-216
Time: Half Day - Morning (0830-1135)

Summary: This tutorial delves into the intriguing research fields of visual domain adaptation and domain generalization. Domain adaptation focuses on transferring knowledge from a source domain to a target domain, assuming access to target data during model training. Conversely, domain generalization poses a more challenging scenario, where no target domain data are available, necessitating the model to generalize to unseen domains without any prior knowledge. Despite extensive study, both domain adaptation and domain generalization encounter practical hurdles, including long-tailed distribution and open-set label space issues. In this tutorial, we will provide a concise yet comprehensive overview of visual domain adaptation and domain generalization. The tutorial primarily emphasizes recent advancements in these areas, highlighting key topics such as class-imbalanced domain adaptation, universal domain adaptation, open-set domain adaptation, and single domain generalization. By exploring these cutting-edge techniques, participants will gain insights into tackling the inherent challenges posed by domain adaptation and generalization. Moreover, this tutorial goes beyond theoretical discussions and demonstrates the practical implications of visual domain adaptation and generalization techniques across various fields. By uncovering promising applications in diverse domains, attendees will discover the immense potential and real-world impact of these techniques.

Notes:

Tutorial: Large-Scale Deep Learning Optimization Techniques

Organizers: James Demmel
Yang You
Location: West 208-209
Time: Half Day - Afternoon (1330-1700)

Summary: Large Transformer models have performed promisingly on a wide spectrum of AI and CV applications. These positive performances have thus stimulated a recent surge of extremely large models. However, training these models generally requires more computation and training time. This has generated interest in both academia and industry in scaling up deep learning (DL) using distributed training on high-performance computing (HPC) resources like TPU and GPU clusters.

However, continuously adding more devices will not scale training as intended, since training at a large scale requires overcoming both algorithmic and systems-related challenges. This limitation prevents DL and CV researchers from exploring more advanced model architectures.

Many existing works investigate and develop optimization techniques that overcome these problems and accelerate large model training at a larger-scale. We categorize these works as improving either model accuracy or model efficiency. One method to maintain or improve model accuracy in a large-scale setting, while still maintaining computing efficiency, is to design algorithms that require less communication and memory demands. It is notable that these are not mutually exclusive goals but can be optimized together to further accelerate training. This tutorial helps enable CV members to quickly master optimizations for large-scale DL training and successfully train large models at large-scale with different optimization techniques in a distributed environment.

Tutorial: Contactless Healthcare Using Cameras and Wireless Sensors

Organizers: Wenjin Wang
Xuyu Wang
Daniel McDuff
Location: East 10
Time: Half Day - Afternoon (1330-1700)

Summary: Extracting health-related metrics is an emerging computer vision research topic that has grown rapidly recently. Without needing physical contact, cameras have been used to measure vital signs remotely (e.g., heart & respiration rates, blood oxygenation saturation, body temperature, etc.) from images/video of the skin or body. This leads to contactless, continuous, and comfortable heath monitoring. Cameras can also leverage computer vision and machine learning techniques to measure human behaviors/activities and high-level visual semantic/contextual information, facilitating better understanding of people and scenes for health monitoring and provides a unique advantage compared to the contact bio-sensors. RF (Radar, WiFi, RFID) and acoustic based methods for health monitoring have also been proposed. The rapid development of computer vision and RF sensing also give rise to new multi-modal learning techniques that expand the sensing capability by combining two modalities, while minimizing the need of human labels. Contactless monitoring will bring a rich set of compelling healthcare applications that directly improve upon contact-based monitoring solutions and improve people's care experience and quality of life, such as in care units of the hospital, sleep/senior centers, assisted-living homes, telemedicine and e-health, fitness and sports, driver monitoring in automotive, etc.

Notes:

Sunday, June 18

NOTE: Workshop rooms are subject to change. Refer to the online site for up-to-date locations. Use the QR code for each workshop to see its schedule. Here is the QR code for the CVPR 2023 Workshops page.

0700–1700 Registration (West Ballroom Foyer)

0700–0900 Breakfast (West Ballrooms A–D)

1000–1045 Morning Break West Ballrooms A–D

1145–1345 Lunch (West Ballrooms A–D)

1500–1545 Afternoon Break (West Ballrooms A–D)

Fair, Data-Efficient, and Trusted Computer Vision

Organizers: Nalini Ratha, Srikrishna Karanam, Ziyan Wu, Mayank Vatsa, Richa Singh, Kuan-Chuan Peng, Michele Merler, Kush R. Varshney, Yiming Ying, Sharath Pankanti

Location: West 217-219

Time: Full Day (0800-1630)

Summary: The CVPR 2023 Workshop on Fair, Data-efficient, and Trusted Computer Vision aims to gather researchers and practitioners from academia and industry to discuss advances in all aspects of fairness, data-efficiency, and trust in computer vision. In addition to invited talks from experts in academia and industry, the workshop will solicit and provide a focused venue for new research ideas that seek to address problems related to topics above in a variety of application areas.

Autonomous Driving

Organizers: Vincent Casser, Alexander Liniger, Henrik Kretzschmar, Jose M. Alvarez, Fisher Yu, Yan Wang, Zhaoqi Leng, Maying Shen, Li Erran Li, Dragomir Anguelov, John Leonard, Luc Van Gool

Location: East Ballroom C

Time: Full Day (0915-1815)

Summary: The CVPR 2023 Workshop on Autonomous Driving (WAD) aims to gather researchers and engineers from academia and industry to discuss the latest advances in perception for autonomous driving. In this full-day workshop, we will host speakers as well as technical benchmark challenges to present the current state of the art, limitations and future directions in the field - arguably one of the most promising applications of computer vision and artificial intelligence. Previous chapters of this workshop attracted hundreds of researchers. This year, multiple industry sponsors are also joining our organizing efforts to push it to a new level.

End-to-End Autonomous Driving: Emerging Tasks and Challenges

Organizers: Hongyang Li, Kashyap Chitta, Holger Caesar, Shenlong Wang, Ziwei Liu, Tai Wang, Enze Xie, Huijie Wang, Yang Li

Location: West 110

Time: Full Day (0900-1800)

Summary: The area of autonomous driving has come to a rapid development to handle complicated scenarios and face the challenge of deploying algorithms to feasible massive production. With aid of various machine learning and computer vision techniques, many autonomous driving problems have been resolved. And yet certain key issues, such as safety and explainability for robust L4 solutions, end-to-end autonomous driving framework (and the benefits), bird's-eye-view perception, etc., have not been fully discussed. We have seen the successful holding of recent events at NeurIPS 2022 (incoming), CVPR 2022 (e.g., Embodied AI, Workshop on Autonomous Driving by Waymo/etc), and believe such a workshop is necessary for both the machine learning and computer vision community. This workshop, besides existing editions held at similar venues, serves a brand-new perspective to discuss broad areas of end-to-end framework design for autonomous driving on a system-level consideration. This workshop aims to bring together leading researchers and practitioners to discuss upcoming paradigms for autonomous vehicles. Central to the program is a series of invited talks and four new challenges in the self-driving domain. Each challenge combines new perspectives of multiple components in perception and planning compared to conventional pipelines. Winners of the challenges will present their results and insights as part of the workshop. We invite researchers around the world to build new algorithms to tackle these challenging, real-world autonomous driving tasks!

Generative Models for Computer Vision

Organizers: Adam Kortylewski, Fangneng Zhan, Lingjie Liu, Vincent Sitzmann, Alan Yuille, Christian Theobalt

Location: East Exhibit Hall B

Time: Full Day (0830-1715)

Summary: Recent advances in generative modeling leveraging generative adversarial networks, auto-regressive models, neural fields and diffusion models have enabled the synthesis of near photorealistic images, drastically increasing the visibility and popularity of generative modeling across the computer vision research community. However, these impressive advances in generative modeling have not yet found wide adoption in computer vision for visual recognition tasks. In this workshop, we aim to bring together researchers from the fields of image synthesis and computer vision to facilitate discussions and progress at the intersection of those two subfields. We investigate the question: "How can visual recognition benefit from the advances in generative image modeling?". We invite a diverse set of experts to discuss their recent research results and future directions for generative modeling and computer vision, with a particular focus on the intersection between image synthesis and visual recognition. We hope this workshop will lay the foundation for future development of generative models for computer vision tasks.

Multimodal Content Moderation

Organizers: Mei Chen
Cristian Canton
Davide Modolo
Maarten Sap
Maria Zontak
Chris Bregler

Location: East 17

Time: Full Day (0800-1800)

Summary: Content moderation (CM) is a rapidly growing need in today's world, with a high societal impact, where automated CM systems can discover discrimination, violent acts, hate/toxicity, and much more, on a variety of signals (visual, text/OCR, speech, audio, language, generated content, etc.). Leaving or providing unsafe content on social platforms and devices can cause a variety of harmful consequences, including brand damage to institutions and public figures, erosion of trust in science and government, marginalization of minorities, geo-political conflicts, suicidal thoughts and more. Besides user-generated content, content generated by powerful AI models such as DALL-E and GPT present additional challenges to CM systems.

With the prevalence of multimedia social networking and online gaming, the problem of sensitive content detection and moderation is by nature multimodal. Moreover, content moderation is contextual and culturally multifaceted, for example, different cultures have different conventions about gestures. This requires CM approach to be not only multimodal, but also context aware and culturally sensitive.

Perception Beyond the Visible Spectrum

Organizers: Riad Hammoud
Michael Teutsch
Angel D. Sappa
Erhan Gundogdu
Yi Ding
Wassim El Ahmar
Erik Blasch

Location: East 15

Time: Full Day (0830-1730)

Summary: The Perception Beyond the Visible Spectrum workshop series (IEEE PBVS) has been one of the key events in the Computer Vision and Pattern Recognition (CVPR) community since its inception in 2004. The main objective is to highlight cutting edge advances and state-of-the-art work being made in the field of computer vision in the non-visible spectrum by analyzing, exploiting, and fusing infrared, thermal, radar, SAR, millimeters wave, or LiDAR sensor data. Applications including autonomous driving, aerial robotics, remote sensing, surveillance, and medical computer vision not only show the need for smart data exploitation methods but also the great benefits when intelligently integrating sensor processing, algorithms, and applications. As a result of the improving sensor technologies and simultaneously dropping sensor costs, the PBVS community has been growing exponentially within the last decade. This 19th IEEE CVPR Workshop on PBVS 2023 fosters connections between communities in the machine vision world ranging from public research institutes to private, defense, and federal laboratories. PBVS brings together academic pioneers, industrial and defense researchers and engineers in the field of computer vision, image analysis, pattern recognition, machine learning, signal processing, artificial intelligence, and sensor exploitation. PBVS 2023 is accompanied by three challenges: the 4th Thermal Image Super-Resolution challenge, the Multi-modal Aerial View Object Classification Challenge, and the Multi-modal Aerial View Imagery Challenge.

LatinX in Computer Vision Research

Organizers: Estefanía Talavera
Fabian Caba
Carlos Hinojosa
Laura Montoya

Location: West 107-108

Time: Full Day (0830-1800)

Summary: The LatinX in Computer Vision (LXCV) Research workshop is a one-day event at CVPR 2023 with invited speakers, oral presentations, and posters. The event brings together faculty, graduate students, research scientists, and engineers for an opportunity to connect and exchange ideas. While all presenters will identify primarily as LatinX, all are invited to attend. The primary objective of this workshop is to enhance the visibility of Latin American researchers in the AI and computer vision field while showcasing their latest findings and cutting-edge research contributions.

Media Forensics

Organizers: Hany Farid
Canton Cristian
Luisa Verdoliva

Location: West 105-106

Time: Full Day (0845-1730)

Summary: Generative adversarial networks and diffusion-based synthesis allow for the rapid and automatic generation of highly realistic images and videos (so-called deep fakes). The increasing prevalence of fraud and misuse associated with such fabricated media, have raised the level of interest in the computer vision community. Both academia and industry have addressed this topic in the past, but only recently, with the emergence of more sophisticated ML and CV techniques, has multimedia forensics become a broad and prominent area of research. The recent appearance of relevant datasets (e.g., DFDC, FaceForensics++) and the widespread concerns surrounding synthetic media and misinformation, have turned the field of media forensics and misinformation into a critical research topic. This workshop aims at bringing a heterogeneous group of specialists from academia and industry together to discuss emerging threats, technologies, and mitigation strategies.

Medical Computer Vision

Organizers: Vasileios Belagiannis
Tal Arbel
Tammy Riklin Raviv
Moti Freiman
Ayelet Akselrod-Ballin
Yuyin Zhou
Nicolas Padoy
Dou Qi
Mathias Unberath
Mert Sabuncu

Location: East 1

Time: Full Day (0800-1600)

Summary: The CVPR MCV workshop provides a unique forum for researchers and developers in academia, industry and healthcare to present, discuss and learn about cutting-edge advances in machine learning and computer vision for medical image analysis and computer assisted interventions. The workshop offers a venue for potential new collaborative efforts, encouraging more dataset and information exchanges for important clinical applications.

New Trends in Image Restoration and Enhancement

Organizers: Radu Timofte
Marcos V. Conde
Florin-Alexandru Vasluianu
Ren Yang
Yawei Li
Kai Zhang
Shuhang Gu
Ming-Hsuan Yang
Lei Zhang
Kyoung Mu Lee
Eli Shechtman
Yulan Guo
Codruta O. Ancuti
Cosmin Ancuti
Chao Dong
Xintao Wang
Sira Ferradans
Tom Bishop
Longguang Wang
Yingqian Wang
Fabio Tosi
Pierluigi Zama Ramirez
Luigi Di Stefano
Luc Van Gool

Location: West 306

Time: Full Day (0800-1900)

Summary: Image and video restoration, enhancement, and manipulation are key computer vision problems, encompassing multiple different tasks, including restoration and completion of image information, enhancement of visual quality, and manipulation of image content to achieve a desired effect. Recent years have witnessed an increased interest from the vision and graphics communities in these fundamental topics of research, which has led to substantial progress in many areas. While image manipulation directly relates to image quality enhancement and editing applications, it also forms an important step in a growing range of applications, including surveillance, automotive, electronics, remote sensing, and medical image analysis. The emergence and ubiquitous use of mobile and wearable devices offer another fertile ground for additional applications and faster methods. This workshop aims to provide an overview of the new trends and advances in areas concerning image restoration, enhancement, and manipulation. This workshop builds upon the success of the past editions of the New Trends in Image Restoration and Enhancement (NTIRE) workshop at CVPR and ACCV, the PIRM workshop at ECCV, the workshop and Challenge on Learned Image Compression (CLIC) editions at CVPR, and Advances in Image Manipulation (AIM) workshops at ICCV and ECCV. This workshop features papers addressing topics related to image and video restoration, enhancement, and manipulation and hosts several challenges covering different tasks within those topics.

3D Vision and Robotics

Organizers: Zhenyu Jiang
Kaichun Mo
Dieter Fox
Kristen Grauman
Li Erran Li
Yuke Zhu

Location: West 109

Time: Full Day (0800-1800)

Summary: 3D perception is critical in robotic applications, such as manipulation and navigation. Understanding the visual world is critical for robots to operate in the real world. In recent years, we have witnessed tremendous progress in deep learning algorithms for processing and making sense of 3D data, such as segmentation and detection. These exciting developments in 3D vision have paved the ground for tackling fundamental challenges in robot perception. Furthermore, connecting 3D vision with robotics will stimulate new research opportunities in active vision, interactive perception, and vision-based decision-making. Nonetheless, a myriad of research challenges and open questions remains. To tackle these challenges, we seek to create a shared forum for interdisciplinary researchers in 3D vision and robotics to share fresh ideas and build new connections.

Biometrics

Organizers: Bir Bhanu
Ajay Kumar

Location: West 113

Time: Full Day (0900-1730)

Summary: The burgeoning use of biometric technologies is fueling an unprecedented demand for reliable authentication and identification methods. The imperative to enhance accuracy, strengthen dependability, and broaden the scope of biometrics for diverse e-business applications is driving cutting-edge research. Biometric technologies are rapidly being implemented in large-scale infrastructure projects such as national ID programs, homeland security applications, and e-commerce solutions. Moreover, biometrics are increasingly used in social welfare programs in countries with large populations such as India, China, and the United States. As evidenced by the widespread use of fingerprint recognition and face recognition on mobile phones, biometrics are now entering the mainstream consumer market like never before. However, many promising biometric applications require accuracy levels that cannot be achieved using existing methods. To address this gap, it is imperative to make fundamental advancements in single biometric modalities as well as in the integration of multiple biometrics. This workshop aims to showcase the latest and most recent research being conducted at academic and research institutions as well as in the industry, highlighting cutting-edge developments in both single and multi-modal biometric technologies.

Computational Cameras and Displays

Organizers: He Sun
Ulugbek S. Kamilov
Salman Asif
Yi Xue

Location: West 117

Time: Full Day (0900-1700)

Summary: Computational photography has become an increasingly active area of research within the computer vision community. Within the few last years, the amount of research has grown tremendously with dozens of published papers per year in a variety of vision, optics, and graphics venues. A similar trend can be seen in the emerging field of computational displays – spurred by the widespread availability of precise optical and material fabrication technologies, the research community has begun to investigate the joint design of display optics and computational processing. Such displays are not only designed for human observers but also for computer vision applications, providing high-dimensional structured illumination that varies in space, time, angle, and the color spectrum. This workshop is designed to unite the computational camera and display communities in that it considers to what degree concepts from computational cameras can inform the design of emerging computational displays and vice versa, both focused on applications in computer vision.

The Computational Cameras and Displays (CCD) workshop series serves as an annual gathering place for researchers and practitioners who design, build, and use computational cameras, displays, and imaging systems for a wide variety of uses. The workshop solicits posters and demo submissions on all topics relating to computational imaging systems.

Neural Architecture Search & Lightweight NAS Challenge

Organizers:
Stephen McGough
Teng Xi
Elliot J. Crowley
Gang Zhang
Amir Atapour-Abarghouei
Errui Ding
Nik Khadijah Nik Aznan
Linchao Zhu
Yifan Sun
David A. Towers
Yi Yang
Jingdong Wang

Location: East 19-20

Time: Full Day (0900-1800)

Summary: Neural Architecture Search (NAS) can be successfully used to automate the design of deep neural network architectures, achieving results that outperform hand-designed models in many modern computer vision tasks. While these recent works are opening up new paths, our understanding on why these specific architectures work well, how similar are the architectures derived from different search strategies, how to design the search spaces, how to search the space in an efficient and unsupervised way, and how to fairly evaluate different auto-designed architectures remains far from complete. In this workshop we will bring together emerging research in the areas of automatic architecture search, optimization, hyperparameter optimization, data augmentation, representation learning and computer vision in order to discuss open challenges and opportunities ahead. This workshop will start with a short tutorial on NAS and its current challenges. We will also have keynotes and presentations from researchers working in the area of NAS covering latest advances and challenges for the future. One of the critiques which can be laid against NAS is that, in general, most approaches have only been developed on a core set of commonly used datasets. We have been running a competition on NAS for unseen and novel datasets. The outcome from our competition will be announced during the workshop along with presentations from the teams who achieved the best outcomes.

