CVPR 2024

IEEE/CVF Conference on Computer Vision and Pattern Recognition

PLATINUM SPONSORS

amazon | science
AMD

ByteDance
captions
Google
Hyperbolic
intel
Lambda
美团
Meta
Microsoft
Qualcomm
SONY
Boston Dynamics AI INSTITUTE

GOLD SPONSORS

Adobe
Alibaba Cloud
ANT RESEARCH
Baidu 百度
comet
Gatik
HYUNDAI
Jane Street
LATITUDE
scale
CHEN TIANQIAO & CHRISSY INSTITUTE
VOXEL51
WAYVE
Weights & Biases
WYZE
ZOOX

SILVER SPONSORS

AKOOL
ANDURIL
ASML
B GARAGE

Disney Research
helm.ai
Determined AI
iMerit
kitware
KUAISHOU
LG
Lightning AI
Creators of PyTorch Lightning
NEC
NEC Laboratories America
nexdata
oppo

PROPHESEE
METAVISION TECHNOLOGIES
sapien
SIMULAR

synthesia
Tencent
WAYMO

Welcome to the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition in Seattle, Washington! 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 keynotes, oral and poster presentations, panels, tutorials, workshops, demos, exhibitions, and social events, all aimed at providing attendees with an exciting and enriching experience. CVPR 2024 is primarily an in-person conference, but for those who are unable to join us physically, we are pleased to offer a virtual component that will provide access to conference papers, posters, videos, and talks.

CVPR 2024 received 11,532 valid paper submissions, a 26% increase from CVPR 2023. The review process was managed by the 6 Program Co-Chairs, 24 Senior Area Chairs, and 477 Area Chairs. During the review phase, each paper received at least 3 reviews from a pool of 9,872 reviewers. As in prior years, after receiving these initial reviews, authors had the opportunity to submit a rebuttal to the reviews. The process concluded with 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, 2,719 papers were accepted, for a 23.6% overall acceptance rate. In keeping with the CVPR tradition, the PCs did not pre-define any target acceptance rate or number of papers to be accepted; the resulting acceptance rate reflects the community consensus, and is consistent with past CVPRs.

All of the 2,719 accepted papers were invited to present posters at CVPR. In addition, 90 (3.3%) papers were selected to be presented as oral talks, based on nominations from Area Chairs, and 324 (11.9%) were selected by ACs to be "highlights" because of their high quality and potential impact. The highlights are flagged with a special annotation in the program. ACs nominated 24 papers to be best paper award candidates, from which a committee convened by the PCs selected the award winners. The award winners will be announced during the conference.

CVPR 2024 brings back the tradition of oral presentations in a three-track configuration. Nevertheless, we kept many of the innovations of CVPR 2023, including single-track panel discussions, "highlights" to indicate top-rated papers, the use of OpenReview for paper submission and management, and the role of Senior Area Chair to help oversee the review process.

We would like to thank everyone involved in making CVPR 2024 a success. This includes the organizing committee, the Senior Area Chairs, Area Chairs, and reviewers, the authors, the demo session participants, and the donors and exhibitors. David Forsyth's service as Senior Advisor to the Program Chairs was incredibly helpful. 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, Mike Weil and Hall Erickson for handling sponsorships and the exhibition, and Luba Elliot as our inaugural AI Art Coordinator. Last but not least, we thank all of you for attending CVPR 2024 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 Seattle during the conference.

Enjoy CVPR 2024. We look forward to meeting you in person!

Program Chairs

Ali Farhadi (University of Washington)
David Crandall (Indiana University)
Imari Sato (National Institute of Informatics)
Jianxin Wu (Nanjing University)
Robert Pless (George Washington University)
Zeynep Akata (University of Tübingen)

General Chairs

Octavia Camps (Northeastern University)
Ramin Zabih (Cornell University)
Rita Cucchiara (University of Modena and Reggio Emilia)
Sudeep Sarkar (University of South Florida)
Walter Scheirer (University of Notre Dame)

General Chairs
Octavia Camps (Northeastern University)
Ramin Zabih (Cornell)
Rita Cucchiara (University of Modena and Reggio Emilia)
Sudeep Sarkar (University of South Florida)
Walter Scheirer (University of Notre Dame)

Program Chairs
Ali Farhadi (University of Washington)
David Crandall (Indiana University)
Imari Sato (National Institute of Informatics)
Jianxin Wu (Nanjing University)
Robert Pless (George Washington University)
Zeynep Akata (University of Tübingen)

Senior Advisor to the Program Chairs
David Forsyth (University of Illinois at Urbana-Champaign)

Technical Chair
Yoshitomo Matsubara (Spiffy AI)

Tutorial Chairs
Katarina Doctor (U.S. Naval Research Lab)
Vitomir Struc (University of Ljubljana)

Workshop Chairs
Abhinav Shrivastava (University of Maryland)
Andrew Owens (University of Michigan)
Antitza Dantcheva (Inria)
Luisa Verdoliva (University Federico II of Naples)

DEI Chairs
Adriana Kovashka (Pitt)
CJ Taylor (Amazon AWS)
Michael King (Fit)
Roni Sengupta (UNC)
Sara Beery (MIT)
Shuran Song (Columbia University)
Tamara L Berg

Conference Ombud
Angjoo Kanazawa (University of California Berkeley)
Derek Hoiem (University of Illinois at Urbana-Champaign)

Demonstration Chairs
Sathyanarayanan N. Aakur (Auburn University)
Shu Kong (University of Macau, Texas A&M University)
Senior PAMI-TC Ombud
David Forsyth (University of Illinois at Urbana-Champaign)
Linda Shapiro (UW Reality Lab University of Washington)

AI Art Curator
Luba Elliott (Independent Curator)

Conference Producer
Nicole Finn (c to c events)

Local Chairs
Ira Kemelmacher-Shlizerman (University of Washington)
Ranjay Krishna (University of Washington)

Publications Chair
Eric Mortensen

Publicity Chairs
Abby Stylianou (Saint Louis University)
Boqing Gong (Google)
Jia-Bin Huang (University of Maryland, College Park)
Kosta Derpanis (York University/Samsung)
Shenghua Gao (ShanghaiTech University)

Finance Chair
Gerard Medioni (Amazon)

Accessibility Chair
Danna Gurari (University of Colorado, Boulder)

Doctoral Consortium Chairs
Aparna Bharati (Lehigh University)
Nathan Jacobs (Washington University in St. Louis)

Website Chair
Mauricio Pamplona Segundo (University of South Florida)

Workflow Chair
Zhenyu (Sherry) Xue (CVPR)

Web Developer
Lee Campbell (Eventhosts)

Virtual Platform Chair
Andreas Geiger (University of Tübingen)

Corporate Relations Chairs
Brian Clipp (Kitware)
Victor Fragoso (Microsoft)

Social Activities Chairs
Giovanni Maria Farinella (University of Catania, Italy)
Vítor Albiero (Meta)
Yale Song (Meta)

MESSAGE FROM THE WORKSHOP CHAIRS

To make it easier to navigate the growing number of CVPR workshops, we have grouped the workshops together into thematic "tracks." The workshops within each track cover closely related topics, and we've also tried our best to avoid having scheduling conflicts between them. We hope you enjoy the workshops!

Workshop Chairs

Antitza Dantcheva, Luisa Verdoliva, Andrew Owens, Abhinav Shrivastava

TABLE OF CONTENTS

Track on Open World Learning

Track on Egocentric & Embodied AI

Track on Video Understanding

Track on Human Understanding

Track on 3D Vision

Track on Neural Rendering

Track on Physics, Graphics, Geometry, AR/VR/MR

Track on Contemporary Discussions and Community Building

Track on Computational Photography

Track on Mobile and Embedded Vision

Track on Content Creation

Track on Biometrics and Forensics

Track on Responsible and Explainable AI

Track on Autonomous Driving

Track on Assistive Technology

Track on Urban Environments

Track on Science Applications

Track on Medical Vision

Track on Foundation Models

Track on Generative Models

Track on Synthetic Data

Track on Applications

Sunday, June 16

11:00 –20:00 **Registration / Badge Pickup** (Summit Lobby)

Monday, June 17

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 2024 Tutorials page.

7:00–17:00 **Registration / Badge Pickup** (Summit Lobby)

7:00–17:00 **Press Room** (Summit 340)

7:00–17:00 **Mother's Room** (Summit 341-adjacent and Summit 441-adjacent)

7:00–17:00 **Prayer or Quiet Room** (Upon Request)

7:00–9:00 **Breakfast** (Summit ExHall 1-2)

8:00–18:00 **TUTORIALS / WORKSHOPS**

10:00–11:00 **Coffee Break** (Arch 4E)

12:00–13:45 **Lunch Summit** (Summit ExHall 1-2)

15:00–16:00 **Coffee Break** (Arch 4E)

TUTORIALS

Tutorial: Deep Stereo Matching in the Twenties

Organizers: Matteo Poggi, Fabio Tosi
Date: Monday, June 17
Time: 9:00 AM–12:00 PM
Location: Arch 213

Summary: For decades, stereo matching has been approached by developing hand-crafted algorithms, focused on measuring the visual appearance between local patterns in the two images and propagating this information globally. Since 2015, deep learning led to a paradigm shift in this field, driving the community to the design of end-to-end deep networks capable of matching pixels. The results of this revolution brought stereo matching to a whole new level of accuracy, yet not without any drawbacks. Indeed, some hard challenges remained unsolved by the first generation of deep stereo models, as they were often not capable of properly generalizing across different domains -- e.g., from synthetic to real, from indoor to outdoor -- or dealing with high-resolution images. This was, however, three years ago. These and other challenges have been faced by the research community in the Twenties, making deep stereo matching even more mature and suitable to be a practical solution for everyday applications. For instance, now we have networks capable of generalizing much better from synthetic to real images, as well as handling high-resolution images or even estimating disparity correctly in the presence of non-Lambertian surfaces -- known to be among the ill-posed challenges for stereo. Accordingly, in this tutorial, we aim at giving a comprehensive overview of the state-of-the-art of deep stereo matching, which architectural designs have been crucial to reach this level of maturity and how to select the best solution for estimating depth from stereo in real applications.

Tutorial: Disentanglement and Compositionality in Computer Vision

Organizers: Xin Jin, Tao Yang, Yue Song, Xingyi Yang, Wenjun (Kevin) Zeng, Nicu Sebe, Xinchao Wang, Shuicheng Yan
Date: Monday, June 17
Time: 9:00 AM–12:00 PM
Location: Arch 2B

Summary: This tutorial aims to explore the concepts of disentanglement and compositionality in the field of computer vision. These concepts play a crucial role in enabling machines to understand and interpret visual information with more sophistication and human-like reasoning. Participants will learn about advanced techniques and models that allow for the disentanglement of visual factors in images and the compositionality of these factors to produce more meaningful representations. All in all, Disentanglement and Composition are believed to be one of the possible ways for AI to fundamentally understand the world, and eventually achieve Artificial General Intelligence (AGI).

Tutorial: Machine Unlearning in Computer Vision: Foundations and Applications

Organizers: Sijia Liu , Yang Liu, Nathalie Baracaldo, Eleni Triantafillou
Date: Monday, June 17
Time: 9:00 AM–12:00 AM
Location: Arch 305

Summary: This tutorial aims to offer a comprehensive understanding of emerging machine unlearning (MU) techniques. These techniques are designed to accurately assess the impact of specific data points, classes, or concepts on model performance and efficiently eliminate their potentially harmful influence within a pre-trained model, in response to users' unlearning requests. With the recent shift to foundation models, MU has become indispensable, as re-training from scratch is prohibitively costly in terms of time, computational resources, and finances. Consequently, the field has expanded beyond the realm of security and privacy (SP) to include the removal of toxic content, copyright material, harmful information, and personally identifying data. Despite increasing research interest, MU for vision tasks remains significantly underexplored compared to its prominence in the SP field. Therefore, it is crucial to meticulously review, thoroughly explore, and comprehensively survey MU for computer vision (CV) through this tutorial. Within this tutorial, we will delve into the algorithmic foundations of MU methods, including techniques such as localization-informed unlearning, unlearning-focused finetuning, and vision model-specific optimizers. We will provide a comprehensive and clear overview of the diverse range of applications for MU in CV. Furthermore, we will emphasize the importance of unlearning from an industry perspective, where modifying the model during its life-cycle is preferable to re-training it entirely, and where metrics to verify the unlearning process become paramount. Our tutorial will furnish the general audience with sufficient background information to grasp the motivation, research progress, opportunities, and ongoing challenges in MU.