Visual Perception via Learning in an Open World

Organizers:
Shu Kong
Yu-Xiong Wang
Deepak Pathak
Andrew Owens
Carl Vondrick
Abhinav Shrivastava
Deva Ramanan
Terrance E. Boult

Location: West 118-120

Time: Full Day (0900-1700)

Summary: Visual perception is indispensable for numerous applications, spanning transportation, healthcare, security, commerce, entertainment, and interdisciplinary research. Visual perception algorithms developed in a closed-world setup often generalize poorly to the real open-world, which contains situations that are never-before-seen, dynamic, vast, and unpredictable. This requires visual perception algorithms to be developed for the open-world, to address its complexities such as recognizing unknown objects, debiasing imbalanced data distributions, leveraging multimodal signals, efficient few-shot learning, etc. Moreover, today's most powerful visual perception models are pretrained in an open-world, e.g., training them on web-scale data consisting of images, languages and so on. We are in the best era to study Visual Perception via Learning in an Open World (VPLOW). Therefore, we are inviting you to our VPLOW workshop, where multiple speakers and challenge competitions will cover a variety of topics of VPLOW. We hope our workshop stimulates fruitful discussions.

EarthVision: Large Scale Computer Vision for Remote Sensing Imagery

Organizers:
Ronny Hänsch
Devis Tuia
Jan Dirk Wegner
Bertrand Le Saux
Nathan Jacobs
Loic Landrieu
Charlotte Pelletier
Hannah R. Kerner
Beth Tellman

Location: West 213-214

Time: Full Day (0900-1745)

Summary: Earth Observation and remote sensing are ever-growing fields of investigation where computer vision, machine learning, and signal/image processing meet. Earth Observation covers a broad range of tasks, from detection to registration, data mining, and multi-sensor, multi-resolution, multi-temporal, and multi-modality fusion and regression, to name just a few. It is motivated by numerous applications such as location-based services, online mapping services, large-scale surveillance, 3D urban modeling, navigation systems, natural hazard forecast and response, climate change monitoring, virtual habitat modeling, food security, etc.

The full-day workshop will provide a forum for presenting original research in computer vision and pattern recognition applied to large-scale remote sensing imagery. The focus will be on recent advancements in automatic analysis of remote sensing imagery for Earth Observation and its impact on geoscience, climate change, sustainable development goals, and the general understanding of the Earth system.

Tracking and Its Many Guises: Tracking Any Object in Open-World

Organizers:
Idil Esen Zulfikar
Ali Athar
Achal Dave
Aljosa Osep
Mark Weber
Yang Liu
Sabarinath Mahadevan
Jonathon Luiten
Tarasha Khurana
Paul Voigtlaender
Pavel Tokmakov
Bastian Leibe
Deva Ramanan

Location: East 11

Time: Full Day (0900-1700)

Summary: Over the course of its rich history, object tracking has been tackled under many disguises: multi-object tracking, single-object tracking, video object segmentation, video instance segmentation, and more. Most such tasks are evaluated on benchmarks limited to a small number of common classes. Practical applicatinos require trackers that go beyond these common classes, detecting and tracking rare and even never-before-seen objects. Our workshop aims at bringing tracking to the open-world, as well as keep discussing developments in long-tail tracking. We have opened two challenges towards this end: (1) Open-World Tracking, which requires building trackers that can generalize to never-before-seen objects, and (2) Long Tail Tracking, which requires building trackers that work for rare objects, that may only contain a few examples in the training set. In addition, we have invited leading experts in the field to present their opinion on the state of the various sub-communities, and on the place of object tracking in the broader video understanding problem. The workshop will culminate in a panel discussion, during which the speakers will attempt to shine some light on the future of both long-tail and open-world tracking through their diverse perspectives.

Computer Vision in the Built Environment for the Design, Construction, and Operation of Buildings

Organizers: Iro Armeni
Martin Fischer
Yasutaka Furukawa
Daniel Hall
Fuxin Li
Michael Olsen
Marc Pollefeys
Yelda Turkan

Location: East 7

Time: Full Day (0900-1800)

Summary: The 3rd Workshop on Computer Vision in the Built Environment connects the domains of Architecture, Engineering, and Construction (AEC) with that of Computer Vision by establishing a common ground of interaction and identify shared research interests. Specifically, this workshop focuses on the as-is semantic status of built environments and the changes that take place within them over time. These topics will be presented from the dual lens of Computer Vision and AEC, highlighting the limitations and bottlenecks related to developing applications for this specific domain. The objective is for attendees to learn more about AEC and the variety of real-world problems that, if solved, could have a tangible impact on this multi-trillion-dollar industry, as well as the overall quality of life across the globe. The workshop will begin by establishing ways to capture the as-is status of a space with expert speakers both from domains. Attendees will be then introduced to the type of information required for the spatiotemporal analysis of our built environment in AEC, with a focus on effective management, safety, and the role of users in this process. Following that, the topic of scene understanding from 3D and 4D reconstructions will be presented. Finally, to close the loop from understanding to designing built environments better and faster, the topic of scene synthesis at a geometric and semantic level will be presented. We will also host the 3rd International Scan-to-BIM competition targeted on acquiring the semantic as-is status of buildings given their 3D point clouds.

New Frontiers in Visual Language Reasoning: Compositionality, Prompts and Causality

Organizers: Vicente Ordonez
Guangrun Wang
Ziliang Chen
Hao Wang
Tianlu Wang
Xiaodan Liang
Liang Lin
Alan Yuille

Location: East 9

Time: Full Day (0915-1730)

Summary: Recent years have seen the stunning powers of Visual Language Pre-training (VLP) models. Although VLPs have revolutionalized some fundamental principles of visual language reasoning (VLR), the other remaining problems prevent them from "thinking" like a human being: how to reason the world from breaking into parts (compositionality), how to achieve the generalization towards novel concepts provided a glimpse of demonstrations in context (prompts), and how to debias visual language reasoning by imagining what would have happened in the counterfactual scenarios (causality). The workshop provides the opportunity to gather researchers from different fields to review the technology trends of the three lines, to better endow VLPs with these reasoning abilities. Our workshop also consists of two multimodal reasoning challenges under the backgrounds of smart education. The challenges are practical and highly involved with our issues, therefore, shedding more insights into the new frontiers of visual language reasoning.

Continual Learning in Computer Vision

Organizers: Gido M. van de Ven
Pau Rodriguez
Vincenzo Lomonaco
Matthias De Lange
Dhireesha Kudithipudi
Xialei Liu
Rahaf Aljundi
Hava Siegelmann
Marc'Aurelio Ranzato
Hamed Hemati
Lorenzo Pellegrini

Location: East 2

Time: Full Day (0830-1730)

Summary: Incorporating new knowledge in existing models to adapt to novel problems is a fundamental challenge of computer vision. Humans and animals continuously assimilate new experiences to survive in new environments and to improve in situations already encountered in the past. Moreover, while current computer vision models must be trained with independent and identically distributed random variables, biological systems incrementally learn from non-stationary data distributions. This ability to learn from continuous streams of data, without interfering with previously acquired knowledge and exhibiting positive transfer is called Continual Learning. The CVPR Workshop on "Continual Learning in Computer Vision" (CLVision) aims to gather researchers and engineers from academia and industry to discuss the latest advances in Continual Learning. In this workshop, there are regular paper presentations, invited speakers, and a technical benchmark challenges to present the current state of the art, as well as the limitations and future directions for Continual Learning, arguably one of the most challenging milestones of AI.

Transformers for Vision

Organizers: Gedas Bertasius
Rohit Girdhar
Zhiding Yu
Jianwei Yang
Gul Varol
Lucas Beyer
Xin Wang
Feng Cheng
Yan-Bo Lin
Md Mohaiminul Islam
Jaemin Cho
Yi-Lin Sung
Lorenzo Torresani
Mohit Bansal
Jose M. Alvarez
Animashree Anandkumar
Joao Carreira

Location: East Ballroom A

Time: Full Day (0750-1730)

Summary: Over the last few years, the field of natural language processing (NLP) has been revolutionized by the emergence of transformer models. Recently, these models have also been successfully applied to various visual recognition problems such as image classification, object detection, action recognition, image/video retrieval, and many more. While many of these models achieve impressive results on their respective tasks, they also come with important technical challenges, including (1) excessive computational cost, (2) data-inefficient learning, (3) suboptimal fusion of different modalities (e.g., video, audio, speech) in multimodal settings, (4) ineffective temporal feature learning in the video domain, etc. Furthermore, the recent discoveries in this area raise many interesting questions: Are vision transformers truly better than CNNs in large-scale regimes? Will transformers replace CNNs in the future, particularly in multimodal domains? Is attention truly all you need, or is it something else? This workshop aims to bring together a diverse set of researchers who will share their latest ideas on solving the challenges of applying transformers to various visual recognition problems.

Fine-Grained Visual Categorization

Organizers: Nico Lang
Elijah Cole
Sara M. Beery
Serge Belongie
Oisin Mac Aodha
Subhransu Maji
Jong-Chyi Su
Kimberly Wilber
Srishti Yadav
Omiros Pantazis
Lukas Picek
Grant Van Horn
Suzanne Stathatos
Xiangteng He

Location: West 210

Time: Full Day (0845-1645)

Summary: Fine-grained categorization, the precise differentiation between similar plant or animal species, disease of the retina, architectural styles, etc., is an extremely challenging problem, pushing the limits of both human and machine ability. In these domains expert knowledge is typically required, and the question that must be addressed is how can we develop systems that can efficiently discriminate between large numbers of highly similar visual concepts. The 10th Workshop on Fine-Grained Visual Categorization (FGVC10) explores topics related to supervised learning, self-supervised learning, semi-supervised learning, matching, localization, domain adaptation, transfer learning, few-shot learning, machine teaching, multimodal learning (e.g., audio and video), 3D-vision, crowd-sourcing, image captioning and generation, out-of-distribution detection, open-set recognition, human-in-the-loop learning, etc., all through the lens of fine-grained understanding. Topics relevant for FGVC10 are neither restricted to vision nor categorization. FGVC10 consists of invited talks from world-renowned computer vision experts and domain experts (e.g., art), poster sessions, challenges, and peer-reviewed extended abstracts. To mark FGVC's 10th anniversary, we have confirmed five panellists for a discussion of the history and future of FGVC. We aim to stimulate debate and to expose the wider computer vision community to new challenging problems which have the potential for large societal impact but do not traditionally receive a significant amount of exposure at other CVPR workshops.

Adversarial Machine Learning on Computer Vision: Art of Robustness

Organizers: Aishan Liu
Jiakai Wang
Francesco Croce
Vikash Sehwag
Yingwei Li
Xinyun Chen
Cihang Xie
Yuanfang Guo
Xianglong Liu
Xiaochun Cao
Dawn Song
Alan Yuille
Philip Torr
Dacheng Tao

Location: East 3

Time: Full Day (0800-1600)

Summary: Deep learning has achieved significant success in multiple fields, including computer vision. However, studies in adversarial machine learning also indicate that deep learning models are highly vulnerable to adversarial examples. Extensive works have demonstrated that adversarial examples challenge the robustness of deep neural networks, which threatens deep-learning-based applications in both the digital and physical worlds. Though harmful, adversarial attacks are also beneficial for deep learning models. Discovering and harnessing adversarial examples properly could be highly beneficial across several domains including improving model robustness, diagnosing model blind spots, protecting data privacy, safety evaluation, and further understanding vision systems in practice. Since there are both the devil and angel roles of adversarial learning, exploring robustness is an art of balancing and embracing both the light and dark sides of adversarial examples. In this workshop, we aim to bring together researchers from the fields of computer vision, machine learning, and security to jointly cooperate with a series of meaningful works, lectures, and discussions. We will focus on the most recent progress and the future directions of both the positive and negative aspects of adversarial machine learning, especially in computer vision. Different from the previous workshops on adversarial machine learning, our proposed workshop aims to explore both the devil and angel characters for building trustworthy deep learning models.

Multi-Modal Learning and Applications

Organizers: Michael Ying Yang
Vittorio Murino
Bodo Rosenhahn
Pietro Morerio
Paolo Rota

Location: West 223-224

Time: Full Day (0915-1800)

Summary: The exploitation of the power of big data in the last few years led to a big step forward in many applications of Computer Vision. However, most of the tasks tackled so far are involving visual modality only, mainly due to the unbalanced number of labelled samples available among modalities, resulting in a huge gap in performance when algorithms are trained separately. Recently, a few works have started to exploit the synchronization of multimodal streams (e.g., audio/video, RGB/depth, RGB/Lidar, visual/text, text/audio, etc.) to transfer semantic information from one modality to another or by learning shared representations, reaching outstanding results. Interesting applications are also proposed in a self-supervised fashion, where multiple modalities can learn correspondences without manual labeling, resulting in a more powerful set of features as compared to those learned by processing the two modalities separately. Particular interest has been devoted to the use of language and vision, e.g., in the context of image/video generation from the text (DALL-E, text2video), audio (wav2clip), or the other way around (image2speech). This workshop aims to generate momentum around this topic of growing interest and to encourage interdisciplinary interaction and collaboration between computer vision, multimedia, remote sensing, and robotics, as well as machine learning communities, which will serve as a forum for research groups from academia and industry.

Topological, Algebraic, and Geometric Pattern Recognition With Applications

Organizers: Tegan Emerson
Henry Kvinge
Timothy Doster
Alexander Cloninger
Bastian A. Rieck
Sarah J. Tymochko

Location: East 16

Time: Full Day (0845-1730)

Summary: Topological, Algebraic, and Geometric Pattern Recognition with Applications (TAG-PRA) aims to gather researchers leveraging these three core fiels of mathematics for pattern recognition applications.

3D Scene Understanding for Vision, Graphics, and Robotics

Organizers: Siyuan Huang
Chuhang Zou
Alexander Schwing
Xiaojian Ma
Hao Su
Yixin Chen
Tengyu Liu
Yixin Zhu
David Forsyth
Derek Hoiem
Song-Chun Zhu

Location: West 220-222

Time: Full Day (0900-1730)

Summary: Tremendous efforts have been devoted to 3D scene understanding over the last decade. Due to their success, a broad range of critical applications like 3D navigation, home robotics, and virtual/augmented reality have been made possible already, or are within reach. These applications have drawn the attention and increased aspirations of researchers from the field of computer vision, computer graphics, and robotics. However, significantly more efforts are required to enable complex tasks and applications like autonomous driving or household assistant robots, where a more comprehensive understanding of the environment compared to what is possible today could be pivotal. This is due to the fact that the aforementioned tasks call for an understanding of 3D scenes across multiple levels, relying on the ability to accurately parse, reconstruct and interact with the physical 3D scene, as well as the ability to jointly recognize, reason and anticipate activities of agents within the scene. Therefore, 3D scene understanding problems become the bridge that connects vision, graphics, and robotics research. A joint effort across those fields is required to address the challenging problems. This workshop aims to foster interdisciplinary communication among researchers working on 3D scene understanding (computer vision, computer graphics, and robotics) so that more attention from the broader community can be drawn to this field. Throughout this workshop, current progress and future directions will be discussed, and new ideas and discoveries in related fields are expected to emerge.

CV4Animals: Computer Vision for Animal Behavior Tracking and Modeling

Organizers: Silvia Zuffi
Helge Rhodin
Angjoo Kanazawa
Shohei Nobuhara
Hyun Soo Park
Sara M. Beery
Anna Zamansky

Location: East 4

Time: Full Day (0920-1745)

Summary: Many biological organisms have evolved to exhibit diverse behaviors. Understanding these behaviors is a fundamental goal of multiple disciplines including neuroscience, biology, animal husbandry, ecology, and animal conservation. These analyses require objective, repeatable, and scalable measurements of animal behaviors that are not possible with existing methodologies that leverage manual encoding from animal experts and specialists. Computer vision is having an impact across multiple disciplines by providing new tools for the detection, tracking, and analysis of animal behavior. This workshop brings together experts across fields to stimulate this new field of computer-vision-based animal behavioral understanding.

Synthetic Data for Autonomous Systems

Organizers: Omar Maher
Alexander Zook
Dengxin Dai
Rareș A. Ambruș

Location: West 302-305

Time: Half Day - Morning (0750-1230)

Summary: This workshop explores challenges and opportunities in using synthetic data to enhance autonomous systems' performance in diverse environments and tasks. We aim to investigate how synthetic data can overcome current machine learning and computer vision limitations.

The half-day hybrid workshop features in-person and streamed online talks with keynotes from leading figures in academia and industry on synthetic data applications in autonomous driving and robotics.

Topics covered include synthetic data practices for embodied foundations, synthesizing humans for outdoor environments, dataset design impact on model performance, synthetic data generation pipelines, unsupervised domain adaptation, Sim2Real gap, generative AI for synthetic data, and democratization of synthetic data.

The workshop aims to share state-of-the-art knowledge and foster lively debate on overcoming current challenges in synthetic data for autonomous systems.

Advances in NeRF for the Metaverse

Organizers: Aayush Prakash
Daeil Kim
Peter Vajda
Fernando de la Torre
Angjoo Kanazawa
Jonathan T. Barron

Location: East Ballroom B

Time: Half Day - Morning (0830-1230)

Summary: A longstanding problem in computer graphics is the realistic rendering of virtual worlds. Generation of highly realistic 3D worlds at scale is an important piece of the Metaverse puzzle. However, creating such worlds and content inside it can be costly and time consuming.

In 2020, the initial work of new techniques around neural volume rendering also known as NeRF (Neural Radiance Fields) has brought an explosion of new work that has direct applicability to the future metaverse. In CVPR 2022, there were more than 50 accepted papers on NeRF improving fidelity, efficiency and scalability. We believe that NeRF is one of the most viable solutions to address the growing content needs of Metaverse.

There have been many recent advances in NeRF that have enabled it to be a strong content generation tool. Some of these advances include but are not limited to a) ability to represent arbitrary scenes including unbounded scenes at city scale, b) ability to run it on mobile devices c) higher fidelity representation of the objects/scene. This workshop is an opportunity to showcase work on NeRF that expands upon key areas that further the Metaverse development.

The aim of this workshop is to bring industry innovators and academic leaders in the world to discuss the problems, applications and in general the state of NeRF technology. Specifically, we would like to cover recent advances in NeRF that expands upon the three areas that need significant gains for the metaverse - scale, efficiency, and fidelity.

Gaze Estimation and Prediction in the Wild

Organizers: Hyung Jin Chang
Xucong Zhang
Shalini De Mello
Thabo Beeler
Seonwook Park
Otmar Hilliges
Ales Leonardis

Location: West 115

Time: Half Day - Morning (0825-1200)

Summary: Intelligent computer systems need to anticipate human intentions to provide appropriate information and efficient interactions. Eye gaze and movement patterns are the clearest indicators of human attention, with many applications for gaze tracking, such as crowd-sourced attention studies, adaptive user interfaces, AR/VR, and driver monitoring. However, low image quality and non-ideal lighting conditions can pose significant challenges outside the laboratory. Deep learning methods have been slow to address these challenges in gaze estimation due to complexity, lack of diverse datasets, and small community size. The GAZE workshop series, including GAZE2019 (at ICCV2019), GAZE2020 (at ECCV2020), GAZE2021 (at CVPR2021), and GAZE2022 (at CVPR2022), have brought together academia and industry to share research achievements and discuss future directions. The upcoming 5th GAZE workshop at CVPR2023 aims to encourage novel strategies for eye gaze estimation and prediction, with a focus on synthetic eye gaze dataset generation and various applications, including VR/AR and driver monitoring.