Tutorial: Recent Advances in Vision Foundation Models

Organizers: Zhengyuan Yang, Linjie Li, Zhe Gan, Chunyuan Li, Jianwei Yang
Date: Monday, June 17
Time: 9:00 AM–5:00 PM
Location: Summit 437-439

Summary: This tutorial covers the advanced topics in designing and training vision foundation models, including the state-of-the-art approaches and principles in (i) learning vision foundation models for multimodal understanding and generation, (ii) benchmarking and evaluating vision foundation models, and (iii) agents and other advanced systems based on vision foundation models.

Tutorial: SCENIC: An Open-Source Probabilistic Programming System for Data Generation and Safety in AI-Based Autonomy

Organizers: Edward Kim, Eric Vin, Kimin Lee, Jinkyu Kim, Hazem Torfah, Marcell Vazquez-Chanlatte, Parasara Sridhar Duggirala, Necmiye Ozay
Date: Monday, June 17
Time: 9:00 AM–12:00 PM
Location: Arch 307-308

Summary: Autonomous systems, such as self-driving cars or intelligent robots, are increasingly operating in complex, stochastic environments where they dynamically interact with multiple entities (human and robot). There is a need to formally model and generate such environments in simulation, for use cases that span synthetic training data generation and rigorous evaluation of safety. In this tutorial, we provide an in-depth tutorial on Scenic, a simulator-agnostic probabilistic programming language to model complex multi-agent, physical environments with stochasticity and spatio-temporal constraints. Scenic has been used in a variety of domains such as self-driving, aviation, indoor robotics, multi-agent systems, and augmented/virtual reality. Using Scenic and associated open source tools, one can (1) model and sample from distributions with spatial and temporal constraints, (2) generate synthetic data in a controlled, programmatic fashion to train and test machine learning components, (3) reason about the safety of AI-enabled autonomous systems, (4) automatically find edge cases, (5) debug and root-cause failures of AI components including for perception, and (6) bridge the sim-to-real gap in autonomous system design. We will provide a hands-on tutorial on the basics of Scenic and its applications, how to create Scenic programs and your own new applications on top of Scenic, and to interface the language to your simulator/renderer of choice.

Tutorial: Efficient Homotopy Continuation for Solving Polynomial Systems in Computer Vision Applications

Organizers: Ben Kimia, Tim Duff, Ricardo Fabbri, Hongyi Fan
Date: Monday, June 17
Time: 1:30 PM–6:00 PM
Location: Summit 447

Summary: Minimal problems and their solvers play an important role in RANSAC-based approaches to several estimation problems in vision. Minimal solvers solve systems of equations, depending on data, which obey a "conservation of number principle": for sufficiently generic data, the number of solutions over the complex numbers is constant. Homotopy continuation (HC) methods exploit not just this conservation principle, but also the smooth dependence of solutions on problem data. The classical solution of polynomial systems using Grobner basis, resultants, elimination templates, etc. has been largely successful in smaller problems, but these methods are not able to tackle larger polynomials systems with a larger number of solutions. While HC methods can solve these problems, they have been notoriously slow. Recent research by the presenters and other researchers has enabled efficient HC solvers with the ability for real-time solutions. The main objective of this tutorial is to make this technology more accessible to the computer vision community. Specifically, after an overview of how such methods can be useful for solving problems in vision (e.g., absolute/relative pose, triangulation), we will describe some of the basic theoretical apparatus underlying HC solvers, including both local and global "probability-1" aspects. On the practical side, we will describe recent advances enabled by GPUs, learning-based approaches, and how to build your own HC-based minimal solvers.

Tutorial: Geospatial Computer Vision and Machine Learning for Large-Scale Earth Observation Data

Organizers: Orhun Aydin, Philipe Dias, Dalton Lunga
Date: Monday, June 17
Time: 1:30 PM–5:00 PM
Location: Summit 448

Summary: The 5Vs of big data, volume, value, variety, velocity, and veracity pose immense opportunity and challenges on implementing local and planet-wide solution from Earth observation (EO) data. EO data, residing at the center of various multidisciplinary problems, primarily obtained through satellite imagery, aerial photography, and UAV-based platforms. Understanding Earth Observation data unlocks this immense data source to address planet-scale problems with computer vision and machine learning techniques for geospatial analysis. This workshop introduces current EO data sources, problems, and image-based analysis techniques. The most recent advances in data, models, and open-source analysis ecosystem related to computer vision and deep learning for EO data will be introduced.

Tutorial: Object-centric Representations in Computer Vision

Organizers: Yanwei Fu, Tong He, Tianjun Xiao, Ke Fan, Francesco Locatello

Date: Monday, June 17

Time: 1:30 PM–6:00 PM

Location: Elliott Bay

Summary: This tutorial discusses the evolution of object-centric representation in computer vision and deep learning. Initially inspired by decomposing visual scenes into surfaces and objects, recent developments focus on learning causal variables from high-dimensional observations like images or videos. The tutorial covers the objectives of OCL, its development, and connections with machine learning fields, emphasizing object-centric approaches, especially in unsupervised segmentation. Advances in encoder, decoder, and self-supervised learning objectives are explored, with a focus on real-world applications and challenges. The tutorial also introduces open-source tools and showcases breakthroughs in video-based object-centric learning. This tutorial will have four talks covering the basic ideas, learning good features for object-centric learning, video based object-centric representation, and more diverse real-world applications.

Tutorial: Robustness at Inference: Towards Explainability, Uncertainty, and Intervenability

Organizers: Ghassan AlRegib, Mohit Prabhushankar

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Summit 440-441

Summary: Neural networks provide generalizable and task independent representation spaces that have garnered widespread applicability in image understanding applications. The complicated semantics of feature interactions within image data has been broken down into a set of non-linear functions, convolution parameters, attention, as well as multi-modal inputs among others. The complexity of these operations has introduced multiple vulnerabilities within neural network architectures. These vulnerabilities include adversarial samples, confidence calibration issues, and catastrophic forgetting among others. Given that AI promises to herald the fourth industrial revolution, it is critical to understand and overcome these vulnerabilities. Doing so requires creating robust neural networks that drive the AI systems. Defining robustness, however, is not trivial. Simple measurements of invariance to noise and perturbations are not applicable in real life settings. In this tutorial, we provide a human-centric approach to understanding robustness in neural networks that allow AI to function in society. Doing so allows us to state the following: 1) All neural networks must provide contextual and relevant explanations to humans, 2) Neural networks must know when and what they don't know, 3) Neural Networks must be amenable to being intervened upon by humans at decision-making stage. These three statements call for robust neural networks to be explainable, equipped with uncertainty quantification, and be intervenable.

Notes:

Tutorial: Contactless AI Healthcare using Cameras and Wireless Sensors

Organizers: Wenjin Wang, Daniel Mcduff, Xuyu Wang
Date: Monday, June 17
Time: 2:00 PM–5:00 PM
Location: Arch 307-308

Summary: Understanding people and extracting health-related metrics is an emerging research topic in computer vision that has grown rapidly recently. Without the need of any physical contact of the human body, cameras have been used to measure vital signs remotely (e.g. heart rate, heart rate variability, respiration rate, blood oxygenation saturation, pulse transit time, body temperature, etc.) from an image sequence of the skin or body, which leads to contactless, continuous and comfortable heath monitoring. The use of cameras also enables the measurement of human behaviors/activities and high-level visual semantic/contextual information leveraging computer vision and machine learning techniques. Understanding of the environment around the people is also a unique advantage of cameras compared to the contact bio-sensors (e.g., wearables), which facilitates better understanding of humans and scene for health monitoring. In addition to the camera based approach, Radio Frequency (RF) based methods for health monitoring have also been proposed, using Radar, WiFi, RFID, and acoustic signals. The contactless monitoring of cameras and RF 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, called "AI health monitoring". In this tutorial, we will give an overview of recent works in this emerging direction.

Tutorial: Edge AI in Action: Practical Approaches to Developing and Deploying Optimized Models

Organizers: Fabricio Batista Narcizo, Elizabete Munzlinger, Anuj Dutt, Shan Shaffi, Narsi Reddy
Date: Monday, June 17
Time: 2:00 PM–5:30 PM
Location: Summit 446

Summary: Edge AI is the application of AI on edge devices, i.e., devices at the periphery of a network, such as cameras, sensors, drones, and robots. Edge AI enables these devices to perform AI tasks autonomously without relying on a connection to the cloud or a central server. This brings higher speed, lower latency, greater privacy, and lower power consumption. However, edge AI also poses many challenges and opportunities for model development and deployment, such as size reduction, compression, quantization, and distillation. This tutorial will provide clear and practical guidance on developing and deploying optimized models for edge AI, emphasizing their real-world applications. Our comprehensive approach covers theoretical and technical aspects, best practices, and real-world case studies. Our primary focus will be on computer vision and deep learning models, which are highly relevant to edge AI applications. Throughout the tutorial, we will demonstrate the use of various tools and frameworks, including TensorFlow, PyTorch, ONNX, OpenVINO, and Qualcomm SNPE. Additionally, we will provide concrete examples of multi-modal AI applications, including gaze correction, body segmentation, hand gesture recognition, and sound localization. These applications leverage input sources like video and sound to create highly interactive and immersive edge AI experiences. Our presentation will encompass developing and deploying multi-modal AI models on Jabra cameras. Furthermore, we will explore integrating with other devices, such as Luxonis OAK-1 MAX and NVIDIA Jetson Nano Developer Kit.

Notes:

WORKSHOPS

Efficient Large Vision Models

Organizers: Amirhossein Habibian, Fatih Porikli, Auke Wiggers, Yun Raymond Fu, Chuo-Ling Chang, Yapeng Tian, Wenming Yang
Date: Monday, June 17
Time: 8:00 AM–12:35 PM
Location: Summit 420-422

Summary: Large vision models (LVMs) are becoming the foundation for many computer vision tasks. CLIP, DINO, SAM and their variations are effectively used to solve various downstream tasks across different image distributions without any fine-tuning. Diffusion based text-to-image generative models such as DALL-E and StableDiffusion have shown great capabilities in image generation, editing and enhancement tasks. The representation capacity of transformer architectures, and their huge number of parameters, empowers LVMs to learn from massive datasets through self-supervised learning without requiring manual annotation. However, this comes at a high computational cost that restricts adaptation of LVMs to settings with limited computational resources. This workshop focuses on enhancing the computational efficiency of LVMs, with the aim of broadening their accessibility to a wider community of researchers and practitioners. We believe that exploring ideas for efficient adaptation of LVMs to downstream tasks and domains without the need for intensive model training and fine-tuning allows the community to conduct research with a limited compute budget. Furthermore, accelerating the inference of LVMs enables adaptation for real-time applications on low compute platforms, including vehicles and phones.

Computer Vision for Mixed Reality

Organizers: Rakesh Ranjan, Peter Vajda, Xiaoyu Xiang, Vikas Chandra, Andrea Colaco
Date: Monday, June 17
Time: 8:00 AM–12:45 PM
Location: Summit 332

Summary: Virtual Reality (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-3 and Apple Vision Pro, now users can build deeply immersive experiences which mix the virtual and the real world into one, often also called Mixed Reality (MR). MR 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, augmented the environment with virtual objects and effects, and all in real-time. We aim to 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.

Domain adaptation, Explainability and Fairness in AI for Medical Image Analysis

Organizers: Dimitris Metaxas, Stefanos Kollias, Xujiong Ye, Francesco Rundo, Dimitrios Kollias
Date: Monday, June 17
Time: 8:00 AM–1:00 PM
Location: Summit 347-348

Summary: The DEF-AI-MIA Workshop focuses on domain adaptation, explainability, fairness, for trustworthiness in AI-enabled medical imaging for digital pathology and radiology. Papers examine: use of self-supervised and unsupervised methods to enforce shared patterns emerging directly from data; development strategies to leverage few (or partial) annotations; interpretability in both model development and/or results obtained; generalizability to data coming from multi-centers, multi-modalities or multi-diseases; robustness to out of distribution data. A Competition is organized focusing on domain adaptation based medical image diagnosis.

AI City Challenge

Organizers: Shuo Wang, Zheng Tang, David Anastasiu, Ming-Ching Chang, Liang Zheng, Anuj Sharma, Pranamesh Chakraborty, Norimasa Kobori, Jun-Wei Hsieh, Rama Chellappa
Date: Monday, June 17
Time: 8:00 AM–5:30 PM
Location: Summit 444

Summary: The eighth AI City Challenge highlighted the convergence of computer vision and artificial intelligence in areas like retail, warehouse settings, and Intelligent Traffic Systems (ITS), presenting significant research opportunities. The 2024 edition featured five tracks, attracting unprecedented interest from 726 teams in 47 countries and regions. Track 1 dealt with multi-target multi-camera (MTMC) people tracking, highlighting significant enhancements in camera count, character number, 3D annotation, and camera matrices, alongside new rules for 3D tracking and online tracking algorithm encouragement. Track 2 introduced dense video captioning for traffic safety, focusing on pedestrian accidents using multi-camera feeds to improve insights for insurance and prevention. Track 3 required teams to classify driver actions in a naturalistic driving analysis. Track 4 explored fish-eye camera analytics using the FishEye8K dataset. Track 5 focused on motorcycle helmet rule violation detection. The challenge utilized two leaderboards to showcase methods, with participants setting new benchmarks, some surpassing existing state-of-the-art achievements.

Multimodal Algorithmic Reasoning Workshop

Organizers: Anoop Cherian, Suhas Lohit, Kuan-Chuan Peng, Honglu Zhou, Moitreya Chatterjee, Kevin Smith, Tim Marks, Joanna Matthiesen, Joshua Tenenbaum

Date: Monday, June 17

Time: 8:25 AM–12:15 PM

Location: Summit 320

Summary: In this workshop, we gather researchers working in neural algorithmic learning, multimodal reasoning, and cognitive models of intelligence to showcase their cutting-edge research, discuss the latest challenges, as well as bring to the forefront problems in perception and language modeling that are often overlooked but are pivotal in achieving true artificial general intelligence. An emphasis of this workshop is on the emerging topic of multimodal algorithmic reasoning, where a reasoning agent is required to automatically deduce new algorithms/procedures for solving real-world tasks, e.g., algorithms that use multimodal foundational models for analysis, synthesis, and planning, new approaches towards solving challenging vision-and-language mathematical (Olympiad type) reasoning problems, procedures for using tools in robotic manipulation, etc. We hope to deep dive into this exciting topic at the intersection of multimodal learning and cognitive science to understand what we have achieved thus far in machine intelligence and what we are lacking in relation to the human way of thinking -- through talks from outstanding researchers and faculty that could inspire the audience to search for the missing rungs on the ladder to true intelligence.

SyntaGen: Harnessing Generative Models for Synthetic Visual Datasets

Organizers: Khoi Nguyen, Anh Tran, Binh-Son Hua, Supasorn Suwajanakorn, Yi Zhou

Date: Monday, June 17

Time: 8:25 AM–12:35 PM

Location: Summit 423-425

Summary: The field of computer vision has undergone a significant transformation in recent years with the advancement of generative models, particularly text-to-image models such as Imagen, Stable Diffusion, and DALLE-3. These models have enabled the creation of synthetic visual datasets that are highly realistic and diverse, complete with annotations and rich variations. These datasets have proven to be extremely valuable in training and evaluating various computer vision algorithms, including object detection and segmentation, representation learning, and scene understanding. The SyntaGen workshop will act as a crucible for an inclusive exchange of ideas, practical insights, and collaborative explorations. By convening experts and enthusiasts from various corners of the field, it strives to propel the development of generative models and synthetic visual datasets to new heights. Through informative talks, poster sessions, paper presentations, and vibrant panel discussions, this workshop endeavors to lay the foundation for innovative breakthroughs that bridge the realms of generative models and computer vision applications.

Scene Graphs and Graph Representation Learning

Organizers: Azade Farshad, Iro Armeni, Federico Tombari, Ehsan Adeli, Nassir Navab

Date: Monday, June 17

Time: 8:30 AM–12:00 PM

Location: Summit 322

Summary: The SG2RL workshop focuses on the topic of scene graphs and graph representation learning for visual perception applications in different domains. Through a series of keynote talks, the audience will learn about defining, generating, and predicting scene graphs, as well as about employing them for other tasks. Oral presentations of accepted submissions to the workshop will further enrich discussed topics with state-of-the-art advancements and engage the community. The objective is for attendees to learn about current developments and application domains of scene graphs and graph representation learning, as well as to draw inspiration and identify commonalities across these domains. Furthermore, this workshop will create an opportunity to discuss limitations, challenges, and next steps from research, practical, and ethical perspectives.

Face Anti-Spoofing

Organizers: Jun Wan, Ajian Liu, Jiankang Deng, Sergio Escalera, Hugo Jair Escalante, Isabelle Guyon, Zhen Lei, Shengjin Wang, Ya-Li Li

Date: Monday, June 17

Time: 8:30 AM–12:00 PM

Location: Arch 201

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 primarily from various attacks. In order to attract researchers and push forward the state of the art in Face Presentation Attack Detection (PAD), we organized four editions of Face Anti-spoofing Workshop and Competition at CVPR 2019, CVPR 2020, ICCV 2021, and CVPR 2023, which together have attracted more than 1200 teams from academia and industry, and greatly promoted the algorithms to overcome many challenging problems. In addition to physical presentation attacks (PAs), such as printing, replay, and 3D mask attacks, digital face forgery attacks (FAs) are still a threat that seriously endangers the security of face recognition systems. FAs aim to attack faces using digital editing at the pixel level, such as identity transformation, facial expression transformation, attribute editing, and facial synthesis. At present, detection algorithms for these two types of attacks, ``Face Anti-spoofing (FAS)'' and ``Deep Fake/Forgery Detection (DeepFake)'', are still being studied as independent computer vision tasks, and cannot achieve the functionality of a unified detection model to respond to both types of attacks simultaneously. To give continuity to our efforts in these relevant problems, we are proposing the 5th Face Anti-Spoofing Workshop@CVPR 2024. We analyze different types of attack clues as the main reason for the incompatibility between these two detection. The spoofing clues based on physical presentation attacks are usually caused by color distortion, screen moire patterns, and production traces. In contrast, the forgery clues based on digital editing attacks are usually changes in pixel values. The fifth competition aims to encourage the exploration of common characteristics in these two types of attack clues and promote the research of unified detection algorithms. Fully considering the above difficulties and challenges, we collect a Unified physical-digital Attack dataset, namely UniAttackData (accepted by IJCAI 2024), for this fifth edition for algorithm design and competition promotion.

Bridging the Gap between Computational Photography and Visual Recognition

Organizers: Nicholas Chimitt, Xingguang Zhang, Ajay Jaiswal, Wes Robbins, Yuecong Xu, Yang Jianfei, Yuan Shenghai, Yang Yizhuo, Hyoungseob Park, Fengyu Yang, Howard Zhang, Rishi Upadhyay, Stanley Chan, Zhangyang Wang, Achuta Kadambi, Alex Wong, Bradley Preece

Date: Monday, June 17

Time: 8:30 AM–12:00 PM

Location: Arch 212

Summary: With the development and successes of computer vision algorithms, application to real-world environments become increasingly feasible for these methodologies. However, real-world applications may involve degradations which are not often included in standard training and testing datasets: poor illumination, adverse weather conditions, aberrated optics, or complicated sensor noise. In addition to this, non-standard imaging solutions such as those for flexible wearables or augmented reality headsets may require unconventional processing algorithms which complicates their path to real-world application. What is the gap in performance for the current state of the art when placed into harsh, real-world environments? The central theme of the workshop in this proposal is to invite researchers to investigate this question and push the state-of-the-art forward for real-world application of recognition tasks in challenging environments. Continuing the history of success at CVPR 2018–2023, we provide its 7th version for CVPR 2024. It will inherit the successful benchmark dataset, platform, and other evaluation tools used in previous UG2+ workshops, as well as broadening its scope with new tracks and applications within the context of real-world application of computer vision algorithms.

Computer Vision for Fashion, Art, and Design

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

Date: Monday, June 17

Time: 8:30 AM–12:00 PM

Location: Summit 334

Summary: Creative domains represent a large part of modern society and have a strong impact on the economy and cultural life. Much of the effort in creative domains such as fashion, art, and design is focused on the creation, consumption, manipulation, and analysis of visual content. In recent years, there has been an explosion of research into the application of machine learning and computer vision algorithms to various aspects of creative domains. For six consecutive years, the CVFAD workshop series has captured important trends and new ideas in this area. At CVPR 2024, we will continue to bring together artists, designers, and computer vision researchers and engineers. We will continue to develop the workshop itself as a space for conversation and idea exchange at the intersection of computer vision and creative applications.

Large Scale Holistic Video Understanding

Organizers: Mohsen Fayyaz, Vivek Sharma, Ali Diba, Shyamal Buch, Juergen Gall, Luc Van Gool, Joao Carreira, Ehsan Adeli, David Ross, Manohar Paluri

Date: Monday, June 17

Time: 8:30 AM–12:00 PM

Location: Summit 429

Summary: In this workshop, we introduce holistic video understanding as a new challenge for video understanding efforts. This challenge focuses on the recognition of scenes, objects, actions, attributes, and events in real-world user-generated videos. Following our previous successful workshop and tutorial where we introduced our new dataset, Holistic Video Understanding~(HVU dataset), this workshop is tailored to bringing together ideas around universal video representation learning. We will discuss current SOTA supervised and unsupervised video representation learning methods and introduce our HVU dataset as a new benchmark for such tasks to better evaluate the universality of the learned video representations in terms of multiple semantic categories and multi-task/multi-label learning. Our goal is to advance universal visual understanding that can tackle real-world problems. As a second step towards this goal, this workshop will continue in the footsteps of its 1st edition and will expand and present HVU as a universal dataset that targets several tasks simultaneously. As a holistic video understanding benchmark, this workshop will integrate joint recognition of all the semantic concepts present in the video by going beyond a single class label per task. We invite the community to help to extend this dataset that will spur research in video understanding as a comprehensive, multi-faceted problem. We also plan to accompany the workshop with the "Second International Challenge on Holistic Video Understanding". Wherein the participants should use the HVU dataset to train video recognition models with different levels of semantic concepts.

MetaFood

Organizers: Yuhao Chen, Jiangpeng He, Fengqing Zhu, Edward Delp, Alexander Wong

Date: Monday, June 17

Time: 8:30 AM–12:30 PM

Location: Arch 309

Summary: Today, computer vision algorithms show near-perfect performance, better than human when there are clear, well curated and enough amount of data. However, there remains a substantial gap when it comes to applying state-of-the-art computer vision algorithms to food data, particularly when dealing with food in its natural, uncontrolled environment, often referred to as "data in the wild." This gap stems from the inherent challenges in noisy, watermarked, and low-quality food data readily available on the internet. The MetaFood Workshop (MTF) invites the CVPR community to engage with the food domain-related challenges. These challenges provide not only a demanding, real testing environment for the development of robust computer vision algorithms, but also an exciting opportunity to develop new algorithms in the fields of food data analysis and food digitization.