Large Scale Holistic Video Understanding

Organizers: Vivek Sharma
Ali Diba
Shyamal Buch
Mohsen Fayyaz
Ehsan Adeli
Luc Van Gool
Jürgen Gall
Rainer Stiefelhagen
David A. Ross
Manohar Paluri

Location: East 8

Time: Half Day - Morning (0830-1200)

Summary: The capabilities of computer systems to classify video from the Internet or analyze human actions in videos have improved tremendously in recent years. Lots of work has been done in the video recognition field on specific video understanding tasks, such as action recognition and scene recognition. Despite substantial achievements in these tasks, the holistic video understanding task has not received enough attention. Currently, video understanding systems specialize in specific fields. For real-world applications, such as analyzing multiple concepts in a video for video search engines and media monitoring systems, or defining a humanoid robot's surrounding environment, a combination of current state-of-the-art methods is necessary. Toward a holistic video understanding (HVU), we present this workshop. Recognizing scenes, objects, actions, attributes, and events in real-world videos is the focus of this challenge. To address such tasks, we introduce our HVU dataset, which is organized hierarchically according to a semantic taxonomy of holistic video understanding. Many real-world conditioned video datasets target human action or sport recognition. Our newly created dataset can help the vision community and attract more attention to the possibility of developing more interesting holistic video understanding solutions. Our workshop will bring together ideas related to multi-label and multi-task recognition in real-world videos. Our dataset will be used to test research efforts.

Ethical Considerations in Creative Applications of Computer Vision

Organizers: Negar Rostamzadeh
Rida Qadri
Mohammad Havaei

Location: West 103-104

Time: Half Day - Morning (0800-1210)

Summary: Creative domains constitute a big part of modern society, having a strong influence on the economy and social life. Computer vision technologies are rapidly being integrated into these domains to, for example, aid in artistic content retrieval and curation, generate synthetic media, or enable new forms of artistic methods and creations. However, creative AI technologies bring with them a host of ethical concerns, ranging from representational harms associated with data augmentation, generation, and analysis of culturally sensitive content to copyright and ownership concerns. This workshop is built on the success and past experiences of the creative computer vision community (Computer Vision for Fashion, Art and Design) workshop series at ECCV 2018, ICCV 2019, and CVPR 2020, and Creative AI workshop series at NeurIPS 2017-2021, as well as the expertise of the Ethical AI scientists, traditional artists, and generative artists. This is the second workshop on "Ethical Considerations in Creative applications of Computer Vision" , and built on the experiences we obtained from organizing the first edition of the workshop.

By proposing this workshop, our aim is to create a platform for interdisciplinary discussions among computer vision researchers, sociotechnical researchers, policy makers, social scientists, artists, and other cultural stakeholders.

Visual Anomaly and Novelty Detection

Organizers: Thomas Brox
Toby P. Breckon
Philipp Seeböck
Paul Bergmann
Latha Pemula

Location: East 13

Time: Half Day - Morning (0830-1230)

Summary: Anomaly detection, and the synonymous topics of novelty and out-of-distribution detection, represent an important and application-relevant challenge within both computer vision and the broader field of pattern recognition. In its simplest formulation, anomaly detection targets the identification of samples which deviate from an obtained approximation to the true distribution of normality for a given dataset. As such anomalies represent unexpected eventualities or outliers in the scope of a given task. The notion of detecting them effectively and efficiently has been sought after for many real-world applications including medical diagnosis, airport security screening, industrial inspection, or crowd control.

We now see the rise of a complex and vibrant set of learning-based paradigms addressing the anomaly detection task - varying across both the fully/semi/un-supervised and few/one/zero shot axes of recent computer vision and pattern recognition research. This workshop brings together researchers of both industry and academia to present and discuss recent developments, opportunities and open challenges in this area. The workshop will also host a challenge for zero-/few-normal-shot anomaly detection, to encourage the development and benchmarking new algorithms for realistic yet challenging tasks.

Catch UAVs That Want to Watch You: Detection and Tracking of Unmanned Aerial Vehicle in the Wild and Anti-UAV Challenge

Organizers: Jian Zhao
Jianan Li
Lei Jin
Jiaming Chu
Zhihao Zhang
Jun Wang
Jianqiang Xia
Kai Wang
Yang Liu
Sadaf Gulshad
Jiaojiao Zhao
Zheng Zhu
Tianyang Xu
Xuefeng Zhu
Shihan Liu
Guibo Zhu
Zechao Li
Zheng Wang
Baigui Sun
Yandong Guo
Shin'ichi Satoh
Junliang Xing
Jane Shen Shengmei

Location: West 121-122

Time: Half Day - Morning (0830-1210)

Summary: Civil unmanned aerial vehicles (UAVs), a.k.a. drones, have been widely used in a broad range of civil application domains. Nevertheless, we should be aware of the potential threat to our lives caused by UAV intrusion since UAVs can also be used to conduct physical attacks (e.g., via explosives) and cyber-attacks (e.g., hacking critical infrastructure). Moreover, unauthorized UAVs sometimes violate aviation safety regulations, thereby bringing hazards to civilian aircraft and passengers and even causing airport disruptions and flight delays. There have been multiple instances of drone sightings halted air traffic at airports, leading to significant economic losses for airlines. It is highly desired to develop anti-UAV techniques to defend against drone accidents. Historically, radar is certainly a compelling technology for detecting traditional incoming airborne threats. However, these comparatively small UAVs are extremely difficult for radar to see because they have very small radar cross-sections, low flight altitudes, and erratic flight paths. Therefore, how to use computer vision and machine learning algorithms to perceive UAVs is a crucial part of the whole UAV-defense system.

The workshop encourages participants to develop automated methods that can detect and track UAVs in thermal infrared videos with high accuracy. Particularly, algorithms that can detect and track fast-moving drones in complex environments are highly expected.

Mobile Intelligent Photography and Imaging

Organizers: Chongyi Li
Shangchen Zhou
Ruicheng Feng
Yuekun Dai
Pengfei Zhu
Qianhui Sun
Chen Change Loy
Wenxiu Sun
Jinwei Gu

Location: East 10

Time: Half Day - Morning (0800-1200)

Summary: Developing and integrating advanced image sensors with novel algorithms in camera systems is prevalent with the increasing demand for computational photography and imaging on mobile platforms. However, the lack of high-quality data for research and the rare opportunity for in-depth exchange of views from industry and academia constrain the development of mobile intelligent photography and imaging (MIPI). The workshop's focus is on MIPI, emphasizing the integration of novel image sensors and imaging algorithms. Together with the workshop, we organize a few exciting challenges and invite renowned researchers from both industry and academia to share their insights and recent work.

Structural and Compositional Learning on 3D Data

Organizers: Kaichun Mo
Kai Wang
Marios Loizou
Despoina Paschalidou
Paul Guerrero
Minhyuk Sung
Melinos Averkiou
Shuran Song
Evangelos Kalogerakis
Luca Carlone
Hao Zhang
Leonidas Guibas

Location: West 205-206

Time: Half Day - Morning (0750-1230)

Summary: Dealing with the huge diversity and complexity of 3D data has become the main research challenge for various applications in computer vision, graphics, and robotics. One key approach that researchers have found promising is to decompose the complex 3D data into smaller and easier composable subcomponents. Scene graphs of objects, part decompositions of 3D objects, and the primitive actions and sub-skills for robotics are a few characteristic examples. Unlike traditional connectionist approaches in deep learning, structural and compositional learning includes components that lean more towards the symbolic end of the spectrum, which leads to many challenging open research questions about how to represent the composable sub-units and how to conduct efficient learning over them. In this workshop series, we aim to bring together researchers from diverse fields and backgrounds to share ideas and jointly discuss structural and compositional learning of 3D data. In conjunction to the workshop, we will host a brand-new challenge BuildingNet on an important task of performing structural decomposition of 3D building shapes.

OmniLabel: Infinite Label Spaces for Semantic Understanding via Natural Language

Organizers: Samuel Schulter
Vijay Kumar B. G.
Yumin Suh
Golnaz Ghiasi
Long Zhao
Qi Wu
Dimitris N. Metaxas

Location: West 207

Time: Half Day - Morning (0800-1200)

Summary: The goal of this workshop is to foster research on the next generation of visual perception systems that reason over label spaces that go beyond a list of simple category names. Modern applications of computer vision require systems that understand a full spectrum of labels, from plain category names ("person" or "cat"), over modifying descriptions using attributes, actions, functions or relations ("women with yellow handbag" , "parked cars" , or "edible item"), to specific referring descriptions ("the man in the white hat walking next to the fire hydrant"). Natural language is a promising direction not only to enable such complex label spaces, but also to train such models from multiple datasets with different, and potentially conflicting, label spaces. Besides an excellent list of invited speakers from both academia and industry, the workshop will present the results of the OmniLabel challenge, which we held with our newly collected benchmark dataset that subsumes generic object detection, open-vocabulary detection, and referring expression comprehension into one unified and challenging task.

Deep Learning in Ultrasound Image Analysis

Organizers: Mengliu Zhao, Amir Ghasemi,
Mike Wong, Zahra Mirikharaji,
Gareth Munro, Jason Vantomme,
Gaurav Handa, Reza Zahiri,
Carlos Alberto da Costa Filho

Location: West 114

Time: Half Day - Morning (0820-1200)

Summary: Ultrasound has become one of the most common imaging modalities in sonar, biomedical imaging, and non-destructive testing (NDT) in the past two decades. While traditional techniques remain useful, modern ultrasound devices can easily collect vast quantities of high-resolution data quickly, making them approachable by deep learning methods. Training datasets for these deep learning methods may contain unique challenges, including extreme data imbalance or the need for multi-task, weakly-supervised and semi-supervised learning. In addition, gaps remain between natural image-derived deep learning algorithms and those for ultrasonic acoustic-derived images, including focused image denoising, image interpretation, uncertainty quantification, and automated system self-awareness. In the last few years, the medical ultrasound field has witnessed the successful application of deep learning in both 2D and 3D to enhance, identify, and significantly speed up the analysis process. However, in the field of NDT, ultrasound analysis comes with its unique challenges. This workshop aims to bring together researchers and experts with biomedical, NDT, and computer vision to explore the future of deep learning and ultrasound image analysis.

New Frontiers for Zero-Shot Image Captioning Evaluation

Organizers: Kyoung Mu Lee, Mark Marsden,
Seung Hwan Kim, Sihaeng Lee,
Alessandra Sala, Pyunghwan Ahn,
Bohyung Han, Sangyun Kim,
Taehoon Kim

Location: West 116

Time: Half Day - Morning (0800-1200)

Summary: The CVPR 2023 Workshop New frontiers for zero-shot Image Captioning Evaluation (NICE) aims to challenge the computer vision community to develop robust image captioning models that advance the state-of-the-art both in terms of accuracy and fairness. Despite recent improvements in large-scale image-text datasets and vision-language models, the current challenges used by the academic community are not sufficient to test the true limits of zero-shot image captioning models. New evaluation datasets are required which contain a larger variety of visual concepts from many domains as well as various image types. Concurrently, as large-scale image-text datasets used to train captioning models have been shown to propagate societal biases, there is an additional need to develop evaluation datasets and methods to identify racial and gender bias in AI generated image captions. Our workshop also includes an open challenge on zero-shot image captioning. Shutterstock provides a new dataset which is open sourced to this community. The new large-scale dataset consists of roughly 26k high quality images with associated curated metadata and it covers more than 20 general categories and a wide breadth of concepts. With this dataset we expect the community to take a longitudinal evaluation across a variety of metrics to comparatively assess performance of different zero-shot image captioning models. In the workshop, we share the results of the challenge and technical contributions of the top-ranking entries.

Monocular Depth Estimation Challenge

Organizers: Jaime Spencer, Wendy Adams,
C. Stella Qian, Andrew J. Schofield,
Chris Russell, James Elder,
Simon Hadfield, Richard Bowden,
Erich Graf

Location: West 208-209

Time: Half Day - Morning (0830-1200)

Summary: Depth estimation is crucial for human perception and daily navigation. Humans rely on stereo vision and motion parallax to estimate depth in their near surroundings. However, these cues become weaker as depth increases. As a result, humans rely profoundly on monocular cues when estimating depth in the far range. Furthermore, humans can perceive depth from purely monocular information, such as paintings, photos, and videos.

Computer vision algorithms for MDE have advanced substantially over the past few years, leveraging large quantities of unlabeled automotive video data. Meanwhile, the benchmarking procedure for these algorithms has remained largely unchanged, relying on simple metrics and sparse LiDAR data. This does not provide detailed insights into the performance of each method, especially if the ground-truth is incorrect.

This workshop will address this problem by providing a carefully-curated human depth perception benchmark on a variety of natural scenes. This will evaluate MDE outside the common automotive domain, testing the generalization to varied real-world scenes and validated using well-established image-/pointcloud-/edge-based metrics, as well as human benchmarks.

The workshop consists of two parts: invited talks discussing current developments in MDE & its evaluation and a challenge organized around a novel benchmarking procedure using the SYNS dataset. The invited keynote speakers are Oisin Mac Aodha, Daniel Cremers and Alex Kendall.

Long-Form Video Understanding and Generation

Organizers: Mike Zheng Shou, Weiyao Wang,
Linchao Zhu, Xiaohan Wang,
Stan Weixian Lei, Hehe Fan,
Difei Gao, Kristen Grauman,
Joya Chen, Matt Feiszli,
Dongxing Mao, Lorenzo Torresani,
Weijia Wu, Karttikeya Mangalam,
Jitendra Malik

Location: West 111-112

Time: Half Day - Morning (0855-1235)

Summary: See workshop's webpage for a description.

Computer Vision for Mixed Reality

Organizers: Rakesh Ranjan
Peter Vajda
Laura Leal-Taixé
Xiaoyu Xiang

Location: West 301

Time: Half Day - Morning (0900-1230)

Summary: VR technologies have the potential to transform the way we use computing to interact with our environment, do our work and connect with each other. VR devices provide users with immersive experiences at the cost of blocking the visibility of the surrounding environment. With the advent of passthrough techniques such as those in Quest Pro, now users can build deeply immersive experiences which mix the virtual and the real world into one, often also called Mixed Reality. Mixed Reality poses a set of very unique research problems in computer vision that are not covered by VR. Our focus is on capturing the real environment around the user using cameras which are placed away from the user's eyes, yet reconstruct the environment with high fidelity, augment the environment with virtual objects and effects, and all in real-time. This would offer the research community to deeply understand the unique challenges of Mixed Reality and research on novel methods encompassing View Synthesis, Scene Understanding, efficient On-Device AI among other things.

Notes:

Quantum Computer Vision and Machine Learning

Organizers: Tolga Birdal, Tongyang Li
Vladislav Golyanik, Jacob Biamonte
Martin Danelljan, Jan-Nico Zaech

Location: West 202-204

Time: Half Day - Afternoon (1300-1800)

Summary: The goal of this workshop is to introduce quantum computation to the realm of computer vision and foster the formation of a community. A concrete summary of the aims are as follows:

- Identify computer vision problems that can be addressed by quantum computers.
- Showcase recent and ongoing progress towards practical quantum computing and computer vision.
- Address and discuss the current state-of-the-art, limitations therein, expected progress and its impact on the computer vision world.
- Enlighten the community to attract further researchers in this direction.

Focal points for discussions and talks include but are not limited to:

- Premises of quantum computation.
- Use of the techniques from quantum mechanics in solving CVML problems, classically.
- Adiabatic quantum computation and use cases in CVML.
- Circuit based quantum computers and their use in CVML.
- Tensor methods in QCVML.
- Review of the upcoming software for programming QC.

Capturing, Interpreting & Visualizing Indoor Living Spaces

Organizers: Naji Khosravan, Enrique Dunn
Ehsan Adeli, Hamid Rezatofighi
Ivaylo Boyadzhiev, Amir Zamir
Chen Chen, Huangying Zhan

Location: East 12

Time: Half Day - Afternoon (Time TBA)

Summary: Motivated by the recent release of datasets such as Zillow Indoor Dataset (ZInD), Apple's ARKit Scenes dataset and Facebook's Habitat-Matterport dataset, in this workshop we would like to bring industry and academia together and encourage both to focus on specific under explored aspects of environment understanding. We welcome innovation in 3 main areas: 1) Data: Ranging from new datasets to new attributes and information being extracted from the current datasets, 2) Modeling: CV/ML models/algorithms to solve one or multiple tasks related to indoor environment understanding. 3) Graphics and Visualization: New reconstruction, lighting, virtual staging/object insertion, etc. We encourage researchers to go beyond "scene understanding" and explore "environment understanding" with a focus on understanding structure through tasks such as 2D/3D room layout estimation, understanding relation of "rooms" for floorplan generation, localization of media within rooms and floorplans, localization of objects within rooms and floorplans. Image, geometric, and semantic information can also be used to reimagine the appearance of home interiors in a photorealistic manner.

Vision Datasets Understanding

Organizers: Fatemeh Saleh, Qiuhong Ke, Liang Zheng, Manmohan Chandraker, Qiang Qiu, Xiaoxiao Sun, Jose Lezama, Yang Yang, Peter Koniusz

Location: West 211

Time: Half Day - Afternoon (1330-1730)

Summary: Data is the fuel of computer vision, on which the state-of-the-art systems are built. A robust object detection system not only needs a strong model architecture and learning algorithms, but also relies on a comprehensive large-scale training set. Despite the pivotal significance of datasets, existing research in computer vision is usually algorithm centric. That is, given fixed training and test data, it is the algorithms or models that are primarily considered for improving. As such, while significant progress has been made in understanding and improving algorithms, there is much less effort in the community made on dataset-level analysis. For example, comparing the number of algorithm-centric works in domain adaptation, the quantitative understanding of the domain gap is much more limited. To further this campaign, this workshop brings together research works and discussions from the dataset perspective and holds a competition on test set difficulty analysis without ground truths.

Pixel-Level Video Understanding in the Wild Challenge

Organizers: Jiaxu Miao, Si Liu, Yunchao Wei, Yi Zhu, Zongxin Yang, Elisa Ricci, Yi Yang, Cees Snoek

Location: West 301

Time: Half Day - Afternoon (1330-1730)

Summary: The CVPR 2023 Workshop will focus on the pixel-level video understanding including video semantic/instance/panoptic segmentation

Face and Gesture Analysis for Health Informatics

Organizers: Zakia Hammal, Mohamed Daoudi

Location: West 212

Time: Half Day - Afternoon (1330-1800)

Summary: Within the past 10 years great strides have been made in the computer vision and machine learning community, as well as sensing technology for the modeling, analysis and synthesis of human verbal and nonverbal behavior for healthcare related applications. For instance, on-board smartphone sensors and wearable devices that track user activity, sleeping and eating habits, blood pressure, heart rate, skin temperature, and movement. However, compared to the advances in sensing technology, the current advances in computer vision and machine learning for verbal and nonverbal analysis has not yet achieved the goal of moving from the laboratory to the real-world healthcare context (e.g., medical setting). Recent advances in computer vision and machine learning for automatic analysis and modeling of human behavior could be used to reliably and objectively measure the physical, mental and social wellness beyond the classical definition of health assessment. The workshop aims to gather researchers working in different domains (from low-level sensing for face, head, and body detection to high-level modeling of complex social and clinically relevant behavior) to discuss the strengths and major challenges in using computer vision and machine learning of automatic modeling of face and gesture for clinical research and healthcare applications

Compositional 3D Vision & 3DCoMPaT Challenge

Organizers: Mohamed Elhoseiny, Yuchen Li, Peter Vajda, Peter Wonka, Natalia Neverova, Habib Slim, Wolfgang Heidrich, Xiang Li

Location: West 205-206

Time: Half Day - Afternoon (1245-1745)

Summary: The C3DV workshop is devoted to the exploration of compositional 3D vision, with a particular focus on recognizing and grounding compositions of materials on parts of 3D objects. The workshop invites research contributions that cover a wide range of topics, including deep learning methods for compositional 3D vision, self-supervised learning, visual relationship detection, zero-shot recognition/detection of compositional 3D visual concepts, novel problems in 3D vision and compositionality, text/composition to 3D generation, text/composition-based editing of 3D scenes/objects, language-guided 3D visual understanding (objects, relationships, ...), transfer learning for compositional 3D Vision, multimodal pre-training for 3D understanding, and other related topics. The workshop offers an inclusive platform for researchers to present and engage in meaningful discussions on various topics related to compositional 3D vision. This is facilitated through keynote presentations, as well as poster presentations, fostering a collaborative environment for the exchange of ideas and insights.

Accessibility, Vision, and Autonomy Meet

Organizers: Eshed Ohn-Bar, Chieko Asakawa, Danna Gurari, Hernisa Kacorri, Kris Kitani

Location: West 111-112

Time: Half Day - Afternoon (1300-1730)

Summary: The goal of this workshop is to gather researchers, students, and advocates who work at the intersection of accessibility, computer vision, and autonomous and intelligent systems. In particular, we plan to use the workshop to identify challenges and pursue solutions for the current lack of shared and principled development tools for vision-based accessibility systems. For instance, there is a general lack of vision-based benchmarks and methods relevant to accessibility (e.g., people using mobility aids are currently mostly absent from large-scale datasets in pedestrian detection). Towards building a community of accessibility-oriented research in computer vision conferences, we also introduce a large-scale fine-grained computer vision challenge. The challenge involves visual recognition tasks relevant to individuals with disabilities. We aim to use the challenge to uncover research opportunities and spark the interest of computer vision and AI researchers working on more robust and broadly usable visual reasoning models in the future. An interdisciplinary panel of speakers will further provide an opportunity for fostering a mutual discussion between accessibility, computer vision, and robotics researchers and practitioners.

End-to-End Autonomous Driving: Perception, Prediction, Planning and Simulation

Organizers: Li Zhang, Jiachen Lu
Xiatian Zhu, Andreas Geiger
Anurag Arnab, Philip Torr
Fatma Güney

Location: East Exhibit Hall A

Time: Half Day - Afternoon (1230-1905)

Summary: A diversity of computer vision capabilities are all critical in building industry-level autonomous driving systems, ranging from 2D to 3D perception, prediction, planning, to scene simulation. This has inspired a surge of relevant research, growing at a fast pace with increasingly accurate and efficient new methods (e.g., BEV-based 3D detection, HDMapNet, NeRF) developed continuously. Much more than simple combination of individual independently developed methods, autonomous driving also requires synergistic integration of different functions as a whole. This however is far away from the current situation that researchers in the sub-fields of perception, planning and simulation make largely limited idea exchange and communication. This calls for a system-level perspective on the advancement of autonomous driving. This workshop aims to provide a platform where researchers from different sub-fields can focus on exchanging the frontier ideas across boundaries, leading to holistic system-aware understanding and systematic research attempts in the future.

4D Hand Object Interaction: Geometric Understanding and Applications in Dexterous Manipulation

Organizers: Li Yi, Sifei Liu
He Wang, Yunze Liu
Xiaolong Wang, Hao Su
Yu-Wei Chao, Shubham Tulsiani
David Fouhey

Location: East 13

Time: Half Day - Afternoon (1300-1800)

Summary: Hand-object interactions (HOI) feature regularly in our daily activities in which we use one hand or two to manipulate objects directly or to use various tools. Through HOI, humans excel in manipulation tasks, master new skills, and adapt to complex and continuously changing environments. Recently, there is a surge of interest in understanding and generating hand-object interaction, to support applications in robotics, augmented reality, and other important fields. Such applications usually require a detailed understanding of hands dynamically interacting with a wide range of objects in 4D (3D space + 1D time), and struggle to synthesize or execute dexterous manipulations in 4D at the same level of skill as humans. The key goal of this workshop is to assemble experts in 4D vision, hand-object interaction, dexterous manipulation, and animation synthesis to synchronize and coordinate the efforts. First, the workshop would help researchers working on 4D HOI understanding to know about dexterous manipulation and tailor their research topics toward interaction-oriented 4D understanding. Second, embodied AI researchers could improve their understanding of the limitation of current 4D HOI perception and design their methods properly. Finally, in the workshop we will also host a competition designed for benchmarking the progress of 4D HOI geometric understanding, to invite more researchers to the field.

Visual Odometry and Computer Vision Applications Based on Location Clues

Organizers: Guoyu Lu, Nicu Sebe
Friedrich Fraundorfer, Chandra Kambhamettu
Yan Yan

Location: West 302-305

Time: Half Day - Afternoon (1245-1830)

Summary: The workshop aims to gather researchers working in different domains (from low-level sensing for face, head, and body detection to high-level modeling of complex social and clinically relevant behavior) to discuss the strengths and major challenges in using computer vision and machine learning of automatic modeling of face and gesture for clinical research and healthcare applications. We invite scientists working in related areas of computer vision and machine learning for face and gesture detection, affective computing, multimodal human behavior modeling, and cognitive behavior to share their expertise and achievements in the emerging field of computer vision and machine learning based face and gesture analysis for health informatics.

High-Fidelity Neural Actors

Organizers: Markos Georgopoulos, Lourdes Agapito
Martin Rünz, Matthias Niessner
Jon Starck

Location: West 121-122

Time: Half Day - Afternoon (1300-1800)

Summary: Reconstructing and animating clothed humans is a research area of increasing academic and industrial interest due to the plethora of applications, such as augmented and virtual reality, that facilitate telepresence in the metaverse. However, representing the spatio-temporal surface dynamics in clothed humans poses a significant challenge, that usually requires manual intervention (e.g., 3D artists) or computationally expensive physics simulations. Thus, synthesising photo-realistic, controllable avatars remains an open problem. To this end, we host a workshop on high-fidelity neural actors, that will bring together experts in the field of neural rendering and digital humans, with the aim to discuss and facilitate progress in the field.

Computer Vision for Fashion, Art, and Design

Organizers: Julia Lasserre, Nour Karessli
Leonidas Lefakis, Reza Shirvany
Loris Bazzani, Ziad Al-Halah
Mariya Vasileva

Location: East Ballroom B

Time: Half Day - Afternoon (1300-1730)

Summary: Creative domains render a big part of modern society, having a strong influence on the economy and cultural life. Much effort within creative domains, such as fashion, art and design, center around the creation, consumption, manipulation and analytics of visual content. In recent years, there has been an explosion of research in applying machine learning and computer vision algorithms to various aspects of the creative domains. The CVFAD workshop series aims to capture important trends and new ideas in this area. At CVPR 2023, CVFAD will continue to bring together artists, designers, and computer vision researchers and engineers, creating a space for conversations and idea exchanges at the intersection of computer vision and creative applications.

Light Fields for Computer Vision: New Applications and Trends in Light Fields

Organizers: Hao Sheng
Yebin Liu
Jingyi Yu
Gaochang Wu

Location: West 215-216

Time: Half Day - Afternoon (1330-1815)

Summary: 4D Light fields can capture both intensity and directions of light rays, and record 3D geometry in a convenient and efficient manner. In the past few years, various areas of research are trying to use light fields to obtain superior performance internal structure information. Light fields have been widely used with remarkable results in some applications like depth estimation, super-resolution and so on. While the attempts in other applications like object detection and semantic segmentation are still in preliminary stage due to the lack of corresponding datasets, and incompatibility between redundant context information and limited memory. Meanwhile, more and more novel and powerful technologies like Neural Radiance Fields and Multiplane Image have been introduced into computer vision, there will be plenty of opportunities and challenges to incorporate them with light fields. To this end, this workshop focuses on two brand new topics. The first is to introduce the light field into more application areas, break through the bottleneck between rich structural information and limited memory, and achieve stable performance. The second is to explore how to introduce emerging technologies from other research fields into light fields to create new technological effects and drive competition. Besides, this workshop also hosts competitions about light field semantic segmentation and depth estimation to invite more researchers to the field.

Precognition: Seeing Through the Future

Organizers: Khoa Luu, Utsav Prabhu
Nemanja Djuric, Hien Van Nguyen
Kris Kitani, Junwei Liang

Location: West 207

Time: Half Day - Afternoon (1300-1700)

Summary: Vision-based detection and recognition studies have been recently achieving highly accurate performance and were able to bridge the gap between research and real-world applications. Beyond these well-explored detection and recognition capabilities of modern algorithms, vision-based forecasting will likely be one of the next big research topics in the field of computer vision. Vision-based prediction is one of the critical capabilities of humans, and the potential success of automatic vision-based forecasting will empower and unlock human-like capabilities in machines and robots.

This workshop aims to facilitate further discussion and interest within the research community regarding this nascent topic. We will discuss recent approaches and research trends not only in anticipating human behavior from videos but also precognition in multiple other visual applications, such as medical imaging, healthcare, human face aging prediction, early even prediction, autonomous driving forecasting, etc.

Notes:

Monday, June 19

NOTE: Tutorial rooms are subject to change. Refer to the online site for up-to-date locations. Use the QR code for each tutorial to see its schedule. Here is the QR code for the CVPR 2023 Tutorials page.

0700–1700 Registration (West Ballroom Foyer)

0700–0900 Breakfast (West Ballrooms A–D)

1000–1045 Morning Break West Ballrooms A–D

1145–1345 Lunch (West Ballrooms A–D)

1500–1545 Afternoon Break (West Ballrooms A–D)

Tutorial: All You Need to Know About Self-Driving

Organizers:	Raquel Urtasun	Sivabalan Manivasagam
	Sergio Casas	Paul Spriesterbach
	Abbas Sadat	Andrei Barsan
Location:	West 302-305	
Time:	Full Day (0900-1800)	

Summary: A full day tutorial covering all aspects of autonomous driving. This tutorial will provide the necessary background for understanding the different tasks and associated challenges, the different sensors and data sources one can use and how to exploit them, as well as how to formulate the relevant algorithmic problems such that efficient learning and inference is possible. We will first introduce the self-driving problem setting and a broad range of existing solutions, both top-down from a high-level perspective, as well as bottom-up from technological and algorithmic points of view. We will then extrapolate from the state of the art and discuss where the challenges and open problems are, and where we need to head towards to provide a scalable, safe and affordable self-driving solution for the future.

Notes:

Tutorial: Large-Scale Visual Localization

Organizers:	Torsten Sattler	Marc Pollefeys
	Yannis Avrithis	Sudipta Sinha
	Eric Brachmann	Giorgos Tolias
	Zuzana Kukelova	
Location:	East 2	
Time:	Half Day - Morning (0830-1215)	

Summary: The tutorial covers the task of visual localization, i.e., the problem of estimating the position and orientation from which a given image was taken. The tutorial's scope includes cases with different spatial/geographical extent, small indoor/outdoor scenes, city-level, and world-level, and localization under changing conditions. In the coarse localization regime, the task is typically handled via retrieval approaches, which is covered in the first part of the tutorial. A typical use case is the following: Given a database of geo-tagged images, the goal is to determine the place depicted in a new query image. Traditionally, this problem is solved by transferring the geo-tag of the most similar database image to the query. The major focus of this part is on the visual representation models used for retrieval, where we include both classical feature-based and recent deep learning-based approaches. The 2nd and 3rd part of the tutorial encompass methods for precise localization with features-based and deep learning approaches, respectively. A typical use-case for these algorithms is to estimate the full 6 Degree-of-Freedom (6DOF) pose of a query image, i.e., the position and orientation from which the image was taken, for applications such as robotics, autonomous vehicles (self-driving cars), Augmented / Mixed / Virtual Reality, loop closure detection in SLAM, and Structure-from-Motion. The final part will cover existing datasets, including their limitations. We provide links to publicly available source code for the discussed approaches.

Tutorial: Hyperbolic Deep Learning in Computer Vision

Organizers:	Pascal Mettes
	Max van Spengler
	Yunhui Guo
	Stella Yu
Location:	West 116-117
Time:	Half Day - Morning (Time TBA)

Summary: Learning in computer vision is all about deep networks and such networks operate on Euclidean manifolds by default. While Euclidean space is an intuitive and practical choice, foundational work on non-visual data has shown that when information is hierarchical in nature, hyperbolic space is superior, as it allows for an embedding without distortion. A core reason is because Euclidean distances scale linearly as a function of their norm, while hyperbolic distances grow exponentially, just like hierarchies grow exponentially with depth. This initial finding has resulted in rapid developments in hyperbolic geometry for deep learning.

Hyperbolic deep learning is booming in computer vision, with new theoretical and empirical advances with every new conference. But what is hyperbolic geometry exactly? What is its potential for computer vision? And how can we perform hyperbolic deep learning in practice? This tutorial will cover all such questions. We will dive into the geometry itself, how to design networks in hyperbolic space, and we show how current literature profits from learning in this space. The aim is to provide technical depth while addressing a broad audience of computer vision researchers and enthusiasts.

Tutorial: Reverse Engineering of Deception: Foundations and Applications

Organizers: Sijia Liu
Xiaoming Liu
Xue Lin

Location: East 7

Time: Half Day - Morning (0900-1200)

Summary: This tutorial will deliver a well-rounded understanding of the emerging field of reverse engineering of deception (RED) techniques, a cutting-edge topic in adversarial machine learning (ML) for reliable computer vision (CV). Past studies have extensively explored the generation, detection, and defense of machine-centric deception (e.g., adversarial attacks that deceive ML models) and human-centric deception (e.g., GAN-created images that mislead human decision-making) in CV. However, RED introduces a new adversarial learning paradigm that automatically uncovers and catalogs attack "fingerprints" found in both machine and human-centric attacks. The RED problem addressed in the tutorial is: Can we reverse-engineer the adversary's knowledge and attack toolchains beyond conventional adversarial detection/defense techniques? To this end, this tutorial will cover the following key aspects: (1) Review RED's definition and formulation, addressing basics and preliminaries. (2) Discuss the challenges and significance of RED, highlighting its connections and differences with conventional adversarial detection/defense techniques in ML. (3) Explore RED for machine-centric adversaries, reviewing recent RED developments on top of a variety of adversarial attacks. (4) Examine RED for human-centric adversaries, reviewing RED methods for the detection and model parsing of GAN-generated fake images. (5) Demonstrate and showcase RED applications in CV.

Tutorial: Object Localization for Free: Going Beyond Self-Supervised Learning

Organizers: Oriane Simeoni
Weidi Xie
Thomas Kipf
Patrick Perez

Location: East 11

Time: Half Day - Morning (0830-1200)

Summary: Object localization in images is a key problem in a wide range of application domains that are embedded in critical settings such as self-driving vehicles or healthcare. However, most efficient solutions able to perform an object localization task follow the standard object detection and semantic segmentation frameworks, meaning that they require large amounts of annotated data for training. Different heuristics and tools can now assist and enhance human annotators, however manual annotation remains a largely heavy and expensive process. Moreover, perception models based on annotations enter a dependence circle of additional annotations for every new object class to detect or new external conditions to cover, e.g. in/outdoor, different times of the day, weathers. Such models struggle in dealing with our open complex world that is evolving continuously. Recent works have shown exciting prospects of avoiding annotations altogether by (1) leveraging self-supervised features, (2) building self-supervised object-centric objectives and (3) combining different modalities. In this context, we propose a half-day tutorial in which we will provide an in-depth coverage of different angles on performing/building-upon object localization with no human supervision.

Tutorial: Polarization-Based Computer Vision

Organizers: Jinwei Ye
Seung-Hwan Baek
Achuta Kadambi
Huaijin Chen

Location: East 19-20

Time: Half Day - Morning (0900-1200)

Summary: Polarization is a fundamental property of light and describes the direction in which the electric field of light oscillates. Polarization, as an intrinsic property of light, provides an extra dimension of information for probing the physical world. Although polarization is often overlooked, it allows for efficient geometry and material analysis beyond the conventional color images. With the snapshot quad-Bayer polarization camera being commercialized, there have been growing interests in using polarization cues to solve a wide range of computer vision problems. Recent advances have demonstrated advantages of using polarization imaging for geometry and material understanding.

In this tutorial, we will cover comprehensive topics in polarization imaging, from the fundamental physical principles to its applications in various computer vision problems. We will specifically focus on recent advances on using polarization imaging for solving the problems of reflectance modeling, 3D reconstruction, and transparent object segmentation. Finally, we will showcase applications of polarization imaging in industry settings.

Tutorial: Recent Advances in Vision Foundation Models

Organizers: Linjie Li
Zhe Gan
Chunyuan Li
Jianwei Yang

Location: East 16

Time: Half Day - Morning (0830-1230)

Summary: Visual understanding at different levels of granularity has been a longstanding problem in the computer vision community. The tasks span from image-level tasks (e.g., image classification, image-text retrieval, image captioning, and visual question answering), region-level localization tasks (e.g., object detection and phrase grounding), to pixel-level grouping tasks (e.g., image instance/semantic/panoptic segmentation). Until recently, most of these tasks have been separately tackled with specialized model designs, preventing the synergy of tasks across different granularities from being exploited.

In light of the versatility of transformers and inspired by large-scale vision-language pre-training, the computer vision community is now witnessing a growing interest in building general-purpose vision systems, also called vision foundation models, that can learn from and be applied to various downstream tasks, ranging from image-level, region-level, to pixel-level vision tasks.

In this tutorial, we will cover the most recent approaches and principles at the frontier of learning and applying vision foundation models, including but not limited to the latest advances such as SAM and GPT-4. Please refer to the tutorial's webpage for more details.

Tutorial: Automatic 3D Modeling of Indoor Structures From Panoramic Imagery

Organizers: Giovanni Pintore
Marco Agus
Enrico Gobbetti

Location: East 15

Time: Half Day - Morning (0900-1230)

Summary: Creating high-level structured 3D models of real-world indoor scenes from captured data and exploiting them are fundamental tasks with important applications in many fields. In this context, 360 capture and processing is very appealing, since panoramic imaging provides the quickest and most complete per-image coverage and is supported by a wide variety of professional and consumer capture devices. Research on inferring 3D indoor models from 360 images has been thriving in recent years, and has led to a variety of very effective solutions. Given the complexity and variability of interior environments, and the need to cope with noisy and incomplete captured data, many open research problems still remain. In this tutorial, we provide an up-to-date integrative view of the field. After introducing a characterization of input sources, we define the structure of output models, the priors exploited to bridge the gap between imperfect input and desired output, and the main characteristics of geometry reasoning and data-driven approaches. We then identify and discuss the main subproblems in structured reconstruction, and review and analyze state-of-the-art solutions for floor plan segmentation, bounding surfaces reconstruction, object detection and reconstruction, integrated model computation, and visual representation generation. We finally point out relevant research issues and analyze research trends.

Tutorial: Multi-Objective Optimization for Deep Learning

Organizers: Vishnu Naresh Boddeti
Zhichao Lu
Qingfu Zhang
and Kalyanmoy Deb

Location: West 113

Time: Half Day - Morning (0900-1200)

Summary: Real-world applications of deep learning must often contend with objectives beyond predictive performance, i.e., more than one equally important and competing objective or criterion. Examples include cost functions pertaining to invariance (e.g., to photometric or geometric variations), semantic independence (e.g., to age or race for face recognition systems), privacy (e.g., mitigating leakage of sensitive information), algorithmic fairness (e.g., demographic parity), generalization across multiple domains, computational complexity (FLOPs, compactness), to name a few. In such applications, achieving a single solution that simultaneously optimizes all objectives is no longer feasible; instead, finding a set of solutions that are representative in describing the trade-off among objectives becomes the goal. Multiple approaches have been developed for such problems, including simple scalarization and population-based methods. This tutorial aims to provide a comprehensive introduction to fundamentals, recent advances, and applications of multi-objective optimization (MOO), followed by hands-on coding examples. Some emerging applications of MOO include (1) hardware-aware neural architecture search; (2) multi-task learning as multi-objective optimization; (3) representation learning for privacy and fairness. We will also summarize potential research directions intersecting MOO and ML/CV research.