New Challenges in 3D Human Understanding

Organizers: Qianli Ma, Siwei Zhang, Rawal Khirodkar, Shashank Tripathi, Georgios Pavlakos, Angjoo Kanazawa, Siyu Tang

Date: Monday, June 17

Time: 8:30 AM–1:00 PM

Location: Summit 440-441

Summary: Understanding humans in 3D remains at the heart of computer vision advancements. The complexities associated with analyzing human behaviors, postures, and interactions are multifaceted, making them a challenging yet crucial aspect of vision research. Recent developments in the field have brought to light new challenges, including multi-human interaction analysis, fine-grained motion capture, clothed body reconstruction, virtual try-ons and tracking under extreme occlusions. This workshop aims to gather researchers in this field, to assess existing methodologies, understand their limitations, and collaboratively discuss novel approaches to push the boundaries of 3D human understanding in computer vision. To that end, this workshop will feature: (1) Invited speakers from both the academic and industrial spheres. (2) An invited poster session that covers diverse relevant works accepted at recent top-tier conferences such as CVPR 2024 and NeurIPS/ICCV/CVPR 2023. (3) A panel discussion featuring invited speakers, aimed at fostering the exchange of ideas, addressing limitations, and identifying promising research directions.

ViLMa – Visual Localization and Mapping

Organizers: Patrick Wenzel, Daniel Cremers, Dima Damen, Lukas Koestler, Stefan Leutenegger, Raquel Urtasun, Niclas Zeller

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 327

Summary: Visual localization and mapping is a fundamental aspect of computer vision, with applications ranging from autonomous robotics to augmented reality. This workshop aims to bring together researchers, practitioners, and enthusiasts in the field to discuss the latest developments, challenges, and applications of visual localization and mapping. The workshop will provide a platform for sharing insights, fostering collaborations, and exploring the cutting-edge research that underpins this crucial area of computer vision.

AIS: Vision, Graphics and AI for Streaming

Organizers: Marcos V. Conde, Radu Timofte, Ioannis Katsavounidis, Ryan Lei, Daniel Motilla, Christos Bampis, Rakesh Ranjan

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Arch 3A

Summary: This workshop focuses on unifying new streaming technologies, computer graphics, and computer vision, from the modern deep learning point of view. Streaming is a massive industry where hundreds of millions of users demand everyday high-quality content on different platforms. Computer vision and deep learning have emerged as revolutionary forces for rendering content, image and video compression, enhancement, and quality assessment. From neural codecs for efficient compression to deep learning-based video enhancement, these advanced techniques are setting new standards for streaming quality and efficiency. Moreover, novel neural representations also pose new challenges and opportunities in rendering streamable content, and allowing to redefine computer graphics pipelines and visual content. We invite researchers and engineers from academia and industry to join us to learn more about the future of streaming.

Computer Vision in the Wild

Organizers: Chunyuan Li, Jianwei Yang, Haotian Liu, Xueyan Zou, Wanrong Zhu, Yonatan Bitton, Jianfeng Gao

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Arch 3B

Summary: A long-standing aspiration in artificial intelligence is to develop general-purpose assistants that can effectively follow users' (multimodal) instructions to complete a wide range of real-world tasks. Recently, the community has witnessed a growing interest in developing foundation models with emergent abilities of multimodal understanding and generation in open-world tasks. While the recipes of using large language models (LLMs) such as ChatGPT to develop general-purpose assistants for natural language tasks have been proved effective, the recipes of building general-purpose, multimodal assistants for computer vision and vision-language tasks in the wild remain to be explored. Recent works show that learning from large-scale image-text data with human feedback in the loop 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, large multimodal models (LMM) such as Flamingo, GPT-4V, Gemini have demonstrated strong zero-shot transfer capabilities on many vision tasks in the wild. The open-source LMMs have also made significant progress, as demonstrated by OpenFlamingo, MiniGPT4 and LLaVA. These models are trained with visual instruction-following data, where human intents are represented in natural language. On the other hand, interactive vision systems such as Segment Anything (SAM) and SEEM have also shown impressive segmentation performance on almost anything in the wild, where human intents are represented in visual prompts, such as click, bounding boxes and text. These vision models with language and multimodal interfaces are naturally open-vocabulary and even open-task models, showing superior zero-shot performance in 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 problems in real-world scenarios, focusing on the challenge of open-world visual task-level transfer. This CVPR 2024 CVinW workshop is a continuation of CVPR 2023 CVinW Workshop and ECCV 2022 CVinW Workshop. The development of LMMs is an emerging new filed, with a vast research exploration space in data collection, modeling, evaluation and new application scenarios. There are many new benchmarks merged to measure their performance from different aspects. To advocate established benchmarks to measure the progress, this workshop welcome authors different benchmark to report results.

AI for Content Creation

Organizers: Deqing Sun, Lingjie Liu, Yuanzhen Li, Sergey Tulyakov, Huiwen Chang, Lu Jiang, Yijun Li, Jun-Yan Zhu, James Tompkin

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 342

Summary: Content creation plays a crucial role in application domains like photography, videography, virtual reality, gaming, art, design, and fashion. Recent progress in machine learning and AI has transformed hours of manual, painstaking content creation work into minutes or seconds of automated or interactive work. For instance, generative modeling approaches can produce photorealistic images of 2D and 3D items such as humans, landscapes, interior scenes, virtual environments, or even industrial designs. New large text and image models that share latent spaces let us imaginatively describe images and have them realized automatically---a capability that until recently was not thought easily possible. Learned priors of images, videos, and 3D data can also be combined with explicit appearance and geometric constraints, perceptual understanding, or even functional and semantic constraints of objects. In addition to creating awe-inspiring artistic images, such techniques offer unique opportunities for generating diverse synthetic training data for downstream computer vision tasks, both in 2D and in 3D domains. The AI for Content Creation workshop explores this exciting and fast-moving research area. We will bring together invited speakers of world-class expertise in content creation, up-and-coming researchers, and authors of submitted workshop papers, to engage in a day filled with learning, discussion, and network building.

AI for 3D Generation

Organizers: Despoina Paschalidou, George Pavlakos, Davis Remple, Angel Xuan Chang, Kai Wang, Amlan Kar, Daniel Ritchie, Kaichun Mo, Manolis Savva, Paul Guerrero, Siyu Tang, Leo Guibas

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit Flex A

Summary: Generating diverse and realistic 3D content is a long-standing problem in Computer Vision and Graphics that has recently gained more attention due to the increased demand for 3D models of digital humans and objects in AR/VR, robotics and gaming applications. Due to the significant progress in generating 3D objects and humans, the research community has shifted their attention towards developing models that can generate 3D environments with multiple humans interacting with other humans or objects in the scene. Despite their promising performance, scaling these approaches to complex scenes with several static and dynamic objects still remains an open challenge. Furthermore, existing generative models do not exert intuitive control, namely controlling what needs to be generated or changed typically requires deep technical knowledge of each model, thus making them impractical to any practitioner. Therefore, in this workshop, we seek to bring together researchers working on generative models for 3D shapes, humans, and scenes, both from the industry and academia, to discuss the latest topics in 3D content creation along with ways to make models more accessible and useful to larger audiences. Furthermore, we seek to better understand the challenges and next steps towards developing generative models capable of producing fully controllable 3D environments containing multiple humans interacting with each other and objects. By inviting speakers from diverse backgrounds, both from industry and academia, we hope to initiate fruitful discussions that will inspire future research to further push the boundaries of AI-generated content creation. Finally, in this workshop, we also hope to address the ethical implications of artificially generated 3D content and to raise awareness of the potential malicious use of the generative technologies.

Dataset Distillation for Computer Vision

Organizers: Saeed Vahidian, Yiran Chen, Bo Zhao, Ramin Hasani, Alexander Amini, Dongkuan (DK) Xu, Ruochen Wang, Vyacheslav Kungurtsev, Xinchao Wang

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 329

Summary: The recent trend of migrating computation from the centralized cloud to distributed edge devices is reshaping the landscape of today's Internet. Distributed machine learning, specifically federated learning (FL), has been envisioned as a key technology for enabling next generation AI at-scale. Moreover, with privacy being a critical concern in data aggregation, FL emerges as a promising solution to such privacy-utility challenges. It pushes the computation towards the consumer's edge devices, where the data is generated. By exchanging statistical information rather than the original data, the participants perform collaborative learning in a distributed fashion. Although FL has become an important privacy-preserving paradigm in various machine learning tasks, the potential of FL in computer vision (CV) 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, image segmentation, and video understanding, compared to the traditional centralized training paradigm. This workshop aims at bringing together researchers and practitioners with a common interest in FL for computer vision. This workshop is an attempt at studying the different synergistic relations in this interdisciplinary area. This day-long event will facilitate interaction among students, scholars, and industry professionals from around the world to discuss future research challenges and opportunities.

Federated Learning for Computer Vision

Organizers: Chen Chen, Matias Mendieta, Salman Avestimehr, Zhengming Ding, Mi Zhang, Ang Li, Bo Li, Yang Liu, Gauri Joshi, Saeed Vahidian

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 325

Summary: The recent trend of migrating computation from the centralized cloud to distributed edge devices is reshaping the landscape of today's Internet. Distributed machine learning, specifically federated learning (FL), has been envisioned as a key technology for enabling next generation AI at-scale. Moreover, with privacy being a critical concern in data aggregation, FL emerges as a promising solution to such privacy-utility challenges. It pushes the computation towards the consumer's edge devices, where the data is generated. By exchanging statistical information rather than the original data, the participants perform collaborative learning in a distributed fashion. Although FL has become an important privacy-preserving paradigm in various machine learning tasks, the potential of FL in computer vision (CV) 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, image segmentation, and video understanding, compared to the traditional centralized training paradigm. This workshop aims at bringing together researchers and practitioners with common interest in FL for computer vision. This workshop is an attempt at studying the different synergistic relations in this interdisciplinary area. This day-long event will facilitate interaction among students, scholars, and industry professionals from around the world to discuss the future research challenges and opportunities.

Adversarial Machine Learning on Computer Vision: Robustness of Foundation Models

Organizers: Aishan Liu, Jiakai Wang, Mo Zhou, Qing Guo, Xiaoning Du, Xinyun Chen, Cihang Xie, Felix Juefei-Xu, Xianglong Liu, VishalM M. Patel, Dawn Song, Alan Yuille, Philip H.S. Torr, Dacheng Tao

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 435

Summary: Artificial intelligence (AI) has entered a new era with the emergence of foundation models (FMs). These models demonstrate powerful generative capabilities by leveraging extensive model parameters and training data, which have become a dominant force in computer vision, revolutionizing a wide range of applications. Alongside their potential benefits, the increasing reliance on FMs has also exposed their vulnerabilities to adversarial attacks. These malicious attacks involve applying imperceptible perturbations to input images or prompts, which can cause the models to misclassify the objects or generate adversary-intended outputs. Such vulnerabilities pose significant risks in safety-critical applications, such as autonomous vehicles and medical diagnosis, where incorrect predictions can have dire consequences. By studying and addressing the robustness challenges associated with FMs, we could enable practitioners to better construct robust, reliable FMs across various domains. The workshop will bring together researchers and practitioners from the computer vision and machine learning communities to explore the latest advances and challenges in adversarial machine learning, with a focus on the robustness of foundation models. The program will consist of invited talks by leading experts in the field, as well as contributed talks and poster sessions featuring the latest research. In addition, the workshop will also organize a challenge on adversarial attacking foundation models. We believe this workshop will provide a unique opportunity for researchers and practitioners to exchange ideas, share latest developments, and collaborate on addressing the challenges associated with the robustness and security of foundation models. We expect that the workshop will generate insights and discussions that will help advance the field of adversarial machine learning and contribute to the development of more secure and robust foundation models for computer vision applications.