Tutorial: Rolling Shutter Camera: Modeling, Optimization, Learning, and Hardware

Organizers: Yuchao Dai, Zhihang Zhong
Yinqiang Zheng, Zhixiang Wang
Bin Fan

Location: East 17

Time: Half Day - Morning (0900-1200)

Summary: This half-day tutorial will cover the latest advances in this area from three aspects, i.e., motion modeling and optimization-based solutions, deep learning-based solutions, and joint hardware and deep learning-based solutions. Specifically, we will first systematically present geometric motion models (like discrete, continuous, and special motions) and optimization-based approaches. Then, we will introduce deep learning-based RS image processing methods, such as RS image correction and RS temporal super-resolution, with new results and benchmarks that have recently appeared. Finally, we will elaborate on the combination of hardware features of RS cameras (e.g., dual RS cameras and global reset feature) and deep learning to boost the correction of RS geometric distortions.

Tutorial: Optics for Better AI: Capturing and Synthesizing Realistic Data for Low-Light Enhancement

Organizers: Yinqiang Zheng
Yunhao Zou
Haiyang Jiang
Ying Fu

Location: West 114-115

Time: Half Day - Morning (0900-1200)

Summary: This half-day tutorial will cover the latest advances in the broad theme of Optics for Better AI, with a specific focus on how to capture and synthesize realistic data for training low-light enhancement deep models. In this tutorial, we will first present the overall pipeline and effects of using realistic data, including (i) Low-light Image Enhancement using Synthesized Data; (ii) Low-light Video Enhancement using Captured Data. Then, we show detailed instructions on noise calibration and construction of optical imaging systems, including (iii) How to Calibrate the Noise Model of a Specific Camera; (iv) How to Construct a Co-axial Imaging System.

Tutorial: Deep Learning Theory for Computer Vision

Organizers: Grigorios Chrysos
Fanghui Liu
Volkan Cevher

Location: West 211

Time: Half Day - Morning (0900-1200)

Summary: What is the interplay of width/depth and how does the initialization affect the robustness to adversarial attacks? What is a principled heuristic for selecting good architectures in Neural Architecture Search (NAS)? What is the role of Fourier features in implicit neural representations (INRs)? In this tutorial, we aim to build a bridge between the empirical performance of neural networks and deep learning theory. We want to make the recent deep learning (DL) theory developments accessible to vision researchers, and motivate vision researchers to design new architectures and algorithms for practical tasks. In the first part of the tutorial, we will discuss popular notions in DL theory, such as lazy training and Neural Tangent Kernel (NTK), or bilevel optimization for adversarial attacks and NAS. Then, we will exhibit how such tools can be critical in understanding the inductive bias of networks.

Tutorial: Prompting in Vision

Organizers: Kaiyang Zhou
Ziwei Liu
Phillip Isola
Hyojin Bahng
Ludwig Schmidt
Sarah Pratt
Denny Zhou

Location: West 223-224

Time: Half Day - Morning (0900-1200)

Summary: Originating from natural language processing, the new paradigm of prompting has recently swept through the computer vision community, bringing disruptive changes to various computer vision applications, such as image recognition and image generation. In comparison to the traditional fixed-once-learned architecture, like a linear classifier trained to recognize a specific set of categories, prompting offers greater flexibility and more opportunities for novel applications. It allows the model to perform new tasks, such as recognizing new categories, by tuning textual instructions or modifying a small number of parameters in the model's input space while keeping the majority of the pre-trained parameters untouched. This paradigm significantly pushes conversational human-AI interaction to unprecedented levels. Within a short period of time, the effectiveness of prompting has been demonstrated in a wide range of problem domains, including image classification, object detection, image generation and editing, video analytics, and robot control. In this tutorial, our aim is to provide a comprehensive background on prompting by building connections between research in computer vision and natural language processing. We will also review the latest advances in using prompting to tackle computer vision problems.

Tutorial: Knowledge-Driven Vision-Language Encoding

Organizers: Manling Li
Xudong Lin
Jie Lei

Location: East 8

Time: Half Day - Morning (0900-1230)

Summary: Does knowledge still have value in current era of large-scale pretraining? In this tutorial, we will comprehensively review existing paradigms for multimedia knowledge discovery and encoding, and focus on their contributions to vision-language pretraining. We categorize the knowledge into internal self-knowledge and external knowledge. Internal knowledge are extracted from text and vision modalities, such as structured entities, relations, events, and event procedures. We will focus on the structural aspects of the knowledge and address two key challenges regarding the acquisition of knowledge and encoding of structure across multiple modalities. External knowledge can be obtained from knowledge bases or language models, and we will exemplify their use to assist in commonsense understanding of vision modalities, with a focus on the temporal and cognitive aspects. The objective of this tutorial is to introduce participants to recent trends and emerging challenges in knowledge-driven vision-language research, as well as learning resources and tools for participants to obtain ready-to-use models, prompting thorough discussions regarding the impact of structured knowledge on text and vision learning.

Tutorial: Few-Shot Learning From Meta-Learning, Statistical Understanding to Applications

Organizers: Yanwei Fu
Da Li
Yu-Xiong Wang
Timothy Hospedales

Location: East 5

Time: Half Day - Morning (0900-1230)

Summary: There is a growing trend of research in few-shot learning (FSL), which involves adapting learned knowledge to learn new concepts with limited few-shot training examples. This tutorial comprises several talks, including an overview of few-shot learning by Dr. Da Li and a discussion of seminal and state-of-the-art meta-learning methods for FSL by Prof. Timothy Hospedales. The tutorial will cover both gradient-based and amortised meta-learners, as well as some theory for meta-learning, and Dr. Yanwei Fu will introduce recent FSL techniques that use statistical methods, such as exploiting the support of unlabeled instances for few-shot visual recognition and causal inference for few-shot learning. Dr. Yu-Xiong Wang will also discuss various applications of FSL in fields beyond computer vision, such as natural language processing, reinforcement learning, and robotics.

Tutorial: Physics-Based Rendering and Its Applications in Computational Photography and Imaging

Organizers: Ioannis Gkioulekas
Adithya Pediredla

Location: East 8

Time: Half Day - Afternoon (1330-1700)

Summary: Physics-based rendering algorithms simulate photorealistic radiometric measurements captured by a variety of sensors, including conventional cameras, time-of-flight sensors, lidar, and so on. They do so by computationally mimicking the flow of light through a mathematical representation of a virtual scene. This capability has made physics-based rendering a key ingredient in inferential pipelines for computational photography, computer vision, and computer graphics applications. For example, forward renderers can be used to simulate new camera systems or optimize the design of existing ones. Additionally, they can generate datasets for further training and optimization of tailored post-processing algorithms, jointly with hardware in an end-to-end fashion. Differentiable renderers can be used to backpropagate through image losses involving complex light transport effects. This makes it possible to solve previously intractable analysis-by-synthesis problems, and to incorporate physics-based simulation modules into probabilistic inference, deep learning, and generative pipelines. The goal of this tutorial is to introduce physics-based rendering, and highlight relevant theory, algorithms, implementations, and current and future applications in computer vision and related areas. This material should help equip computer vision researchers and practitioners with the necessary background for utilizing state-of-the-art rendering tools in a variety of exciting applications in vision, graphics, computational photography, and computational imaging.

Tutorial: Neural Search in Action

Organizers: Yusuke Matsui
Martin Aumuller
Han Xiao

Location: West 113

Time: Half Day - Afternoon (1330-1630)

Summary: Neural search, a technique for efficiently searching for similar items in deep embedding space, is the most fundamental technique for handling large multimodal collections. With the advent of powerful technologies such as foundation models and prompt engineering, efficient neural search is becoming increasingly important. For example, multimodal encoders such as CLIP allow us to convert various problems into simple embedding-and-search. Another example is the way to feed information into LLMs; currently, vector search engines are a promising direction. Despite the above attention, it is not obvious how to design a search algorithm for given data. In this tutorial, we will focus on "million-scale search", "billion-scale search", and "query language" to show how to tackle real-world search problems.

Tutorial: Full-Stack, GPU-Based Acceleration of Deep Learning

Organizers: Maying Shen Pavlo Molchanov
Hongxu Yin Jose M. Alvarez
Jason Clemons Jan Kautz

Location: East 11

Time: Half Day - Afternoon (1330-1700)

Summary: This tutorial focuses on describing techniques to allow deep learning practitioners to accelerate the training and inference of large deep networks while also reducing memory requirements across a spectrum of off-the-shelf hardware for important applications such as autonomous driving and large language models. Topics include, but are not limited to:

- Deep learning specialized hardware overview. We review the architecture of the most used deep learning acceleration hardware, including the main computational processors and memory modules.
- How deep learning is performed on this hardware. We cover aspects of algorithmic intensity and an overview of theoretical aspects of computing. Attendees will learn how to estimate processing time and latency by looking only at hardware specs and the network architecture.
- Best practices for acceleration. We provide an overview of best practices for designing efficient neural networks including channel number selection, compute heavy operations, or reduction operations among others.
- Existing tools for model acceleration. In this part we will focus on existing tools to accelerate a trained neural network on GPU devices. We will particularly discuss operation folding, TensorRT, ONNX graph optimization, sparsity.
- Research overview of recent techniques. In the last part, we will focus on recent advanced techniques for post training model optimization including pruning, quantization, model distillation or NAS among others.

Tutorial: Hands-On Egocentric Research With Project Aria From Meta

Organizers: Edward Miller Richard Newcombe
Pierre Moulon Vasileios Balntas
Prince Gupta Xiaqing Pan
Rawal Khirodkar

Location: East 12

Time: Half Day - Afternoon (1330-1700)

Summary: Project Aria is a research device from Meta, which is worn like a regular pair of glasses, and enables researchers to study the future of always-on egocentric perception. In this tutorial, we will introduce two exciting new datasets from Project Aria: Aria Digital Twin, a real-world dataset with hyper-accurate digital counterpart; and Aria Synthetic Environments, a procedurally-generated synthetic Aria dataset for large-scale ML research. Each dataset will be presented with corresponding challenges, which we believe will be powerful catalysts for research. In addition to introducing new datasets and research challenges, we will also provide a hands-on demonstration of newly open-sourced tools for working with Project Aria, and demonstrate how the Project Aria ecosystem can be used to accelerate open research into egocentric perception tasks such as visual and non-visual localization and mapping, static and dynamic object detection and spatialization, human pose and eye-gaze estimation, and building geometry estimation.

Tutorial: Exploring Synthetic Data as an Enterprise Capability for Training and Validating CV Systems

Organizers: Nathan Kundtz
Matt Robinson
Dan Hedges

Location: East 18

Time: Half Day - Afternoon (1330-1630)

Summary: With the rise of edge computing, increase in remote sensing information, and ubiquitous adoption of computer vision systems throughout retail and manufacturing markets, organizations are increasingly relying on the accuracy and reliably of training Artificial Intelligence and Machine Learning systems to analyze and extract information from data captured using physical sensors and sensor platforms. Real data sets often fail to capture rare events or assets, are inaccurately labeled, and the collection of real sensor data can have cost, privacy, security, and safety issues.

Synthetic data offers the opportunity to design and label datasets for specific algorithmic training needs. Synthetic imagery designed to emulate ground-based video systems or remotely sensed satellite imagery, for example, can be generated to show real world locations populated with objects that are hard to find or that don't yet exist. Accurately labeled, simulated datasets can be created to fit a wide range of potential real-world scenarios in which AI/ML systems will be deployed, thereby enabling teams to train and test these systems before being deployed in production environments.

This tutorial will include an introduction to creating, using, and iterating on synthetic data using the open Rendered.ai synthetic data platform. We will also feature a demonstration using NVIDIA Omniverse Replicator in the AWS cloud. The tutorial will define physics-based synthetic data, discuss differences with Generative AI, and introduce concepts for designing synthetic data.

Monday, June 19

NOTE: Workshop rooms are subject to change. Refer to the online site for up-to-date locations. Use the QR code for each workshop to see its schedule. Here is the QR code for the CVPR 2023 Workshops page.

0700–1700 Registration (West Ballroom Foyer)

0700–0900 Breakfast (West Ballrooms A–D)

1000–1045 Morning Break West Ballrooms A–D

1145–1345 Lunch (West Ballrooms A–D)

1500–1545 Afternoon Break (West Ballrooms A–D)

Vision-Centric Autonomous Driving

Organizers: Yue Wang
Hang Zhao
Vitor Guizilini
Xin Wang
Katherine Driggs-Campbell

Location: West 202-204

Time: Full Day (0800-1700)

Summary: With the commercialization of autonomous driving and assisted driving systems, the demand for high-performance, efficient, and scalable machine learning solutions is becoming more urgent than ever before. Visual perception is a key research area of self-driving that is always attracting a lot of attention since 1) visual data provides much richer information than other sensors; 2) there is an abundance of existing visual data of driving for machine learning; and 3) cameras are affordable and pervasive on vehicles as well as other robotic systems. This workshop embraces topics around vision-centric and data-driven autonomous driving technologies, including vision-only or sensor fusion-based perception, self- and semi-supervised visual learning, visual perception simulation, and data-driven motion prediction and planning.

Vision-based InduStrial InspectiON

Organizers: Meng Cao
Haoping Bai
Shancong Mou
Oncel Tuzel
Tatiana Likhomanenko
Ramazan Gokberk Cinbis

Location: West 208

Time: Full Day (0830-1800)

Summary: The VISION workshop aims to provide a platform for the exchange of scholarly innovations and emerging practical challenges in Vision-based Industrial Inspection. Through a series of keynote talks, technical presentations, and challenge competition, this workshop is intended to (i) bring together researchers from the interdisciplinary research communities related to computer vision-based inspection; (ii) connect researchers and industry practitioners to synergize recent research progress and current needs in industrial practice.

Bridging the Gap Between Computational Photography and Visual Recognition

Organizers: Zhiyuan Mao
Wuyang Chen
Abdullah AlShabili
Zhenyu Wu
Xingguang Zhang
Ajay Kumar Jaiswal
Yunhao Ba
Howard Zhang
Stanley Chan
Zhangyang Wang
Achuta Kadambi
Alex Wong
Kevin Miller
Jiaying Liu
Walter Scheirer
Wenqi Ren

Location: West 107-108

Time: Full Day (0830-1700)

Summary: The rapid development of computer vision algorithms increasingly allows automatic visual recognition to be incorporated into a suite of emerging applications. Some of these applications have less-than-ideal circumstances such as low-visibility environments, causing image captures to have degradations. In other more extreme applications, such as imagers for flexible wearables, smart clothing sensors, ultra-thin headset cameras, implantable in vivo imaging, and others, standard camera systems cannot even be deployed, requiring new types of imaging devices. Computational photography addresses the concerns above by designing new computational techniques and incorporating them into the image capture and formation pipeline. This raises a set of new questions. For example, what is the current state-of-the-art for image restoration for images captured in non-ideal circumstances? How can inference be performed on novel kinds of computational photography devices? Continuing the success of the 1st (CVPR'18), 2nd (CVPR'19), 3rd (CVPR'20), 4th (CVPR'21), and 5th (CVPR'22) UG2 Prize Challenge workshops, we provide its 6th version for CVPR 2023. It will inherit the successful benchmark dataset, platform and evaluation tools used by the previous UG2 workshops, but will also look at brand new aspects of the overall problem, significantly augmenting its existing scope.

Computer Vision for Microscopy Image Analysis

Organizers: Mei Chen
Daniel J. Hoeppner
Dimitris N. Metaxas
Steve Finkbeiner

Location: East 10

Time: Full Day (0800-1800)

Summary: High-throughput microscopy enables researchers to acquire thousands of images automatically over a matter of hours. This makes it possible to conduct large-scale, image-based experiments for biological discovery. The main challenge and bottleneck in such experiments is the conversion of "big visual data" into interpretable information and hence discoveries. Visual analysis of large-scale image data is a daunting task. Cells need to be located and their phenotype (e.g., shape) described. The behaviors of cell components, cells, or groups of cells need to be analyzed. The cell lineage needs to be traced. Not only do computers have more "stamina" than human annotators for such tasks, they also perform analysis that is more reproducible and less subjective. The post-acquisition component of high-throughput microscopy experiments calls for effective and efficient computer vision techniques.

This workshop intends to draw more visibility and interest to this challenging yet fruitful field, and establish a platform to foster in-depth idea exchange and collaboration.

Safe Artificial Intelligence for All Domains

Organizers: Timo Sämann
Oliver Wasenmüller
Markus Enzweiler
Peter Schlicht
Johannes Otterbach
Christian Wojek
Stefan Milz
Oliver Grau
Thomas Stauner
Joachim Sicking
Claus Bahlmann

Location: East 13

Time: Full Day (0900-1700)

Summary: After the success of ML and AI-based approaches in outperforming traditional vision algorithms, recently a lot of research effort is dedicated to understanding of the limitations and the general behavior of AI methods in a broad range of computer vision applications. Specifically for a successful introduction of ML and AI in a wider range of products, safety is often a top priority. Being able to ensure safety of ML based computer vision is key to unlock its potential in a broad range of safety related applications and future products. In domains like automotive, aviation and the medical domain, it paves the way towards systems with a greater degree of autonomy and assistance for humans.

The workshop focuses on bringing together researchers, engineers, and practitioners from academia, industry, and government to exchange ideas, share their latest research, and discuss the latest trends and challenges in this field. The workshop also aims to foster collaboration between different stakeholders, including computer vision researchers, machine learning experts, robotics engineers and safety experts, to create a comprehensive framework for developing safe AI systems for all domains. Overall, the SAIAD workshop aims to advance the state-of-the-art in safe AI, address the most pressing challenges, and provide a platform for networking and knowledge sharing among the experts in this field.

Deep Learning for Geometric Computing

Organizers: Dena Bazazian
Ilke Demir
Bernhard Egger
Geraldine Morin
Kathryn Leonard
Adarsh Krishnamurthy
Silvia Sellán

Location: East Ballroom C

Time: Full Day (0900-1700)

Summary: Computer vision approaches have made tremendous efforts toward understanding shape from various data formats, especially since entering the deep learning era. Although accurate results have been obtained in detection, recognition, and segmentation, there is less attention and research on extracting topological and geometric information from shapes. These geometric representations provide compact and intuitive abstractions for modeling, synthesis, compression, matching, and analysis. Extracting such representations is significantly different from segmentation and recognition tasks, as they contain both local and global information about the shape.

To advance the state of the art in topological and geometric shape analysis using deep learning, we aim to gather researchers from computer vision, computational geometry, computer graphics, and machine learning in this third edition of "Deep Learning for Geometric Computing" workshop at CVPR 2023. The workshop encapsulates competitions with prizes, proceedings, keynotes, paper presentations, and a fair and diverse environment for brainstorming about future research collaborations.

Secure and Safe Autonomous Driving

Organizers: Chejian Xu
Wenhao Ding
Haohong Lin
Mansur Arief
Jiawei Zhang
Hazem Torfah
Alberto L. Sangiovanni-Vincentelli
Sanjit A. Seshia
Ding Zhao
Bo Li

Location: West 301

Time: Full Day (0845-1700)

Summary: Despite the great success achieved by machine learning recently, extensive studies have shown that machine learning algorithms are vulnerable to adversarial attacks or natural distribution shifts, which has raised great concerns when deploying machine learning algorithms for real-world applications, especially in safety-critical domains such as autonomous driving (AD). While there have been significant advances in AD (e.g., perception, planning and control, etc.), the security and safety of these algorithms are often challenged by various realistic safety-critical scenarios.