Foundation Models

Organizers: Teng Xi, Hisham Cholakkal, Gang Zhang, Fahad Khan, Errui Ding, Ming-Hsuan Yang, Xian Sun, Joost van de Weijer, Linchao Zhu, Salman Khan, Yifan Sun, Rao Mohammed Anwer, Yi Yang, Mubarak Shah, Edith Ngai, Song Bai, Jingdong Wang

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 434

Summary: Recent years have witnessed remarkable advancements in foundation models for natural language processing and computer vision. Trained on large-scale diverse datasets, these models serve as a basis for various downstream tasks, offering adaptability and robustness. The Second Workshop on Foundation Models aims to provide a platform for exploring the latest research and practical applications of vision foundation models and large language models (LLMs). Our workshop will focus on bridging the gap between cutting-edge research in foundation models and their real-world applications across different domains, including healthcare, earth sciences, remote sensing, biology, science, agriculture, and climate sciences. We encourage submissions exploring theoretical insights into vision foundation models (VFMs) and large language models, efficient foundation model architectures, and hybrid network designs combining VFMs with convolutional and graph-based models. Additionally, contributions addressing challenges and strategies involved in leveraging VFMs and LMMs for tasks such as image and video generative models, unsupervised, weakly, and semi-supervised learning settings, multi-modal foundation models, and improving semantic understanding of visual content in multi-spectral data are welcomed. The workshop also aims to accelerate the adoption of foundation models in various industrial applications by investigating inherent biases, blind spots, and strategies for safe and privacy-preserving deployment. We welcome papers reporting experimental results on accelerating the training and inference time of foundation models, as well as their adaptation for low-level and high-level vision problems, mobile devices, and embodied AI.

Efficient and On-Device Generation

Organizers: Felix Juefei-Xu, Tingbo Hou, Licheng Yu, Ruiqi Gao, Xiaoliang Dai, Huiwen Chang, Bichen Wu, Chenlin Meng, Ning Zhang, Yanwu Xu, Xi Yin, Sirui Xie, Camille Couprie, Yunzhi Zhang, Andrew Brown, Yang Zhao, Ali Thabet, Zhisheng Xiao, Peizhao Zhang, Peter Vajda

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 432

Summary: The First Workshop on Efficient and On-Device Generation (EDGE) at CVPR 2024 will focus on the latest advancements of generative AI in the computer vision domain, with an emphasis on efficiencies across multiple aspects. We encourage techniques that enable generative models to be trained more efficiently and/or run on resource-constrained devices, such as mobile phones and edge devices. Through these efforts, we envision a future where these permeating generative AI capabilities become significantly more accessible with virtuous scalability and plateauing carbon footprint.

Mobile AI Workshop and Challenges

Organizers: Andrey Ignatov, Radu Timofte

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Arch 211

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 the majority of 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.

VAND 2.0: Visual Anomaly and Novelty Detection

Organizers: Thomas Brox, Toby Breckon, Guansong Pang, Yedid Hoshen, Philipp Seeboċk , Paul Bergmann, Latha Pemula

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 330

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 they 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, but not limited to medical diagnosis, airport security screening, industrial inspection, and crowd control with varying degrees of success. However, anomaly detection is far from a simple task due to the challenges of accounting for all forms with which an anomaly may be present. For the established supervised techniques, this can lead to a lack of exposure to certain types of anomalies during training leading to heavy classification bias and over-fitting. Subsequently, it is typically impossible for any given dataset to account for all forms within which an anomaly may present itself within a given context or task as essentially they represent an unbounded (open set) distribution of possible deviations from the distribution of normality. To these ends, we now see the rise of a complex and vibrant set of learning-based paradigms that address 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.

Physics Based Vision meets Deep Learning

Organizers: Shaodi You, Ying Fu, Yu Li, Boxin Shi, Jose Alvarez, Coert van Gemeren

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 333

Summary: The goal of the 4th workshop on physics based vision meets deep learning is to bring together researchers and engineers from both academia and industry to discuss the current state of the art in the field and future challenges of integrating classic physics based approaches into the modern data-driven deep learning ones. This successful workshop series started in 2017 and, for this edition, maintains the core organizers and includes newer organizers. For this edition, we introduce recent ML advances (such as stable diffusion and LLM) for physics based vision (such as under water imaging, bad weather, low light enhancement, light field, hyper spectral imaging). We also introduce a low-light enhancement and detection challenge.

Fair, Data-efficient, and Trusted Computer Vision

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

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Arch 303

Summary: As the computer vision research community makes rapid progress in producing algorithms with human-level performance, it is extremely critical that we take a step back and assess, and consequently promise to the consumer world, what this objective performance reported in academic literature means in the context of real-world systems and applications. As a concrete example, it is one thing for a social media organization to use an algorithm to automatically identify a person of interest in pictures uploaded to its platform. On the other hand, the use of algorithms in making life-changing decisions in areas such as healthcare (e.g., should a certain treatment be administered?) or jurisprudence (e.g., should this person be released from prison?) is a totally different ballgame. At the very least, the following questions will be asked of the algorithm/system by the user: Why is the algorithm predicting X? How sure is the algorithm of this prediction/decision? Why should I trust the algorithm? How can I be sure the algorithm has been fair in the process leading up to its prediction/decision? Is the algorithm biased? Answers to these questions can have profound consequences depending on the application (e.g., accidents and autonomous vehicles, life/death for a patient, incarceration/freedom for an accused). Consequently, as artificial intelligence (AI) is seeing increasing adoption in a variety of daily-life applications, addressing the underlying themes of the questions above has become a matter of urgent importance. In light of these issues, we seek to provide a focused venue for academic and industry researchers and practitioners to discuss research challenges and solutions associated with learning computer vision models with the overarching requirements of fairness, data efficiency, and trustworthiness.

CV4Science

Organizers: David Fouhey, Katie Bouman, Subhransu Maji

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Arch 2A

Summary: Our workshop aims to bring together people working at the intersection of computer vision and the sciences. While many computer vision researchers are working at the intersection of computer vision and a science discipline, these efforts are often not highlighted at CVPR. As a result researchers remain disconnected, unaware of each others' work, and miss opportunities to learn from each other. Early career researchers walking the poster floor of CVPR can get the mistaken impression that computer vision researchers do not work on or care about these problems. We aim to highlight work in this space and are interested in any topic that covers both computer vision and the sciences: Computer vision topics in this area often include (but are not limited to): reconstruction, recognition, segmentation and counting, human-in-the loop efforts, low-shot learning, domain adaptation and sim2real, video analysis, joint design of hardware and software. Science topics include (but are not limited to): astrophysics via a variety of instrument types (radio, light, spectropolarimetry), chemistry, biology, neuroscience, and ecology.

CV4Animals: Computer Vision for Animal Behavior Tracking and Modeling

Organizers: Urs Waldmann, Shangzhe Wu, Gengshan Yang, Anna Zamansky

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Arch 214

Summary: Many biological organisms have evolved to exhibit diverse behaviors, and 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. Recently, computer vision has been making a significant 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.

Urban Scene Modeling: Where Vision Meets Photogrammetry and Graphics

Organizers: Ruisheng Wang, Jack Langerman, Ilke Demir, Qixing Huang, Florent Lafarge, Dmytro Mishkin, Tolga Birdal, Hui Huang, Shangfeng Huang, Daoyi Gao, Xiang Ma, Hanzhi Chen, Clement Mallet, Caner Korkmaz, Yang Wang, Marc Pollefeys

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 443

Summary: Rapid urbanization poses social and environmental challenges. Addressing these issues effectively requires access to accurate and up-to-date 3D building models, obtained promptly and cost-effectively. Urban modeling is an interdisciplinary topic among computer vision, graphics, and photogrammetry. The demand for automated interpretation of scene geometry and semantics has surged due to various applications, including autonomous navigation, augmented reality, smart cities, and digital twins. As a result, substantial research effort has been dedicated to urban scene modeling within the computer vision and graphics communities, with a particular focus on photogrammetry, which has coped with urban modeling challenges for decades. This workshop is intended to bring researchers from these communities together. Through invited talks, spotlight presentations, a workshop challenge, and a poster session, it will increase interdisciplinary interaction and collaboration among photogrammetry, computer vision and graphics. We also solicit original contributions in the areas related to urban scene modeling.

Long-form Video Understanding: Towards Multimodal AI Assistant and Copilot

Organizers: Mike Zheng Shou, Linchao Zhu, Difei Gao, Joya Chen, Stan Lei, Jitendra Malik, Weiyao Wang, Xiaohan Wang, Hehe Fan, Kristen Grauman, Matt Feiszli, Lorenzo Torresani, Karttikeya Mangalam, Jay Wu, Forrest Iandola, Xiuyu Li, Zhen Dong, Kurt Keutzer, Ziqi Huang, Fan Zhang, Ziwei Liu, Wenhao Chai, Enxin Song

Date: Monday, June 17

Time: 8:30 AM–5:30 PM

Location: Summit 430

Summary: The exemplary performance of multimodal large models in image domains, such as GPT-4V, has unveiled the potential for AI assistants capable of aiding users in their daily lives. However, to construct AI assistants that can provide efficient and accurate assistance in complex tasks, a fundamental ability to understand intricate long-form videos is essential. This topic holds tremendous interest for researchers working across a wide spectrum of video tasks including action recognition, video QA, video summarization, and procedure planning, thereby making this workshop highly relevant and timely. Our workshop not only includes enlightening talks to point out future directions but also identifies proper benchmarks and provides data annotations that can significantly advance the model development for multimodal assistants and copilots capable of handling long-form videos.

New Trends in Image Restoration and Enhancement Workshop and Challenges

Organizers: Radu Timofte, Zongwei Wu, Marcos V. Conde, Florin Vasluianu, Ren Yang, Yawei Li, Bin Ren, Nancy Mehta, Zheng Chen, Yulun Zhang, Kai Zhang, Longguang Wang, Yingqian Wang, Yulan Guo, Zhi Jin, Shuhang Gu, Zhilu Zhang, Wangmeng Zuo, Ming-Hsuan Yang, Kyoung Mu Lee, Codruta Ancuti, Cosmin Ancuti, Nicolas Chahine, Sira Ferradans, Xiaohong Liu, Xin Li, Kun Yuan, Zhibo Chen, Jie Liang, Lei Zhang, Xiaoning Liu, Tom Bishop, Fabio Tosi, Pierluigi Zama Ramirez, Luigi Di Stefano, Egor Ershov, Luc Van Gool

Date: Monday, June 17

Time: 8:30 AM–6:00 PM

Location: Arch 204

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. The 9th edition of the New Trends in Image Restoration and Enhancement (NTIRE) workshop provides an overview of the new trends and advances in areas concerning image restoration, enhancement and manipulation. NTIRE 2024 workshop features invited talks from distinguished researchers, 74 paper presentations addressing topics related to image and video restoration, enhancement and manipulation and hosts 17 associated challenges gauging the state-of-the-art for various tasks: dehazing, night photography, compressed image enhancement, shadow removal, super-resolution, stereo super-resolution, efficient super-resolution, light field super-resolution, depth estimation, quality assessment of portraits, AI and user-generated (video) contents, bracketing restoration and enhancement, low light enhancement, restoration in the wild, raw super-resolution, burst alignment and learned ISP.