In this workshop, we aim to explore and discuss recent research and summarize potential future directions for secure and safe AD algorithms. In particular, we will host different invited talks, paper submissions, panel discussions, and a safe AD competition based on our unified platform SafeBench, which is developed to integrate different types of safety-critical testing scenarios, scenario generation algorithms, and other variations such as driving routes and environments, to provide comprehensive learning and testing environment for AD algorithms.

We will bring together experts from computer vision, reinforcement learning, security, and trustworthy machine learning communities, in an attempt to highlight recent work in this area as well as to clarify the foundations of secure autonomous driving. We hope this workshop will help to chart out important directions for future work and cross-community collaborations.

Federated Learning for Computer Vision

Organizers: Chen Chen
Salman Avestimehr
Zhengming Ding
Mi Zhang
Ravikumar Balakrishnan
Nageen Himayat
Ang Li
Lingjuan Lyu
Naji Khosravan
Seyyedali Hosseinalipour
Lichao Sun

Location: West 217-219

Time: Full Day (0830-1730)

Summary: Federated Learning (FL) has become an important privacy-preserving paradigm in various machine learning tasks. However, the potential of FL in computer vision applications, such as face recognition, person re-identification, and action recognition, is far from being fully exploited. Moreover, FL has rarely been demonstrated effectively in advanced computer vision tasks such as object detection and image segmentation, compared to the traditional centralized training paradigm. This workshop aims at bringing together researchers and practitioners with common interests in FL for computer vision and studying the different synergistic relations in this interdisciplinary area. The day-long event will facilitate interaction among students, scholars, and industry professionals from around the world to discuss future research challenges and opportunities.

Event-Based Vision

Organizers: Guillermo Gallego, Cornelia Fermuller, Davide Scaramuzza, Davide Migliore, Kostas Daniilidis

Location: West 209

Time: Full Day (0800-1800)

Summary: This workshop is dedicated to event-based cameras, smart cameras, and algorithms processing data from these sensors. Event-based cameras are bio-inspired sensors with the key advantages of microsecond temporal resolution, low latency, very high dynamic range, and low power consumption. Because of these advantages, event-based cameras open frontiers that are unthinkable with standard frame-based cameras (which have been the main sensing technology for the past 60 years). These revolutionary sensors enable the design of a new class of algorithms to track a baseball in the moonlight, build a flying robot with the agility of a bee, and perform structure from motion in challenging lighting conditions and at remarkable speeds. These sensors became commercially available in 2008 and are slowly being adopted in computer vision and robotics. In recent years they have received attention from large companies, e.g., the event-sensor company Prophesee collaborated with Intel and Bosch on a high spatial resolution sensor, Samsung announced mass production of a sensor to be used on hand-held devices, and they have been used in various applications on neuromorphic chips such as IBM's TrueNorth and Intel's Loihi. The workshop also considers novel vision sensors, such as pixel processor arrays, which perform massively parallel processing near the image plane. Because early vision computations are carried out on-sensor, the resulting systems have high speed and low-power consumption, enabling new embedded vision applications.

Computer Vision in Sports

Organizers: Rikke Gade, Adrian Hilton, Thomas B. Moeslund, James J. Little, Graham Thomas, Michele Merler

Location: West 214

Time: Full Day (0900-1730)

Summary: Sports is said to be the social glue of society. It allows people to interact irrespective of their social status, age etc. With the rise of the mass media, significant resources have been channeled into sports in order to improve understanding, performance and presentation. For example, areas like performance assessment are now finding applications in broadcast and other media, driven by the increasing use of online sports viewing which provides all sorts of performance statistics available to viewers. Computer vision has recently started to play an important role in sports as seen in for example football where computer vision-based graphics in real-time enhances different aspects of the game. Vision algorithms have a huge potential in many aspects of sports ranging from automatic annotation of broadcast footage, through to better understanding of sport injuries, and enhanced viewing. So far, the use of computer vision in sports has been scattered between different disciplines. The ambition of this workshop is to bring together practitioners and researchers from different disciplines to share ideas and methods on current and future use of computer vision in sports. The workshop program consists of oral and poster presentations of peer-reviewed papers, as well as invited talks from both industry and academia, and challenge results.

AI for Content Creation

Organizers: Deqing Sun, Seungjun Nah, Huiwen Chang, James Tompkin, Lu Jiang, Ting-Chun Wang, Yijun Li, Fitsum Reda, Lingjie Liu, Jun-Yan Zhu

Location: East Exhibit Hall A

Time: Full Day (0900-1815)

Summary: The AI for Content Creation (AI4CC) workshop brings together researchers in computer vision, machine learning, and AI. Content creation is required for simulation and training data generation, media like photography and videography, virtual reality and gaming, art and design, and documents and advertising (to name just a few application domains). Recent progress in machine learning, deep learning, and AI techniques has allowed us to turn hours of manual, painstaking content creation work into minutes or seconds of automated or interactive work. For instance, generative adversarial networks (GANs) can produce photorealistic images of 2D and 3D items such as humans, landscapes, interior scenes, virtual environments, or even industrial designs. Neural networks can super-resolve and super-slomo videos, interpolate between photos with intermediate novel views and even extrapolate, and transfer styles to convincingly render and reinterpret content. In addition to creating awe-inspiring artistic images, these offer unique opportunities for generating additional and more diverse training data. AI for content creation lies at the intersection of the graphics, the computer vision, and the design community. However, researchers and professionals in these fields may not be aware of its full potential and inner workings. We hope that the workshop will serve as a forum to discuss the latest topics in content creation and the challenges that vision and learning researchers can help solve.

Mobile AI

Organizers: Andrey Ignatov, Radu Timofte

Location: Virtual (AM); West 114-115 (PM)

Time: Full Day (0800-1800)

Summary: Over the past years, mobile AI-based applications are becoming more and more ubiquitous. Various deep learning models can now be found on any mobile device, starting from smartphones running portrait segmentation, image enhancement, face recognition and natural language processing models, to smart-TV boards coming with sophisticated image super-resolution algorithms. The performance of mobile NPUs and DSPs is also increasing dramatically, making it possible to run complex deep learning models and to achieve fast runtime in most tasks. While many research works targeted at efficient deep learning models have been proposed recently, the evaluation of the obtained solutions is usually happening on desktop CPUs and GPUs, making it nearly impossible to estimate the actual inference time and memory consumption on real mobile hardware. To address this problem, we introduce the first Mobile AI Workshop, where all deep learning solutions are developed for and evaluated on mobile devices. Due to the performance of the last-generation mobile AI hardware, the topics considered in this workshop will go beyond the simple classification tasks, and will include such challenging problems as image denoising, HDR photography, accurate depth estimation, learned image ISP pipeline, real-time image and video super-resolution.

Computer Vision in the Wild

Organizers: Jianwei Yang
Haotian Zhang
Haotian Liu
Xiuye Gu
Chunyuan Li
Neil Houlsby
Jianfeng Gao

Location: East Ballroom B

Time: Full Day (Time TBA)

Summary: State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concepts.

Recent works show that learning from large-scale image-text data is a promising approach to building transferable visual models that can effortlessly adapt to a wide range of downstream computer vision (CV) and multimodal (MM) tasks. For example, CLIP, ALIGN and Florence for image classification, ViLD, RegionCLIP, GLIP and OWL-ViT for object detection, GroupViT, OpenSeg, MaskCLIP, X-Decoder, Segment Anything (SAM) and SEEM for segmentation, LLaVA for langauge-and-image instruction-following chatbots built towards multimodal GPT-4 capabities. These vision models with language or interactive interface are naturally open-vocabulary recogntion models, showing superior zero-shot and few-shot adaption performance on various real-world scenarios.

We host this "Computer Vision in the Wild (CVinW)" workshop, aiming to gather academic and industry communities to work on CV and MM problems in real-world scenarios, focusing on the challenge of open-set/domain visual recognition at different granularities and efficient task-level transfer. To measure the progress of CVinW, we develop new benchmarks for image classification, object detection and segmentation to measure the task-level transfer ablity of various models/methods over diverse real-world datasets, in terms of both prediction accuracy and adaption efficiency.

Embodied AI

Organizers: Claudia Perez-D'Arpino
Anthony Francis
Luca Weihs
Lamberto Ballan
Yonatan Bisk
Angel X. Chang
Devendra Singh Chaplot
Changan Chen
Matt Deitke
David R. Hall
R. Devon Hjelm
Chengshu Li
Oleksandr Maksymets
Katherine Metcalf
Soeren Pirk
Mike Roberts
Mohit Shridhar
Andrew Szot
Jesse Thomason
Naoki Yokoyama

Location: East Ballroom A

Time: Full Day (0900-1730)

Summary: The goal of the Embodied AI workshop is to bring together researchers from computer vision, language, graphics, and robotics to share and discuss the latest advances in embodied intelligent agents. This year's workshop will focus on the three themes of: Minds live in bodies, and bodies move through a changing world. The goal of embodied artificial intelligence is to create agents, such as robots, which learn to creatively solve challenging tasks requiring interaction with the environment. While this is a tall order, fantastic advances in deep learning and the increasing availability of large datasets like ImageNet have enabled superhuman performance on a variety of AI tasks previously thought intractable. Computer vision, speech recognition and natural language processing have experienced transformative revolutions at passive input-output tasks like language translation and image processing, and reinforcement learning has similarly achieved world-class performance at interactive tasks like games.

Learning With Limited Labelled Data for Image and Video Understanding

Organizers: Mennatullah Siam
Xin Wang
Pau Rodriguez
Issam Hadj Laradji
Katerina Fragkiadaki
David Vazquez
Boris N. Oreshkin
Richard P. Wildes
He Zhao
Konstantinos G. Derpanis

Location: East 3

Time: Full Day (0900-1710)

Summary: Deep learning has been widely successful in a variety of computer vision tasks such as object recognition, object detection, and semantic segmentation. It also has been deployed with success in learning spatiotemporal features for video segmentation/detection and action recognition tasks. However, one of the major bottlenecks of deep learning in both image and video understanding tasks is the need for large-scale labelled datasets. Collecting and annotating such datasets can be labor intensive and costly. In many scenarios of practical interest only a few labelled examples of novel categories may be available at model training time. Currently available large-scale data typically cover relatively narrow sets of categories and are constrained by licensing. As such, they are often hard to naively apply to practical problems. It is especially problematic in developing countries that do not have the required resources to collect large scale labelled datasets for new tasks. The goal of this workshop is to explore approaches that learn from limited labelled data, or with side information such as text data, or using data with weak/self supervision, with special focus on video understanding tasks. This will be the second L3D-IVU workshop in conjunction with CVPR, where it had a great success and wide interest last year from multiple researchers as it explores the intersection of learning with limited labelled data and video understanding.

Efficient Deep Learning for Computer Vision

Organizers: Bichen Wu
Peter Vajda
Peizhao Zhang
Xiaoliang Dai
Tao Xu
Andrew Howard
Chas H. Leichner
Kurt Keutzer
Yung-Hsiang Lu
Kate Saenko
Ping Hu
Dilin Wang

Location: West 118-120

Time: Full Day (0850-1835)

Summary: As computer vision algorithms, models, and systems become increasingly more powerful in understanding and generating visual contents, computer vision research has not sufficiently considered compute efficiency — speed or computation time, power/energy, memory footprint, model size, or carbon emission; and data efficiency — the amount of training data or labels needed to train models. Nevertheless, addressing all these metrics is essential if advances in Computer Vision are going to be widely available on mobile and AR/VR devices. In this year's ECV workshop, our topics include but are not limited to the following: Efficient neural architecture; Compression, quantization and hardware acceleration, data-efficient learning, efficient generative models & 3D models, mobile and AR/VR applications

AI City Challenge

Organizers: Milind Naphade
Shuo Wang
Zheng Tang
David C. Anastasiu
Ming-Ching Chang
Pranamesh Chakraborty
Liang Zheng
Anuj Sharma
Stan Sclaroff
Rama Chellappa

Location: East 4

Time: Full Day (0800-1730)

Summary: The AI City Challenge Workshop aims to advance the application of AI in physical environments, from retail and warehouse operations to transportation outcomes. By reducing friction in these environments, AI can facilitate speedier check-outs, improve traffic efficiency, enhance road safety, and more. This year, the workshop focuses on two domains: brick-and-mortar retail and Intelligent Traffic Systems (ITS). In brick-and-mortar retail, AI can be used to improve multi-camera people tracking and automated checkout. In ITS, AI can be used to retrieve tracked vehicles by natural language, analyze naturalistic driver data, and improve traffic safety. We solicit original contributions in these and related areas where computer vision, natural language processing, and deep learning have shown promise in achieving large-scale practical deployment that will help make our environments smarter and safer. To accelerate the research and development of techniques, the 7th edition of this Challenge pushes the research and development in multiple directions. We released a brand-new dataset for multi-camera people tracking where a combination of real and synthetic data was provided for training and evaluation. The synthetic data were generated by the NVIDIA Omniverse Platform that creates highly realistic characters and environments as well as a variety of random lighting, perspectives, avatars, etc. We also expand the diversity of Traffic related tasks such as helmet safety and the diversity of datasets including data from traffic cameras in India.

Joint Ego4D and EPIC Workshop on Egocentric Vision

Organizers: Dima Damen
Kristen Grauman
Giovanni Maria Farinella
Rohit Girdhar
Michael Wray
Antonino Furnari
David Crandall
Jitendra Malik
Andrew Westbury
Kris Kitani
James Rehg
Bernard Ghanem
C.V. Jawahar
Yoichi Sato
Mike Zheng Shou
Sanja Fidler
Christian Micheloni
David Fouhey
Pablo Arbelaez
Jianbo Shi
Hyun Soo Park
Vamsi Krishna K. Ithapu
Lorenzo Torresani
Richard Newcombe

Location: West 111-112

Time: Full Day (0830-1845)

Summary: This joint full-day workshop is the longstanding event that brings together the strongly growing egocentric computer vision community, offering the 3rd Ego4D edition and the 11th Egocentric Perception, Interaction and Perception (EPIC) edition. This year, 17 Ego4D benchmark and 9 EPIC benchmark winners and findings will be presented throughout the day, ranging from social interactions, episodic memory, hand-object interactions, long-term tracking, video object segmentations and audio-based interaction recognition. In addition to the recurring Ego4D and EPIC challenges, new challenges are associated with recently released benchmarks EgoTracks, PACO, EPIC-KTICHENS VISOR and EPIC-Sounds.

Additionally, the day will include accepted abstracts, invited CVPR papers and 5 keynotes by Andrea Vedaldi (Oxford and Meta), Hyun Soo Park (UMinnesota), David Fouhey (UMich) and Suraj Nair (Stanford). Check the program for details.

Open-Domain Reasoning Under Multi-Modal Settings

Organizers: Tejas Gokhale
Man Luo
Kenneth Marino
Pratyay Banerjee
Zhiyuan Fang
Yezhou Yang
Chitta Baral

Location: West 201

Time: Full Day (0830-1700)

Summary: AI has undergone a paradigm shift in the past decade -- the connection between vision and language (V+L) is now an integral part of AI, with deep impact beyond vision and NLP -- robotics, graphics, cybersecurity, and HCI are utilizing V+L tools and there are direct industrial implications for software, arts, and media. The link between vision and language is much more complex than simple image--text alignment – the use of language for reasoning beyond the visible (e.g., physical, spatial, commonsense, and embodied reasoning) is being pursued. Open-Domain Reasoning in Multi-Modal Settings (ODRUM 2023) provides a platform for discussions on multimodal (vision+language) topics with special emphasis on reasoning capabilities.

The aim of ODRUM 2023 is to address the emerging topic of visual reasoning using multiple modalities (text, images, videos, audio, etc.). The workshop will feature invited talks by experts in the realm of reasoning such as: embodied AI, navigation, learning via interaction and collaboration with humans, building large V+L that can perform multiple tasks, visual grounding, and the use of language to instruct robots. Participants and speakers will converge for a panel discussion to discuss the importance of reasoning (a core AI topic that has a rich and long history since the 1950s) to computer vision, relevance to recent progress in visual reasoning, discuss trends and challenges in open-domain reasoning, from different perspectives of NLP, vision, machine learning, and robotics researchers.

Explainable AI for Computer Vision

Organizers: Sunnie S. Y. Kim
Vikram V. Ramaswamy
Ruth C. Fong
Filip Radenovic
Abhimanyu Dubey
Deepti Ghadiyaram

Location: West 121-122

Time: Full Day (0900-1730)

Summary: Explainability of computer vision systems is critical for people to effectively use and interact with them. The 2nd Explainable AI for Computer Vision (XAI4CV) workshop seeks to contribute to the development of more explainable CV systems by: (1) initiating discussions across researchers and practitioners in academia and industry to identify successes, failures, and priorities in current XAI work; (2) examining the strengths, weaknesses, and underlying assumptions of proposed XAI methods and establish best practices in evaluation of these methods; and (3) discussing the various nuances of explainability and brainstorm ways to build explainable CV systems that benefit all involved stakeholders.

Visual Pre-Training for Robotics

Organizers: Ilija Radosavovic
Tete Xiao
Lerrel Pinto
Mathilde Caron
Amy Zhang
Shuran Song
Pieter Abbeel
Trevor Darrell

Location: West 220-222

Time: Full Day (0900-1730)

Summary: The great vision scientist, James J. Gibson, famously said "We see in order to move and we move in order to see." But can we learn to see before we learn to move? And how far can we move if we first learn to see?

This interdisciplinary workshop will focus on visual pre-training for robotics. The goals of this workshop are (1) to present key questions and the state of the art on visual pre-training for robotics, and (2) to encourage the wider computer vision community to consider future original contributions in this space.

We are excited to have a lineup of speakers from computer vision, machine learning, and robotics. We hope that this workshop will help attract the broader CVPR community to this important and exciting topic.

Vision for All Seasons: Adverse Weather and Lighting Conditions

Organizers: Dengxin Dai
Christos Sakaridis
Haoran Wang
Lukas Hoyer
Wim Abbeloos
Daniel Olmeda Reino
Jiri Matas
Bernt Schiele
Luc Van Gool

Location: East 9

Time: Full Day (0900-1730)

Summary: Adverse weather and illumination conditions (e.g., fog, rain, snow, low light, nighttime, glare and shadows) create visibility problems for the sensors that power automated systems. Many outdoor applications such as autonomous cars and surveillance systems are required to operate smoothly in the frequent scenarios of bad weather. While rapid progress is being made in this direction, the performance of current vision algorithms is still mainly benchmarked under clear weather conditions (good weather, favorable lighting). Even the top-performing algorithms undergo a severe performance degradation under adverse conditions. The aim of this workshop is to promote research into the design of robust vision algorithms for adverse weather and lighting conditions.

Embedded Vision

Organizers: Nabil Belbachir
Tse-Wei Chen
Branislav Kisacanin
Marius Leordeanu

Location: West 213

Time: Full Day (0830-1700)

Summary: For over 20 years, this workshop has been the place to exchange experiences and learn the latest science and art of embedded vision. Embedded vision is an active field of research, bringing together efficient learning models with fast computer vision and pattern recognition algorithms, to tackle many areas of robotics and intelligent systems that are enjoying an impressive growth today. Such strong impact comes with many challenges that stem from the difficulty of understanding complex visual scenes under the tight computational constraints required by real-time solutions on embedded devices. The Embedded Vision Workshop will provide a venue for discussing these challenges by bringing together researchers and practitioners from the different fields outlined above. Such a topic is directly aligned with the topics of interest of the CVPR community. In addition to regular papers, this year's workshop is hosting five keynotes on trending topics such are vision systems in autonomous driving, in autonomous aquaculture operations and in machine perception.

Sight and Sound

Organizers: Andrew Owens
Andrew Zisserman
Antonio Torralba
Jiajun Wu
Kristen Grauman
Jean-Charles Bazin
William T. Freeman
Triantafyllos Afouras
Arsha Nagrani
Ruohan Gao
Hang Zhao

Location: West 207

Time: Full Day (0900-1800)

Summary: In recent years, there have been many advances in learning from visual and audio data. While traditionally these two modalities have been studied independently, researchers have increasingly been creating multimodal audio-visual models that learn from both at once. This has led to many developments in topics such as audio-visual speech understanding, action recognition, and multimodal self-supervised learning. This workshop will cover recent advances in audio-visual learning. It will also touch on higher-level questions, such as what information sound conveys that vision doesn't, the merits of sound versus other modalities (e.g., language) in self-supervised learning, and the role of sound in egocentric video understanding.

Visual Copy Detection

Organizers: Edward Pizzi
Hiral Patel
Gheorghe Postelnicu
Sugosh Nagavara Ravindra
Giorgos Kordopatis-Zilos
Symeon Papadopoulos
Giorgos Tolias
Ioannis Patras
Priya Goyal
Canton Cristian
Matthijs Douze

Location: East 12

Time: Half Day - Morning (0900-1200)

Summary: The Visual Copy Detection Workshop (VCDW) explores the task of identifying copied images and videos, robust to common transformations. This task is central to social problems facing online services where users share media, such as combating misinformation and exploitative imagery, as well as enforcing copyright. Recently, copy detection methods have been used to identify and promote original content, and to reduce memorization in both predictive and generative models. The workshop will explore technical advances in copy detection as well as the applications that motivate this research. The workshop will feature the Video Similarity Challenge, a copy detection challenge in the video domain, including presentations by challenge participants.

Foundation Models Challenge

Organizers: Teng Xi
Yifan Sun
Gang Zhang
Yi Yang
Errui Ding
Edith Ngai
Linchao Zhu
Jingdong Wang

Location: East 1

Time: Half Day - Morning (0900-1230)

Summary: Foundation model has attracted great interest from both the academia and the industry. By its early definition, the foundation model is a large artificial intelligence model trained on a vast quantity of unlabeled data at scale and can be adapted to a wide range of downstream tasks. Recent realistic applications further encourage using both the labeled and unlabeled data, therefore generalizing the concept of foundation model. This evolution is natural because besides the unlabeled data, many labeled datasets (from public or private resources) are large-scale and can bring substantial benefit to downstream tasks as well. In this workshop, we advocate the generalized foundation model with two considerations: 1) due to the combination of labeled and unlabeled data, it enlarges the potential benefit of large-scale pretraining, and 2) it is more flexible and efficient for downstream task adaptation. For example, a recent foundation model UFO trained with labeled datasets can be trimmed into a specific model for the already-seen sub-task without any adaptation cost.

Image Matching: Local Features and Beyond

Organizers: Vasileios Balntas
Luca Morelli
Fabio Bellavia
Fabio Remondino
Vincent Lepetit
Weiwei Sun
Jiri Matas
Eduard Trulls
Dmytro Mishkin
Kwang Moo Yi

Location: West 109-110

Time: Half Day - Morning (0800-1230)

Summary: Matching two or more images across wide baselines is a core computer vision problem with many applications. Until recently one of the last bastions of traditional handcrafted methods, they too have begun to be replaced with learned alternatives. Interestingly though, these new solutions often still rely on design intuitions behind handcrafted methods. Our field is in a transition stage, and our workshop aims to bring together researchers across academia and industry to assess its true state. We focus on what works and doesn't in practice, and for that purpose we hold an open challenge co-located with the workshop.

Women in Computer Vision

Organizers: Ivaxi Sheth
Asra Aslam
Doris Antensteiner
Ziqi Huang
Marah Halawa
Sachini A. Herath
Xin Wang
Naga Vara Aparna Akula
Sunnie S. Y. Kim

Location: West 205-206

Time: Half Day - Morning (0830-1300)

Summary: Computer vision has become one of the largest computer science research communities. We have made tremendous progress in recent years over a wide range of areas. However, despite the expansion of our field, the percentage of women researchers in both academia and industry is still relatively low. As a result, many women students and researchers in computer vision do not have a lot of opportunities to meet with other women and may feel isolated. The goals of this workshop are to: Raise visibility of women computer vision researchers through invited talks by leading women researchers in the field. Provide opportunities for junior women students and researchers to present their work via oral/poster sessions and travel awards. Exchange experience and career advice between women students and researchers. The half-day Women in Computer Vision (WiCV) workshop is a gathering for researchers of all genders and career stages. All are welcome and encouraged to attend the workshop. Travel grants will be offered to selected women presenters of oral and poster sessions.

Reconstruction of Human-Object Interaction

Organizers: Xi Wang
Nikos Athanasiou
Gerard Pons-Moll
Otmar Hilliges
Kaichun Mo
Xianghui Xie
Chun-Hao Paul Huang
Bharat Lal Bhatnagar

Location: East 18

Time: Half Day - Morning (0820-1230)

Summary: This half-day Rhobin workshop will provide a venue to present and discuss state-of-the-art research in the reconstruction of human-object interactions from images. The focus will be on recent developments in human-object interaction learning and its impact on 3D scene parsing, building human-centric robotic assistants, and the general understanding of human behaviors. Humans are an essential component of the interaction. Hence, it is crucial to estimate the human pose, shape, and motion as well as objects that are being interacted with accurately to achieve a realistic interaction. 3D Human Pose and Motion estimation from images or videos have attracted a lot of interest. However, in most cases, the task does not explicitly involve objects and the interaction with them. Whether it is 2D detection and/or monocular 3D reconstruction, objects and humans have been studied separately. Humans are in constant contact with the world as they move through it and interact with it. Considering the interaction between them can marry the best of both worlds.

VizWiz: Describing Images and Videos Taken by Blind People

Organizers: Danna Gurari
Daniela Massiceti
Abigale Stangl
Ed Cutrell
Samreen Anjum
Jeffrey Bigham
Chongyan Chen

Location: West 210

Time: Half Day - Morning (0815-1200)

Summary: Our goal for this workshop is to educate researchers about the technological needs of people with vision impairments while empowering researchers to improve algorithms to meet these needs. A key component of this event will be to track progress on four dataset challenges, where the tasks are to answer visual questions, ground answers, detect salient objects, and recognize objects in few-shot learning scenarios. The second key component of this event will be a discussion about current research and application issues, including invited speakers from both academia and industry who will share their experiences in building today's state-of-the-art assistive technologies as well as designing next-generation tools.

Computer Vision for Physiological Measurement

Organizers: Wenjin Wang
Sander Stuijk
Daniel McDuff
Yuzhe Yang

Location: East 14

Time: Half Day - Morning (0800-1200)

Summary: Measuring physiological signals from the human face and body using cameras is an emerging research topic that has grown rapidly in the last decade. Avoiding mechanical contact of skin, remote cameras have been used to measure vital signs (e.g. heart rate, heart rate variability, respiration rate, blood oxygenation saturation, pulse transit time, body temperature, etc.) from an image sequence registering a human skin or body. This leads to contactless, continuous and comfortable heath monitoring, which improves user experience/clinical workflow and eliminates potential risks of infection/contamination caused by contact bio-sensors. Imaging methods for recovering vital signs also present new opportunities for machine vision applications that require better understanding of human physiology (e.g. affective computing and cognitive recognition).The CVPM workshop aims to unite the researchers working in this field, and those who can directly/indirectly benefit from and/or contribute to it (including CV and AI researchers, doctors/clinicians, medical experts and psychologists). Although targeted at computer vision audiences, and aimed at promoting advancements in methods, a unique aspect of this workshop is that it brings a rich set of compelling applications (e.g., from video health monitoring to affective computing to face anti-spoofing and biometric security) that attracts broader audiences from fields beyond computer science.

Learning 3D With Multi-View Supervision

Organizers: Abdullah J. Hamdi, Jinjie Mai
Silvio Giancola, Jesus Zarzar
Guocheng Qian, Matthias Müller
Sara Rojas Martinez, Bernard Ghanem

Location: East Exhibit Hall B

Time: Half Day - Morning (0800-1215)

Summary: Many of the recent advances in 3D vision have focused on the direct approach of applying deep learning to 3D data (e.g., 3D point clouds, meshes, voxels). Another way of using deep learning for 3D understanding is to project 3D into multiple 2D images and apply 2D networks to process the 3D data indirectly. Tackling 3D vision tasks with indirect approaches has two main advantages: (i) mature and transferable 2D computer vision models (CNNs, Transformers, Diffusion, etc.), and (ii) large and diverse labeled image datasets for pre-training (e.g., ImageNet). Furthermore, recent advances in differentiable rendering allow for end-to-end deep learning pipelines that render multi-view images of the 3D data and process the images by CNNs/transformers/diffusion to obtain a more descriptive representation for the 3D data. However, several challenges remain in this multi-view direction, including handling the intersection with other modalities like point clouds and meshes and addressing some of the problems that affect 2D projections like occlusion and view-point selection. We aim to enhance the synergy between multi-view research across different tasks by inviting keynote speakers from across the spectrum of 3D understanding and generation, mixing essential 3D topics (like multi-view stereo) with modern generation techniques (like NeRFs).

Agriculture-Vision: Challenges & Opportunities for Computer Vision in Agriculture

Organizers: Jennifer Hobbs, Melba Crawford
Naira Hovakimyan, Edward Delp
Humphrey Shi, Jing Wu

Location: West 212

Time: Half Day - Morning (0800-1200)

Summary: The 4th Agriculture-Vision Workshop supports the development computer vision techniques to identify field issues to aid farmers in decision making, track crop development at international scales to address poverty and supply chain issues, and enable sustainability efforts to address challenges related to climate change. While agriculture related vision tasks benefit directly from the larger body of research in computer vision, they also require directed research and adaptation of approaches due to the size, complexity, and ambiguity of the available data. This workshop seeks to bring together researchers across disciplines including computer vision, agronomy and crop science, remote sensing, robotics, soil science, climate science, and others.

Multi-Agent Behavior: Properties, Computation and Emergence

Organizers: Markus Marks, Yisong Yue
Jennifer J. Sun, Pietro Perona
Ann Kennedy

Location: West 215-216

Time: Half Day - Morning (0800-1230)

Summary: Interactions between multiple agents can happen on various spatio-temporal scales, from two humans dancing, tens of cars organizing at an intersection, hundreds of fish organizing in a formation to trillions of moving nanoparticles interacting in a tumor environment. In each case the behavior of the agents is shaped by their interactions with other agents in the environment, such that the behavior of an individual cannot be understood in isolation. The purpose of this workshop is to provide a forum for exchanging perspectives on how the behavior of the interacting agents is defined, interpreted, measured, and modeled. A panel of speakers from a variety of disciplines will present their work and discuss the key goals of multi-agent behavior research as it applies to their own field. By identifying common challenges and themes across fields, we aim to foster new cross-disciplinary approaches to the modeling and analysis of multi-agent behavior.

Affective Behavior Analysis In-the-Wild

Organizers: Dimitrios Kollias, Alan S. Cowen
Panagiotis Tzirakis, Stefanos Zafeiriou
Alice Baird

Location: West 306

Time: Half Day - Morning (0800-1230)

Summary: The ABAW Workshop has a unique aspect of fostering cross-pollination of different disciplines, bringing together experts (from academia & industry) and researchers of computer vision and pattern recognition, artificial intelligence, machine learning, HCI, multimedia, robotics and psychology. The diversity of human behavior, the richness of multi-modal data that arises from its analysis, and the multitude of applications that demand rapid progress in this area ensure that our event provides a timely and relevant discussion and dissemination platform.

Scholars and Big Models — How Can Academics Adapt?

Organizers: Anand Bhattad, Angjoo Kanazawa
Unnat Jain, David Forsyth
Sara M. Beery

Location: East Exhibit Hall B

Time: Half Day - Afternoon (1245-1805)

Summary: In the wake of big vision models' success, computer vision has experienced rapid growth and increased attention, raising concerns about its impact on the CVPR academic community. Grad students feel discouraged because they lack access to immense compute in academia that power these big models, while senior researchers witness exponential growth beyond their wildest dreams. We are at a pivotal moment in the history of CVPR, prompting several critical questions:

- How can we discuss and address concerns arising from the rapid changes due to massively distributed training of big models?
- As big models become more prevalent, what can we do to reduce the barriers for entry, open access, and equity of opportunities?
- When the SOTA is being pushed frequently, how can the community support PhD students and researchers not feel discouraged?
- Are a few research directions affecting the questions most of our community focuses on?
- What creative solutions can benefit grad students, assistant professors, senior faculty members, and other community members?
- How can we redesign collaborations, access to computational resources, and curricula to adapt to this evolving landscape?

To address these questions, this workshop features diverse talks and panels from researchers in academia and industry. Our speakers and panelists will share candid thoughts on the challenges and opportunities posed by big models. To improve inclusion, equity, and diversity, we hope to gain a deeper understanding of the current landscape and possible strategies to adapt to and thrive in it.

Photogrammetric Computer Vision

Organizers: Ewelina Rupnik, Jianzhu Huai
Ronny Hänsch, Rongjun Qin
Mozhdeh Shahbazi

Location: West 211

Time: Half Day - Afternoon (1300-1800)

Summary: PCV explores the intersection of photogrammetry and computer vision. While both fields relate to image processing and analysis, their focus is different. Computer vision interprets visual information in a broad sense, and photogrammetry is concerned with the development of methods for engineering purposes such as mapping, surveying, and high-precision metrology. The scope of PCV includes, but is not limited to: feature extraction, matching, and sensor orientation and sensor fusion, Structure from motion and SLAM, stereo (multi-view) and surface reconstruction, 3D point cloud processing, classification, multi-temporal analysis, dynamic scene understanding, 3D scene analysis and semantic segmentation.

More importantly, the workshop provides a forum for collaboration between the computer vision and photogrammetry communities to discuss modern challenges and ideas, propose new and contemporary benchmarks, elaborate on the overlap with machine learning, mathematics, and boost the development in the highly challenging and quickly evolving field of photogrammetric computer vision.

Omnidirectional Computer Vision

Organizers: Kaavya Rekanar, Marc Eder
Ciarán Eising, Pierre Moulon
Li Guan, Varun Ravi Kumar
Stefan Milz, Ganesh Sistu
Jonathan Horgan, Shubhankar Borse
Senthil Yogamani, Fatih Porikli

Location: East 2

Time: Half Day - Afternoon (1300-1800)

Summary: Our objective is to provide a venue for novel research in omnidirectional computer vision with an eye toward actualizing these ideas for commercial or societal benefit. As omnidirectional cameras become more widespread, we want to bridge the gap between the research and application of omnidirectional vision technologies. Omnidirectional cameras are already widespread in a number of application areas such as automotive, surveillance, photography, simulation and other use-cases that benefit from large field of view. More recently, they have garnered interest for use in virtual and augmented reality. We want to encourage the development of new models that natively operate on omnidirectional imagery as well as close the performance gap between perspective-image and omnidirectional algorithms. Our workshop seeks to provide a link between the formative research that supports these advances and the realization of commercial products that leverage this technology. We want to encourage the development of new algorithms and applications for this imaging modality that will continue to drive this engine of progress.

Face Anti-Spoofing Challenge

Organizers: Jun Wan, Hugo Jair Escalante
Ajian Liu, Isabelle Guyon
Sergio Escalera

Location: East 1

Time: Half Day - Afternoon (1330-1730)

Summary: In recent years, the security of face recognition systems has been increasingly threatened. Face Anti-spoofing (FAS) is essential to secure face recognition systems from various attacks. In order to attract researchers and push forward the state of the art in Face Presentation Attack Detection (PAD), we organized three previous editions of Face Anti-spoofing Workshop and Competition, which promoted the algorithms to overcome many challenging problems. However, long-distance face presentation attack based on surveillance is still a threat. Specifically, compared with traditional FAS (e.g., phone unlocking, face payment, and self-service security inspection), FAS in long-distance such as station squares, parks, and self-service supermarkets are equally important, but it has not been sufficiently explored yet. This fourth edition of the Face Anti-Spoofing Workshop and Challenge is to provide continuity to our effort in this relevant problem. Unlike the previous editions, where faces were identified by posing in specific situations at a close distance, the 2023 challenge will focus on more general surveillance and in the wild scenarios, and alleviating the performance degradation of PAD technology in the case of low face resolution, occlusion interference, non-frontal perspective, and other natural person behaviors. Considering the above difficulties and challenges, two datasets are released for this fourth edition for algorithm design and competition promotion: 1) a large-scale High Fidelity Mask dataset based on Surveillance Scenes and 2) a large-scale in-the-wild dataset.

ScanNet 3D Scene Understanding

Organizers: Angela Dai
Angel X. Chang
Manolis Savva
Matthias Niessner

Location: West 205-206

Time: Half Day - Afternoon (1330-1730)

Summary: 3D scene understanding for indoor environments is becoming an increasingly important area. Application domains such as augmented and virtual reality, computational photography, interior design, and autonomous mobile robots all require a deep understanding of 3D interior spaces, the semantics of objects that are present, and their relative configurations in 3D space. We aim to highlight methods and advances not only for traditional 3D semantic scene understanding tasks, but also under modern practical challenges, such as limited data learning on the ScanNet Data-Efficient Benchmark.

This year, we also introduce a new challenge to expand semantic vocabulary by an order of magnitude than previous: the Large-Vocabulary 3D Understanding in the ScanNet200 Challenge.

Machine Visual Common Sense: Perception, Prediction, Planning

Organizers: Yining Hong, Qinhong Zhou
Zhenfang Chen, Chuang Gan
Bo Wu, Joshua Tenenbaum
Mingyu Ding, Antonio Torralba

Location: West 116-117

Time: Half Day - Afternoon (Time TBA)

Summary: Over the years, there have been a variety of visual reasoning tasks that evaluate machines' ability to understand and reason about visual scenes. However, these benchmarks mostly focus on classification of objects and items that exist in a scene. Common sense reasoning – an understanding of what might happen next, or what gave rise to the scene – is often absent in these benchmarks. Humans, on the other hand, are highly versatile, adept in numerous high-level cognition-related visual reasoning tasks that go beyond pattern recognition and require common sense (e.g., physics, causality, functionality, psychology, etc).

In order to design systems with human-like visual understanding of the world, we would like to emphasize benchmarks and tasks that evaluate common sense reasoning across a variety of domains, including but not limited to:

- Intuitive Physics: A general understanding and expectations about the physical world (e.g., how things support, collide, fall, contain, become unstable etc.)
- Intuitive Psychology & Social Science: A basic understanding of inter-relations and interaction of agents; An understanding of instrumental actions (e.g., assistance, imitation, speech etc.); The ability to reason about hidden mental variables that drive observable actions.
- Affordance & Functionality: What actions of agents can be applied to objects; What functions objects provide for the agents.
- Causality & Counterfactual Thinking: Understanding of causes and effects; Mental representations of alternatives to past or future events, actions, or states.