Foundation Models for Medical Vision

Organizers: Jun Ma, Vishal M. Patel, Julia A Schnabel, Yuyin Zhou, Bo Wang

Date: Monday, June 17

Time: 8:30 AM–6:00 PM

Location: Summit 324

Summary: The rapid growth of foundation models in various domains has been transformative, bringing unprecedented capabilities and advances in automated understanding. Medical vision, a pivotal segment of computer vision, is poised to greatly benefit from these advancements. This workshop delves into the integration and application of foundation models specific to the realm of medical imaging. We will cover state-of-the-art techniques for diverse medical data, such as echocardiogram, fundus, pathology, and radiology, as well as the practical challenges of implementing these models in clinical settings. Through expert-led sessions, interactive discussions, and international competitions, we aim to offer attendees a comprehensive understanding of the potential impact foundation models could have on the future of medical diagnostics and patient care.

Multimodal Content Moderation

Organizers: Mei Chen, Cristian Canton, Davide Modolo, Maria Zontak, Maarten Sap, Matthew Lease

Date: Monday, June 17

Time: 8:30 AM–6:00 PM

Location: Arch 304

Summary: Content moderation (CM) is a rapidly growing need in today's industry, 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-4 Vision present additional challenges to CM systems. With organizers across industry and academia, speakers who are experts across relevant disciplines investigating technical and policy challenges, the Workshop on Multimodal Content Moderation (MMCM) aims to strengthen and nurture the community for interdisciplinary cross-organization knowledge sharing to push the envelope of what is possible, and improve the quality of multimodal sensitive content detection and moderation that will benefit the society at large.

Notes:

Joint Egocentric Vision

Organizers: Antonino Furnari, Angela Yao, Xin Wang, Tushar Nagarajan, Huiyu Wang, Jing Dong, Jakob Engel, Siddhant Bansal, Takuma Yagi

Date: Monday, June 17

Time: 8:30 AM–6:15 PM

Location: Summit 428

Summary: This joint workshop aims to be the focal point for the egocentric computer vision community to meet and discuss progress in this fast growing research area, addressing egocentric vision in a comprehensive manner including key research challenges in video understanding, multi-modal data, interaction learning, self-supervised learning, AR/VR with applications to cognitive science and robotics. Wearable cameras, smart glasses, and AR/VR headsets are gaining importance for research and commercial use. They feature various sensors like cameras, depth sensors, microphones, IMUs, and GPS. Advances in machine perception enable precise user localization (SLAM), eye tracking, and hand tracking. This data allows understanding user behavior, unlocking new interaction possibilities with augmented reality. Egocentric devices may soon automatically recognize user actions, surroundings, gestures, and social relationships. These devices have broad applications in assistive technology, education, fitness, entertainment, gaming, eldercare, robotics, and augmented reality, positively impacting society. Previously, research in this field faced challenges due to limited datasets in a data-intensive environment. However, the community's recent efforts have addressed this issue by releasing numerous large-scale datasets covering various aspects of egocentric perception, including HoloAssist, Aria Digital Twin, Aria Synthetic Environments, Ego4D, Ego-Exo4D, and EPIC-KITCHENS. The goal of this workshop is to provide an exciting discussion forum for researchers working in this challenging and fast-growing area, and to provide a means to unlock the potential of data-driven research with our datasets to further the state-of-the-art.

Tool-Augmented VIsion

Organizers: Ahmet Iscen, Gul Varol, Pan Lu, Ziniu Hu, Mathilde Caron, Alireza Fathi, Minsu Cho

Date: Monday, June 17

Time: 8:45 AM–12:45 PM

Location: Summit 321

Summary: Recent vision-language models (VLMs), such as CLIP, Flamingo, PaLI, have shown strong capabilities to memorize large amounts of world knowledge when scaled to tens of billions of parameters and trained on web-scale data. Although these models have achieved remarkable results across various benchmarks, they tend to struggle on tasks that require 1)) seeking the answers from external sources 2) long-tail knowledge and 3) fine-grained understanding. Recently, there has been a growing interest in retrieval and tool-augmented models that rely on non-parametric, external knowledge sources to address these limitations. In this inaugural edition of the TAVI workshop, we aim to bring together a diverse group of researchers who will share their recent work on this exciting and increasingly popular topic with our computer vision community.

Learning 3D with Multi-View Supervision

Organizers: Abdullah Hamdi, Guocheng Qian, Chuanxia Zheng, Silvio Giancola, Sara Rojas Martinez, Jinjie Mai, Yash Bhalgat, Bernard Ghanem

Date: Monday, June 17

Time: 8:45 AM–5:30 PM

Location: Summit 331

Summary: Following the success of the first Workshop for Learning 3D with Multi-View Supervision held during CVPR 2023, we are excited to bring forth the second iteration of this workshop for CVPR 2024. With the growing interest and advancements in the domain, this year's workshop promises more depth, diverse topics, and inclusive participation. It would cover various topics that involve multi-view deep learning for core 3D understanding tasks (recognition, detection, segmentation) and methods that use posed or un-posed multi-view images for 3D reconstruction and generation. A set of new topics of interest will be added such as dynamic multi-view datasets and generative 4D models that leverage multi-view representation. The detailed topics covered in the workshop include the following: Multi-View for 3D Understanding, Deep Multi-View Stereo, Multi-View for 3D Generation and Novel View Synthesis, Dynamic Multi-View Datasets and 4D Generative models.

AI4Space

Organizers: Tat-Jun Chin, Gabriele Meoni, Djamila Aouada, Tae Ha Park, Rajat Talak, Viorela Ila, Nicolas Longépé, Aunkumar Rathinam

Date: Monday, June 17

Time: 9:00 AM–12:10 PM

Location: Arch 205

Summary: The space sector is experiencing significant growth. Currently planned activities and utilisation models also greatly exceed the scope, ambition and/or commercial value of space missions in the previous century, e.g., autonomous satellites for earth observation, on-orbit servicing, intelligent rovers for planetary exploration, and space traffic management and debris mitigation. Achieving these ambitious goals requires surmounting non-trivial technical obstacles. AI4Space focuses on the role of AI, particularly computer vision and machine learning, in helping to solve those technical hurdles. The workshop will highlight the space capabilities that draw from and/or overlap significantly with vision and learning research, outline the unique difficulties presented by space applications to vision and learning, and discuss recent advances towards overcoming those obstacles.

Foundation Models for Autonomous Systems

Organizers: Hongyang Li, Kashyap Chitta, Holger Caesar, German Ros, Christos Sakaridis, Anthony Hu

Date: Monday, June 17

Time: 9:00 AM–6:00 PM

Location: Summit 442

Summary: Autonomous systems, such as robots and self-driving cars, have rapidly evolved over the past decades. Recently, foundation models have emerged as a promising approach to building more generalist autonomous systems due to their ability to learn from vast amounts of data and generalize to new tasks. The motivation behind this workshop is to explore the potential of foundation models for autonomous agents and discuss the challenges and opportunities associated with this approach. Besides, the workshop comprises an Autonomous Grand Challenge with seven new tasks that push the boundary of existing perception, prediction, and planning pipelines.

Causal and Object-Centric Representations for Robotics

Organizers: Efstratios Gavves, Roozbeh Mottaghi, Francesco Locatello, Andrii Zadaianchuk, Ruta Desai, Priyam Parashar, Siddharth Patki, Phillip Lippe

Date: Monday, June 17

Time: 9:00 AM–5:00 PM

Location: Arch 210

Summary: Current approaches in computer vision and machine learning primarily rely on identifying statistical correlations within massive datasets. This reliance limits their efficacy in areas that necessitate generalization through higher-order cognition, such as domain generalization and planning. A foundational approach to overcome these limitations involves incorporating principles of causality into the processing of large datasets. Similar to classic AI methodologies, causal inference usually assumes that the causal variables of interest are provided externally. However, real-world data, often encapsulated in high-dimensional, low-level observations (e.g., RGB pixels in a video), generally lacks organization into meaningful causal units. • Causal Representation Learning proposes a promising approach by integrating principles of causality, enabling models to discern cause-and-effect relationships and thereby generate controllable representations. • From another perspective, Object-centric Representation Learning focuses on decomposing sensory inputs, such as images and videos, into set-based representations where distinct vectors represent different objects. • Thus, Robotics and Embodied AI that require compositional and controllable scene representations can highly benefit from object-centric and causal representations. This workshop aims to bring together researchers from structured (object-centric and causal) representation learning and robotics-oriented computer vision. To help integrate ideas from these areas, we invite researchers from Embodied AI, Causality and Representation Learning. We hope that this creates opportunities for discussion, presenting cutting-edge research, establishing new collaborations and identifying future research directions.

EarthVision: Large Scale Computer Vision for Remote Sensing Imagery

Organizers: Ronny Hänsch, Devis Tuia, Jan Dirk Wegner, Bertrand Le Saux, Loïc Landrieu, Charlotte Pelletier, Hannah Kerner

Date: Monday, June 17

Time: 9:00 AM–5:30 PM

Location: Arch 310

Summary: Earth Observation (EO) and remote sensing are ever-growing fields of investigation where computer vision, machine learning, and signal/image processing meet. The general objective of the domain is to provide large-scale and consistent information about processes occurring at the surface of the Earth by exploiting data collected by airborne and spaceborne sensors. 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 sheer amount of data calls for highly automated scene interpretation workflows. Earth Observation and in particular the analysis of spaceborne data directly connects to 34 indicators out of 40 (29 targets and 11 goals) of the Sustainable Development Goals defined by the United Nations. The aim of EarthVision to advance the state of the art in machine learning-based analysis of remote sensing data is thus of high relevance. It also connects to other immediate societal challenges such as monitoring of forest fires and other natural hazards, urban growth, deforestation, and climate change.

Biometrics

Organizers: Bir Bhanu, Ajay Kumar

Date: Monday, June 17

Time: 9:00 AM–5:30 PM

Location: Arch 203

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 existing methods cannot achieve. To address this gap, it is imperative to make fundamental advancements in single biometric modalities and 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 industry, highlighting cutting-edge developments in both single and multi-modal biometric technologies.

Prompting in Vision

Organizers: Kaiyang Zhou, Amir Bar, Ziwei Liu, Yossi Grandelsman, Yixuan Li, Hyojin Bahng, Linjie Li, Amir Globerson, Yuanhan Zhang, Bo Li, Jingkang Yang, Xinlong Wang

Date: Monday, June 17

Time: 9:00 AM–5:30 PM

Location: Summit 335-336

Summary: Building general-purpose computer vision models is a multifaceted challenge that requires a system capable of understanding and interpreting a wide array of visual problems. Drawing inspiration from the field of NLP, the concept of "prompting" has been identified as a promising method for adapting large vision models to perform various downstream tasks. This adaptation process is streamlined by integrating a prompt during the inference stage. Prompts can take several forms in the context of computer vision. They can be as straightforward as providing visual examples of the input and the desired output, thereby giving the model a clear reference for what it needs to accomplish. Alternatively, prompts can be more abstract, such as a series of dots, boxes, or scribbles that guide the model's attention or highlight features within an image. Beyond these visual cues, prompts can also include learned tokens or indicators that are associated with particular outputs through the model's training process. Moreover, prompts can be constructed using language-based task descriptions. In this scenario, textual information is used to direct the model's processing of visual data, bridging the gap between visual perception and language understanding. This workshop aims to provide a platform for pioneers in prompting for vision to share recent advancements, showcase novel techniques and applications, and discuss open research questions about how the strategic use of prompts can unlock new levels of adaptability and performance in computer vision.

Autonomous Driving

Organizers: Vincent Casser, Alex Liniger, Jose Alvarez, Maying Shen, Nigamaa Nayakanti, Jannik Zuern, Dragomir Anguelov, John Leonard, Luc Van Gool

Date: Monday, June 17

Time: 9:00 AM–6:00 PM

Location: Summit 345-346

Summary: The CVPR 2024 Workshop on Autonomous Driving (WAD) brings together leading researchers and engineers from academia and industry to discuss the latest advances in autonomous driving. Now in its 7th year, the workshop has been continuously evolving with this rapidly changing field and now covers all areas of autonomy, including perception, behavior prediction and motion planning. In this full-day workshop, our keynote speakers will provide insights into the ongoing commercialization of autonomous vehicles, as well as progress in related fundamental research areas. Furthermore, we will host a series of technical benchmark challenges to help quantify recent advances in the field, and invite authors of accepted workshop papers to present their work.

Sight and Sound

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

Date: Monday, June 17

Time: 9:00 AM–6:00 PM

Location: Summit 326

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.

CV 20/20: A Retrospective Vision

Organizers: Anand Bhattad, Aditya Prakash, Unnat Jain, Svetlana Lazebnik

Date: Monday, June 17

Time: 12:45 PM–6:05 PM

Location: Arch 201

Summary: In the current fast-paced world of computer vision research, we hardly find time to pause and reflect. While conferences and workshops may be a good resource for such reflection, members of our community get busy publicizing new methods and constant benchmark-chasing efforts. Reflective and retrospective dialogs stay restricted to small groups on dinner tables and, often, exclusive circles. We wish to create a wide and open space for students and early-stage researchers to witness their community leaders reflect and share insights not about the current and "state-of-the-art". Instead, insights from looking into the past, highlighting high-level trends, their (or the community's) mistakes, and their unique experiences and diverse voices. This workshop will be a platform for critical reflection and deep discussion on past, present, and future trends in our field. We sincerely believe this well-rounded view of history, the hits, the misses, and touching upon lessons from the past will not only enrich the collective wisdom of the community but also pave the way for more informed and innovative future research in computer vision.

GenAI Media Generation Challenge for Computer Vision

Organizers: Sam Tsai, Ji Hou, Bichen Wu, Xiaoliang Dai, Kevin Chih-Yao Ma, Matthew Yu, Wang Rui, Tianhe Li, Simran Motwani, Ajay Menon, Kunpeng Li, Tao Xu, Jialiang Wang, Karthik Sivakumar, Peter Vajda, Peizhao Zhang, Ning Zhang, Sergey Tulyakov, Zijian He, Roshan Sumbaly

Date: Monday, June 17

Time: 1:00 PM–5:30 PM

Location: Summit 423-425

Summary: In the wake of the rapid advancements in generative model capabilities, the field of text-guided image generation and editing has seen unprecedented progress. This has resulted in the creation of images with unparalleled quality, aesthetics, and adherence to text guidance. Yet, a significant challenge remains: the absence of a universally-accepted, easily accessible benchmark for evaluating in-depth capabilities of the generated models. This issue stems from the lack of a comprehensive, large-scale evaluation dataset, standardized evaluation protocols, and the insufficiency of current automatic metrics. We are proud to present the GenAI Media Generation Challenge (MAGIC) that moves us towards addressing these issues. We host two challenge tracks: (1) text to image generation, and (2) text guided image editing, and provide the following set of data and resources: (a) Benchmark Datasets: diverse and comprehensive datasets will be released publicly for both tracks, ensuring that all participants have access to the same information and resources. (b) Evaluation Protocol: A standardized evaluation protocol and metrics will be established and shared with all participants. This metric will be designed keeping in mind the nuances and specifics of the tasks, ensuring fair and objective evaluation. (c) Human Annotations for all Submissions: Recognizing the importance of meticulous human oversight, we commit to offering the requisite human annotation resources as stipulated by the evaluation protocol for the duration of the competition. (d) Baselines: For aiding participants and setting a preliminary standard, baseline results will be shared. In this workshop, beyond presenting the challenge results and winners, we have also invited domain experts to give talks on media generation and evaluation. We will also host a panel discussion on evaluating generative AI solutions and tracking progress.

Image Matching: Local Features and Beyond

Organizers: Fabio Bellavia, Jiri Matas, Dmytro Mishkin, Luca Morelli, Fabio Remondino, Weiwei Sun, Amy Tabb, Eduard Trulls, Kwang-Moo Yi

Date: Monday, June 17

Time: 1:00 PM–5:45 PM

Location: Summit 323

Summary: Matching two or more images across wide baselines is a core computer vision problem, with applications to stereo, 3D reconstruction, re-localization, SLAM, and retrieval. Until recently one of the last bastions of traditional handcrafted methods, they too have begun to be replaced with learned alternatives. Interestingly, these new solutions still rely heavily on design intuitions behind handcrafted methods. Our workshop, held every year at CVPR since 2019, explores this transition period, bringing together researchers across academia and industry to assess the state of the field. We aim to establish what works, what doesn't, what's missing, and which research directions are most promising, while focusing on experimental validation. We feature invited talks, paper talks, and an open Kaggle challenge on 3D reconstruction.

Populating Empty Cities – Virtual Humans for Robotics and Autonomous Driving

Organizers: Kwan-Yee Lin, Wayne Wu, Bolei Zhou, Matthias Nießner, Stella X. Yu

Date: Monday, June 17

Time: 1:00 PM–5:45 PM

Location: Summit 334

Summary: Human beings, are the most centric and core elements in the real world. Intelligent machines (e.g., autonomous vehicles and robots) should be social-aware in the human-populated world. Perceiving humans is thus critical for robotics and autonomous driving. However, training autonomous machines in the real world with humans involved is impossible in scales – the difficulties in capturing the diversity of human behaviours and consequent changes of environments, and the considerations in human safety issues when interacting. In recent years, simulation environments emerged as a promising way to train autonomous systems. However, the environments act like ghost cities – human simulation is not included. Now, the robotics and autonomous driving domains face a common inflection point. On the one hand, the advent of new data, representations, and methodologies to build virtual humans has opened up new avenues of human simulation. The relatively innovative approach has seen great progress in various computer vision and computer graphics tasks, particularly in human rendering, reconstruction, animation, and motion synthesis with realism and fast speed. On the other hand, there is no paradigm for the combination of virtual humans with autonomous systems. An intriguing question has surfaced — can the progress in virtual humans bring a new revolution in robotics and autonomous driving? In this workshop, we aim to unfold the pioneer opinions and discussions about how will virtual humans function in robotics and autonomous driving. Two critical questions centered on virtual humans will be discussed – (1) What are the present and the future of virtual humans? (2) Will and how virtual humans play roles in robotics and autonomous driving?

TDLCV: Topological Deep Learning for Computer Vision

Organizers: Tolga Birdal, Mustafa Hajij, Michael Schaub, Theodore Papamarkou, Karthikeyan Natesan Ramamurthy, Nina Miolane, Ghada Zamzmi, Claudio Battiloro

Date: Monday, June 17

Time: 1:00 PM–5:45 PM

Location: Summit 328

Summary: Topology (and related fields of mathematics) offers a principled, novel way to work with higher-order relations within the data under interrogation. Based on the principles of topology, Topological Deep Learning (TDL) is becoming a rapidly growing frontier that leverages generalisation of graphs, such as simplicial complexes, cell complexes, and hypergraphs, to extract more global information from data, generalising most data domains encountered in scientific computations. Deep learning models developed to learn from data supported on these topological domains constitute the essence of TDL, which facilitates the analysis of higher-order network data, enabling the representation of relations between multiple entities. Consequently, computational models supported on these higher-order domains offer the possibility to tackle novel applications across various fields, such as computer vision. On the other hand, TDL also encapsulates tools of topology that are used to characterise deep neural networks. For instance, it is known that the topology of training dynamics is highly correlated to generalisation performance.

Data Curation and Augmentation in Enhancing Medical Imaging Applications

Organizers: Shuoqi Chen, Jihun Yoon, Rogerio Nespolo, Rohit Jena, Wanwen Chen, Dominik Rivoir

Date: Monday, June 17

Time: 1:00 PM–6:00 PM

Location: Summit 347-348

Summary: Medical imaging is a key component of modern healthcare, facilitating a wide array of diagnostic and therapeutic applications. Data-driven computer vision and AI solutions for medical imaging thereby represent a great potential to make a real-life impact by improving patient care. However, safety requirements associated with healthcare pose major challenges for this research field, especially regarding data curation. Collection and annotation of medical data is often resource-intensive due to the need for medical expertise. At the same time, data quality is of the highest importance to ensure safe and fair usage in clinical settings. As a result, efficient data curation and validation as well as learning from small data are important areas of research. Synthetic data generation and augmentation are further promising directions, which themselves, however, pose challenges regarding quality, bias, and utility. In addressing these demands, data engineering emerges as a crucial driver in advancing medical imaging research into deployment. Nevertheless, it is challenging to fulfill all the needs of task-specific applications via traditional methods. To bridge the gap, this workshop aims to encourage the discussion on topics related to data curation and augmentation for medical applications to tackle the challenges of limited or imperfect data in the real-world medical application.

Rhobin Challenge on Reconstruction of Human-Object Interaction

Organizers: Xi Wang, Xianghui Xie, Nikos Athanasiou, Shashank Tripathi, Ilya A. Petrov, Bharat Lal Bhatnagar, Kaichun Mo, Julien Valentin, Dimitrios Tzionas, Otmar Hilliges, Luc Van Gool, Gerard Pons-Moll

Date: Monday, June 17

Time: 1:20 PM–6:00 PM

Location: Summit 427

Summary: Following the success of the first Rhobin workshop at CVPR'23, this second half-day Rhobin workshop will continue providing a venue to present and discuss state-of-the-art research in the reconstruction of human-object interactions from images. The focus of this second workshop will go beyond image-based interaction reconstruction, extend to interaction tracking over time, and seek connections to relevant topics such as egocentric vision and dynamic scene interactions. The second Rhobin challenge will feature five tracks in total with two new tasks on human-object interaction tracking and image-based contact estimation, using two new datasets InterCap and DAMON along with BEHAVE.

Multimodalities for 3D Scenes

Organizers: Changan Chen, Angel Chang, Krishna Murthy, Alexander Richard, Kristen Grauman

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Arch 2B

Summary: Human sensory experiences such as vision, audio, touch, and smell are the natural interfaces to perceive the world around us and reason about our environments. Understanding the 3D environments around us is important for many applications such as video processing, robotics, or augmented reality. While there have been a lot of efforts in understanding 3D scenes in recent years, most works (workshops) focus on mainly using vision to understand 3D scenes. However, vision alone does not fully capture the properties of 3D scenes, e.g., the materials of objects and surfaces, the affordance of objects, and the acoustic properties. In addition, humans use language to describe 3D scenes, and understanding 3D scenes from languages is also of vital importance. We believe the future is to model and understand 3D scenes and objects with rich multi-sensory inputs, including but not limited to vision, language, audio, and touch. The goal of this workshop is to unite researchers from these different sub-communities and move towards scene understanding with multi-modalities. We want to share the recent progress of multimodal scene understanding, and also to discuss which directions the field should investigate next.

DeepFake Analysis and Detection

Organizers: Lorenzo Baraldi, Alessandro Nicolosi, Dmitry Kangin, Tamar Glaser, Plamen Angelov, Tal Hassner

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Arch 205

Summary: The second Workshop and Challenge on DeepFake Analysis and Detection (DFAD) focuses on the development of benchmarks and tools for Fake data Understanding and Detection, with the final goal of protecting from visual disinformation and misuse of generated images and text, and to monitor the progress of existing and proposed solutions for detection. Moreover, with the growing amount of generation models, the challenge of generated content detection should be generalizable to content generated by models that were unseen during the training phase. It fosters the submission of works that identify novel ways of understanding and detecting fake data, especially through new machine learning approaches capable of mixing syntactic and perceptive analysis. In parallel to soliciting the submission of relevant scientific works, the Workshop hosts a competition on deepfake detection. This is organised with the support of the ELSA project–the European Lighthouse on Secure and Safe AI, which builds on and extends the existing internationally recognized and excellently positioned ELLIS (European Laboratory for Learning and Intelligent Systems) network of excellence. The objective of the challenge is to monitor and evaluate the development of algorithms for deep fake detection, in terms of efficacy and explainability.