RetailVision – Revolutionizing the World of Retail

Organizers: Ehud Barnea, Zhao Deli
Yosi Keller, Yanheng Wei
Marina Paolanti, Danny Barash

Location: West 215-216

Time: Half Day - Afternoon (1300-1800)

Summary: The rapid development in computer vision and machine learning has caused a major disruption in the retail industry in recent years. In addition to the rise of the web and online shopping, traditional markets also quickly embrace AI-related technology solutions at the physical store level. Following the introduction of computer vision to the world of retail a new set challenges emerged in both the physical and online domains. The physical domain exhibits challenges such as shopper and product interaction detection, detection of products in crowded store displays, fine-grained classification of many visually similar classes, as well as dynamically adapting to changes in data in terms of class appearance variation over time, and new classes that may appear in the images before they are labeled in the dataset. The online domain contains similar challenges, but with their own twist. Product search and recognition is performed on more than 100,000 classes, and also incorporates textual captions describing the products, and text by users during their search. All of these challenges are at the heart of the computer vision community, and this workshop aims to present the progress in these challenges and encourage the forming of a community for retail computer vision.

Multimodal Learning for Earth and Environment

Organizers: Miriam Cha, Morgan J. Schmidt
Gregory Angelides, Nathaniel Maidel
Mark T. Hamilton, Phillip Isola
Andy Soszynski, Taylor Perron
Brandon M. Swenson, William T. Freeman

Location: West 109-110

Time: Half Day - Afternoon (1300-1700)

Summary: The Multimodal Learning for Earth and Environment Workshop (MultiEarth 2023) is the second annual CVPR workshop aimed at leveraging the significant amount of remote sensing data that is continuously being collected to aid in the monitoring and analysis of the health of Earth ecosystems. The goal of the workshop is to bring together the Earth and environmental science communities as well as the multimodal representation learning communities to examine new ways to leverage technological advances in support of environmental monitoring. In addition, through a series of public challenges, the MultiEarth Workshop hopes to provide a common benchmark for remote sensing multimodal information processing. These challenges are focused on the monitoring of the Amazon rainforest and include deforestation estimation, fire detection, cross-modal image translation, and environmental change projection.

Dynamic Scene Reconstruction

Organizers: Armin Mustafa, Christian Richardt, Marco Volino, Adrian Hilton, Dan Casas

Location: West 223-224

Time: Half Day - Afternoon (1330-1720)

Summary: Reconstruction of general dynamic scenes is motivated by potential applications in film and broadcast production together with the ultimate goal of automatic understanding of real-world scenes from distributed camera networks. With recent advances in hardware and the advent of virtual and augmented reality, dynamic scene reconstruction is being applied to more complex scenes with applications in Entertainment, Games, Film, Creative Industries and AR/VR/MR. This workshop aims to give an overview of the advances of computer vision algorithms in dynamic scene reconstruction to the target audience and will identify future challenges.

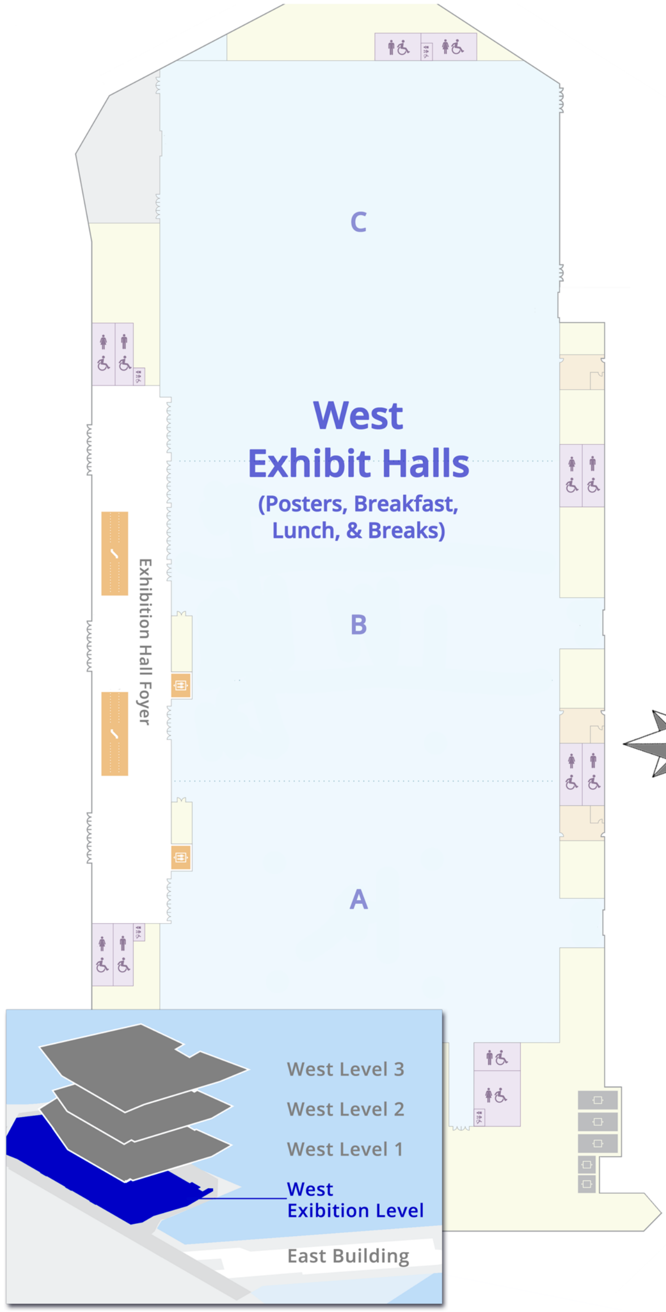
Notes:

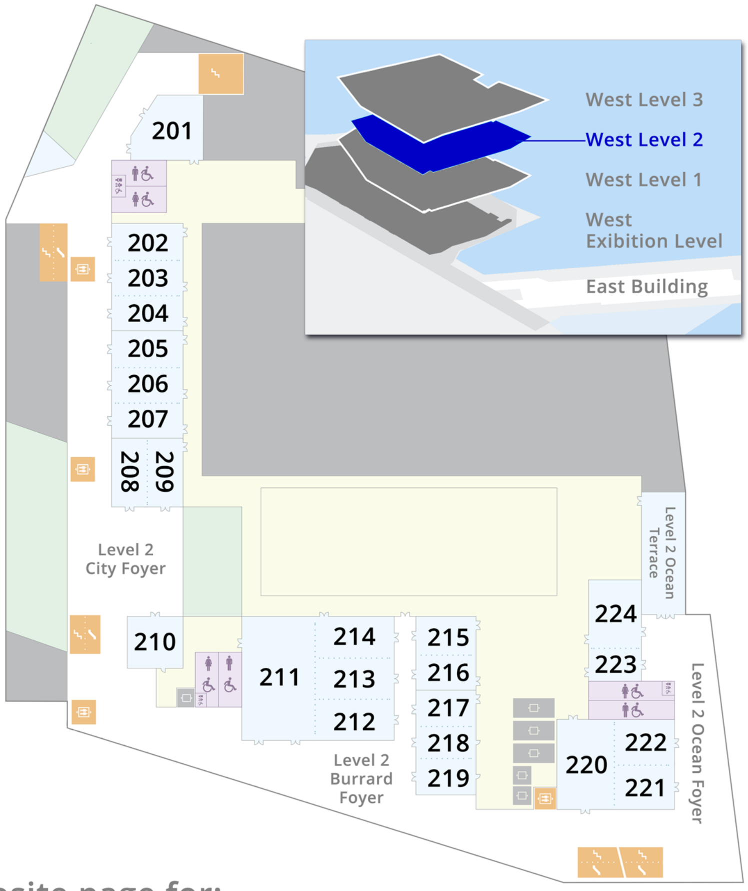
JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

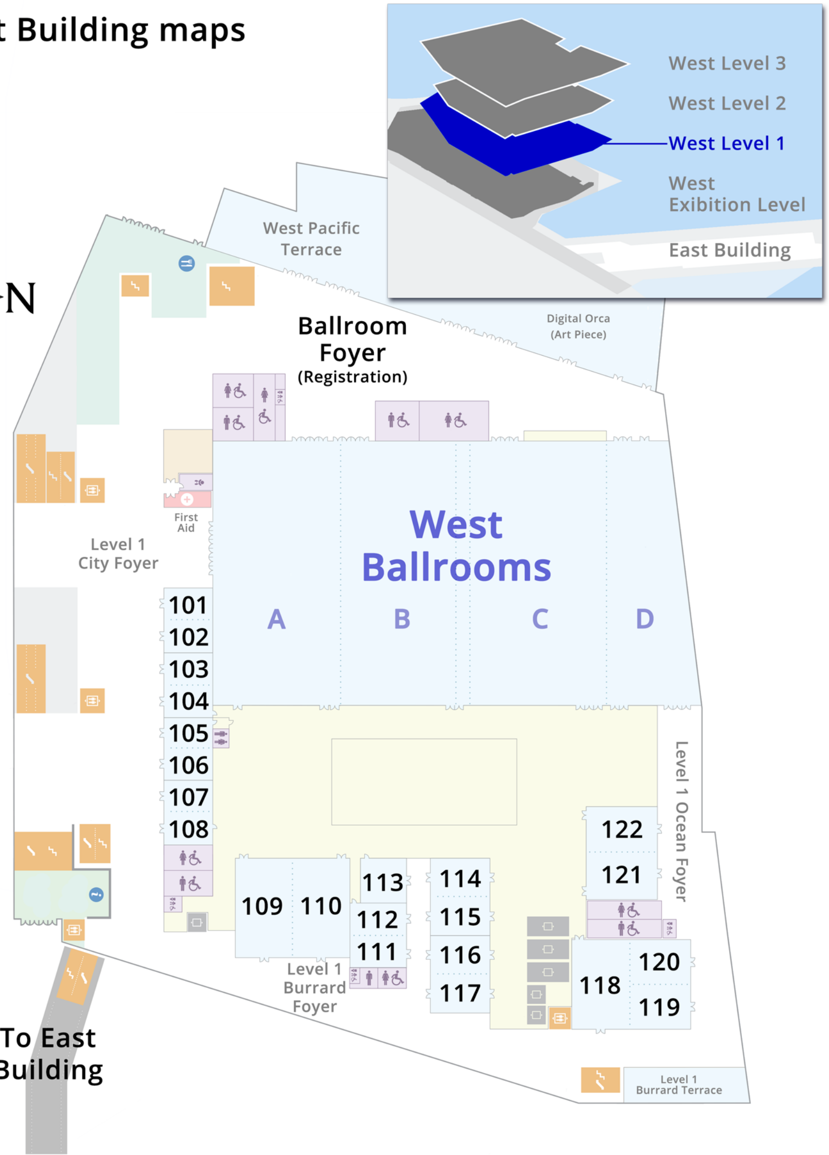
VANCOUVER CONVENTION CENTRE





See opposite page for:

West Level 3 and East Building maps



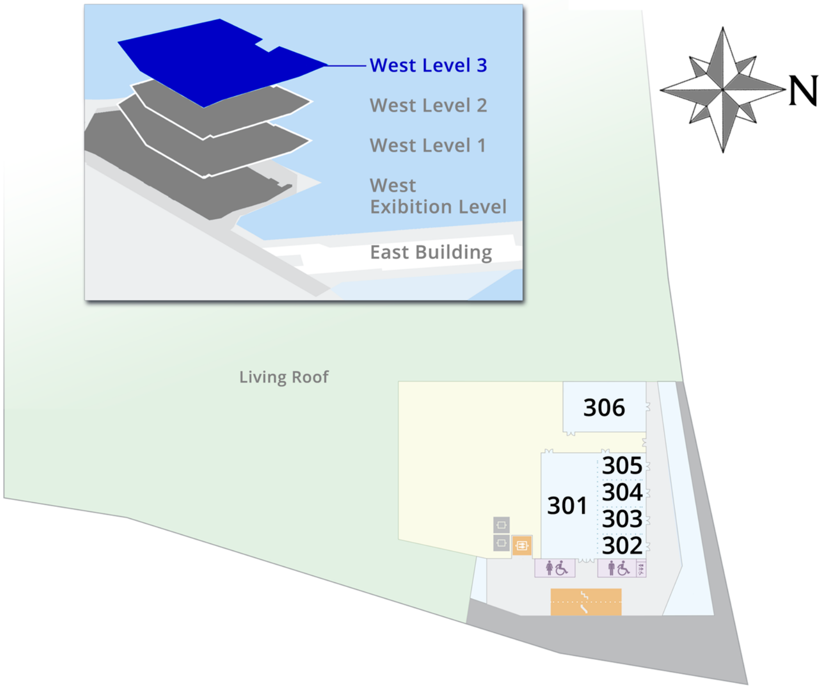
CVPR 2023 Sponsors

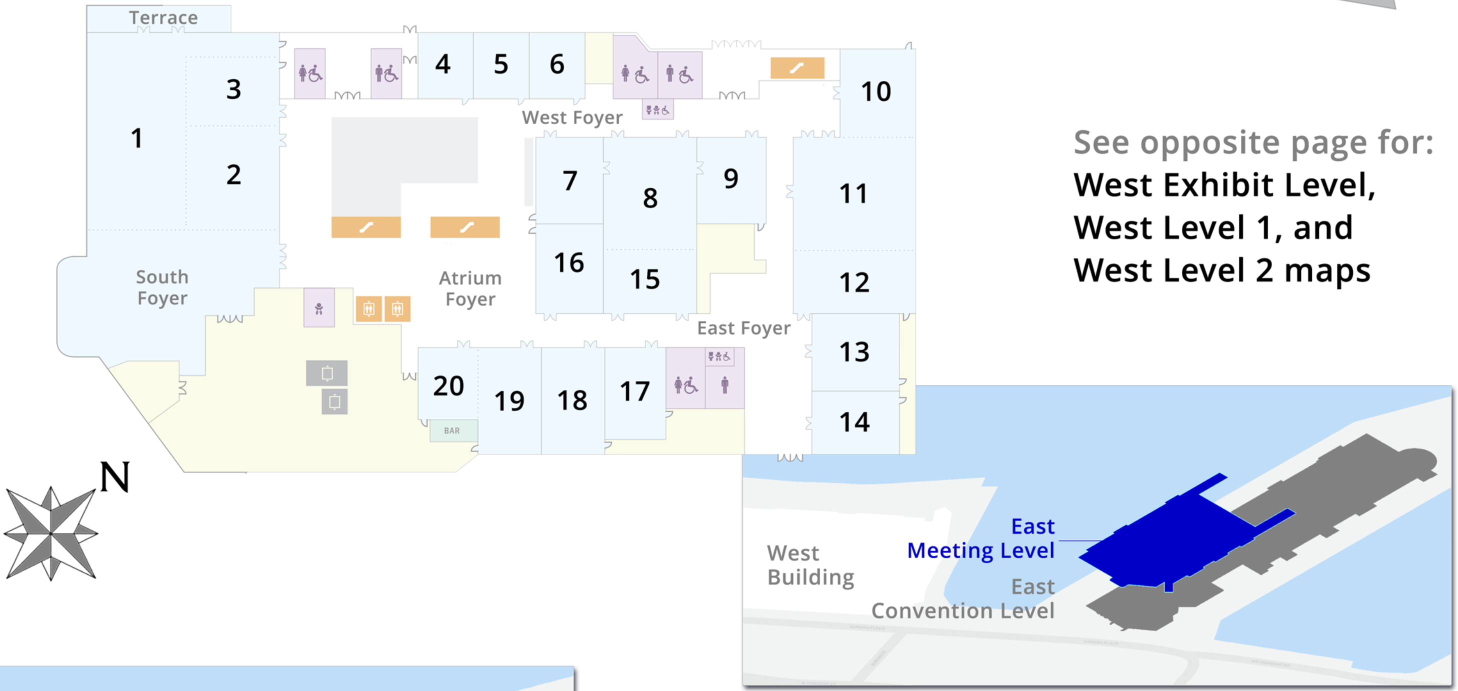
IEEE COMPUTER SOCIETY

CVF

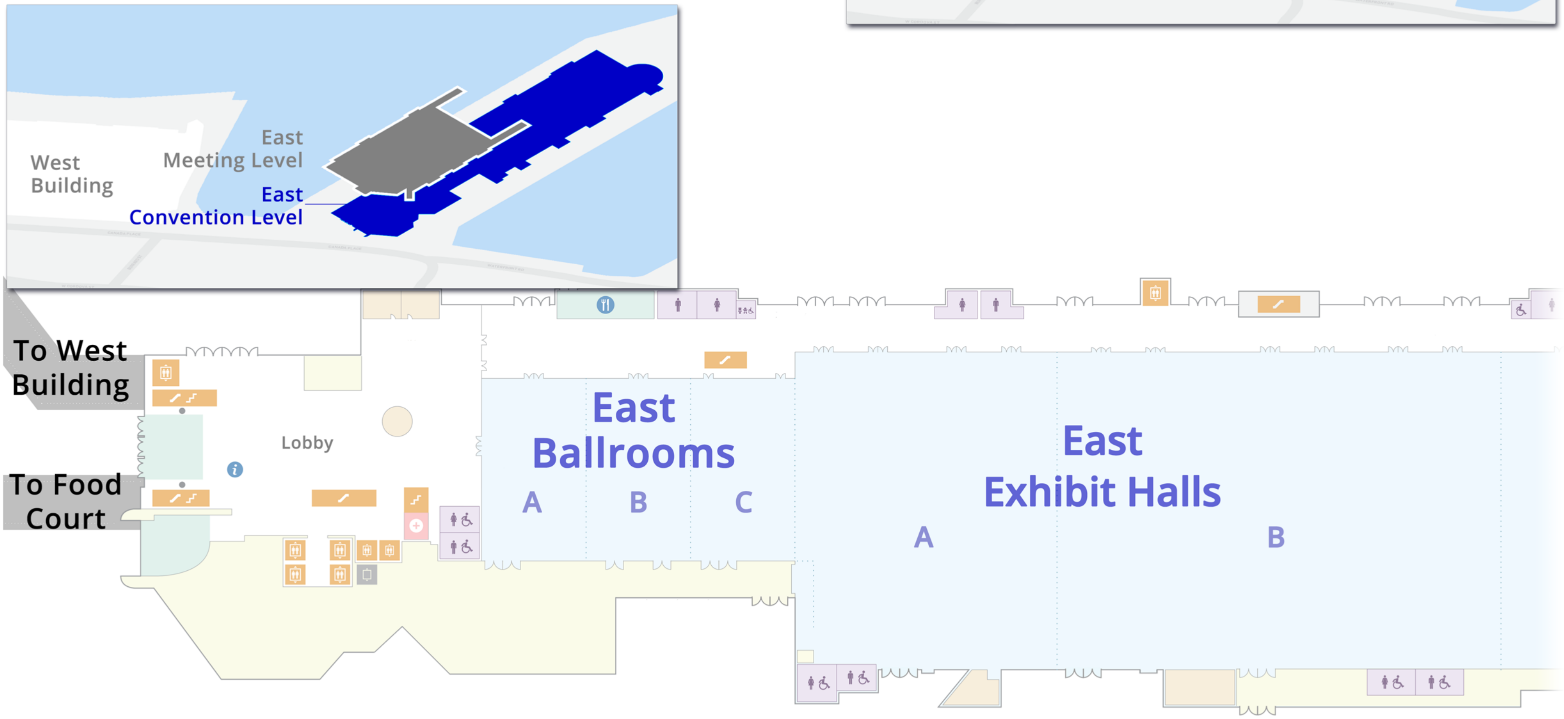
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

VANCOUVER
CONVENTION
CENTRE





See opposite page for:
West Exhibit Level,
West Level 1, and
West Level 2 maps


West
Building
East
Meeting Level
East
Convention Level
To West
Building
To Food
Court
Lobby
East
Ballrooms
A
B
C
East
Exhibit Halls
A
B