Graphic Design Understanding and Generation

Organizers: Kota Yamaguchi, Yizhi Wang, Naoto Inoue, Mayu Otani, Xueting Wang

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Summit 344

Summary: The workshop on Graphic Design Understanding and Generation (GDUG) aims to bring together researchers, creators, and practitioners to discuss the important concepts, technical perspectives, limitations, and ethical considerations surrounding recognition and generative approaches to graphic design. While recent advances in generative AI are making impressive strides in creative domains, there is a disconnect between raster-based approaches and the real-world workflow that involves vector graphics, such as the creation of a website, posters, online advertisements, social media posts, infographics, or presentation slides, where creators do not paint pixels but instead work with layered objects, stylistic attributes, and typography. In addition, despite the richness of what humans perceive from visual presentation, there is no universal metric for evaluating the quality of graphic design. The GDUG workshop aims to identify, discuss, and address these issues in the graphic design workflow.

Neural Architecture Search

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

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Summit 420–422

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. Details of how to participate in the competition, finishing in September, will be presented during the workshop.

Neural Rendering Intelligence

Organizers: Fangneng Zhan, Anpei Chen, Adam Kortylewski, Ayush Tewari, Siyu Tang, Christian Theobalt

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Summit 332

Summary: Neural rendering has demonstrated significant success across various fields, including computer vision, computer graphics, and robotics. The scope and definition of neural rendering have considerably widened, finding applications in numerous downstream tasks. These applications extend beyond merely demonstrating the capability to fit a specific scene; they also uncover the intelligence that arises from neural rendering techniques. As a case in point, several studies have attempted to reconstruct 3D models or render novel views from a single image using generative models, showcasing remarkable generalization abilities. This workshop is designed to promote discussions on the latest developments in neural rendering and the emergent rendering intelligence. We have gathered a diverse group of researchers who will present their most recent findings and perspectives on neural rendering. By organizing this workshop, we aim to lay a strong foundation for the future evolution of neural rendering and recognize its unique contribution to the scientific understanding and advancement of 3D intelligence.

Ethical Considerations in Creative Applications of Computer Vision

Organizers: Negar Rostamzadeh, Ziad Al-Halah, Remi Denton, Harry Jiang, Atieh Taheri, Cindy Bennett

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Arch 213

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. While artists and illustrators communities are affected by these technologies, their voices are often underrepresented in discussions. Conceptual artists have only been recently part of these conversations through events like CVPR Art Galleries in 2023 and 2024, and a range of workshops on creativity. However, these communities are not speaking on behalf of artists at large who primarily work in commercial and small-market settings as opposed to galleries or academic spaces. Our aim is to create a platform for interdisciplinary discussions among computer vision researchers, sociotechnical researchers, policy makers, social scientists, artists, illustrators and other cultural stakeholders. Our workshop will encourage retrospective discussions, position papers examining the social impacts of research in creative applications of computer vision, ethical considerations in this domain including but not limited to artwork attributions, cultural appropriation, environmental impacts of generative arts, biases embedded in generative arts, dynamics of art marketplaces/platforms, and policy perspectives on creative AI. Finally, this year, we are inviting the community to submit their retrospective papers on Art Galleries and Exhibitions in academic conferences to generate conversations on improving the community culture.

Computer Vision for Physiological Measurement

Organizers: Wenjin Wang, Daniel McDuff, Sander Stuijk

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Arch 305

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). In addition to vital signs monitoring, cameras also enable the analysis of high-level image/video semantics and context by leveraging computer vision (CV) and artificial intelligence (AI) techniques, such as facial expression analysis for pain/discomfort/delirium detection; emotion recognition for depression analysis; body motion for sleep staging; activity recognition for patient actigraphy or gait analysis; clinical workflow monitoring and optimization; etc. Camera-based monitoring will bring a rich set of compelling CV and healthcare applications that directly improve upon the human's life and care experience, such as in hospital care units, sleep/senior centers, assisted-living homes, telemedicine and e-health, home-based baby and elderly care, fitness and sports, driver monitoring in automotive, AR/VR entertainment, etc. Contactless health monitoring of cameras have been used to control the pandemic, such as vital signs screening, home-based monitoring, social distancing alarm, etc.

Vision Datasets Understanding and DataCV Challenge

Organizers: Fatemeh Saleh, Liang Zheng, Qiang Qiu, José Lezama, Xin Zhao, Piotr Koniusz, Qiuhong Ke, Manmohan Chandraker, Yue Yao, Ruining Yang, Jiajun Ding

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Summit 436

Summary: Data is the fuel of computer vision, on which 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. Comparing the number of algorithm-centric works in domain adaptation, the quantitative understanding of the domain gap is much more limited. As a result, there are currently few investigations into the representations of datasets, while in contrast, an abundance of literature concerns ways to represent images or videos, essential elements in datasets. The 3rd VDU workshop aims to bring together research works and discussions focusing on analyzing vision datasets, as opposed to the commonly seen algorithm-centric counterparts.

Pixel-level Video Understanding in the Wild Challenge

Organizers: Henghui Ding, Jiaxu Miao, Yunchao Wei, Zongxin Yang, Nikhila Ravi, Yi Yang, Si Liu, Yi Zhu, Elisa Ricci, Cees Snoek, Song Bai, Philip Torr

Date: Monday, June 17

Time: 1:30 PM–5:30 PM

Location: Summit 429

Summary: Pixel-level Scene Understanding is one of the fundamental problems in computer vision, aiming at recognizing object classes, masks and semantics of each pixel in the given image and video. Given that the real world is complex, dynamic, and essentially video-based, it is more reasonable and practical to focus on video segmentation in complex environments for real-world applications. To this end, we introduce two large-scale datasets, MOSE and MeViS, designed to enhance video segmentation in challenging complex environments. This workshop includes four challenge tracks based on the MOSE, MeViS, VSPW, and VIPSeg datasets, aiming at performing the challenging yet practical Pixel-level Video Understanding in the Wild (PVUW).

Virtual Try-On

Organizers: Vidya Narayanan, Sunil Hadap, Javier Romero, Kathleen Lewis, Hanbyul Joo, Alla Sheffer, Hao (Richard) Zhang

Date: Monday, June 17

Time: 1:30 PM–6:00 PM

Location: Arch 309

Summary: Virtual Try-On is an emerging consumer application that enables users to perceive products on their unique bodies in a virtual or mixed reality space. The retail e-commerce industry is beginning to heavily adopt these technologies within their offerings enabling their users to visualize products especially in the beauty, fashion, and accessories space before they make purchases, and provide opportunities to customize and personalize products. In principle, try-on experiences can have significant environmental impact by reducing the need to return products, improving satisfaction of purchased ones and improving accessibility. Enabling these applications requires solving diverse challenges in the space of computer vision, 3D modeling and reconstruction, geometry processing and learning, generative AI and perception. This is an active and multi-disciplinary area of research. The primary goal of this inaugral workshop is to bring together expert academic and industry researchers as well as young researchers working in this space to present, discuss and understand the state of the art and open challenges in this area that are core to enabling a convincing, useful and safe try-on experience.

Notes:

Tuesday, 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 2024 Tutorials page.

7:00–17:00	**Registration / Badge Pickup** (Summit Lobby)
7:00–17:00	**Press Room** (Summit 340)
7:00–17:00	**Mother's Room** (Summit 341-adjacent and Summit 441-adjacent)
7:00–17:00	**Prayer or Quiet Room** (Upon Request)
7:00–9:00	**Breakfast** (Summit ExHall 1-2)
8:00–18:00	**TUTORIALS / WORKSHOPS**
10:00–11:00	**Coffee Break** (Arch 4E)
12:00–13:45	**Lunch Summit** (ExHall 1-2)
15:00–16:00	**Coffee Break** (Arch 4E)

TUTORIALS

Tutorial: 3D/4D Generation and Modeling with Generative Priors

Organizers: Hsin-Ying Lee, Peiye Zhuang, Chaoyang Wang
Date: Tuesday, June 18
Time: 8:30 AM–12:00 PM
Location: Summit 440-441

Summary: In today's metaverse, where the digital and physical worlds blend seamlessly, capturing, representing, and analyzing 3D structures is vital. Advances in 3D and 4D tech have revolutionized gaming, AR, and VR, offering immersive experiences. 3D modeling bridges reality and virtuality, enabling realistic simulations and AR overlays. Adding time enhances experiences with lifelike animations and object tracking, shaping digital interactions.

Traditionally, 3D generation involved directly manipulating data, evolving alongside 2D techniques. Recent breakthroughs in 2D diffusion models have enhanced 3D tasks using large-scale image datasets. Methods like Score Distillation Sampling improve quality. However, biases in 2D data and limited 3D info pose challenges.

Generating 3D scenes and reducing biases in 2D data for realistic synthesis are ongoing challenges. Our tutorial explores techniques for diverse scenes and realism, including 3D/4D reconstruction from images and videos. Attendees learn about various generation methods, from 3D data training to leveraging 2D models, gaining a deep understanding of modern 3D modeling.

In summary, our tutorial covers the breadth of 3D/4D generation, from basics to the latest. By tackling scene-level complexities and using 2D data for realism, attendees gain insight into the evolving 3D modeling landscape in the metaverse.

Tutorial: Edge-Optimized Deep Learning: Harnessing Generative AI and Computer Vision with Open-Source Libraries

Organizers: Samet Akcay, Paula Ramos, Ria Cheruvu, Alexander Kozlov, Zhen Zhao, Zhuo Wu, Raymond Lo, Yury Gorbachev
Date: Tuesday, June 18
Time: 8:30 AM–5:00 PM
Location: Summit 436

Summary: This tutorial aims to guide researchers and practitioners in navigating the complex deep learning (DL) landscape, focusing on data management, training methodologies, optimization strategies, and deployment techniques. It highlights open-source libraries like the OpenVINO toolkit, OpenVINO Training eXtensions (OTX), and Neural Network Compression Frameworks (NNCF) in streamlining DL development. The tutorial covers how OTX 2.0 simplifies the DL ecosystem (Computer Vision) by integrating various frameworks and ensuring a consistent experience across different platforms (MMLab, Lightning, or Anomalib). It also demonstrates how to fine-tune generative AI models, specifically Stable Diffusion SD with LoRA, and the benefits of customized models in reducing latency and enhancing efficiency. The tutorial explores fine-tuning visual prompting tasks, including Segment Anything Model (SAM). It explains how to fine-tune a SD model with custom data using multiple acceleration methods, and how to deploy the fine-tuned model using OpenVINO Transformation Passes API. Lastly, the tutorial focuses on model optimization capabilities for the inference phase, with the OpenVINO toolkit and OTX library integrating with NNCF to refine neural networks and improve inference speed, especially on edge devices with limited resources. The tutorial includes demos showcasing how OpenVINO runtime API enables real-time inference on various devices.

Tutorial: Generalist Agent AI

Organizers: Naoki Wake, Zane Durante, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Yejin Choi, Demetri Terzopoulos, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, Jianfeng Gao, Qiuyuan Huang
Date: Tuesday, June 18
Time: 8:30 AM–12:00 AM
Location: Summit 446

Summary: Generalist Agent AI (GAA) is a family of systems that generate effective actions in an environment based on the understanding of multimodal sensory input. While these systems are expanding into various fields with the advent of large foundation models, they share common interests such as data collection, benchmarking, and ethical perspectives. In this tutorial, we focus on several representative research areas of GAA, including gaming, robotics, and healthcare, and aim to provide comprehensive knowledge on the common concerns discussed in these fields. We expect the participants to learn the fundamentals of GAA and gain insights to further advance their research. Specific learning outcomes include:

- GAA Overview: A deep dive into its principles and roles in contemporary applications, providing attendees with a thorough grasp of its importance and uses.
- Methodologies: Detailed examples of how LLMs and VLMs enhance GAAs, illustrated through case studies.
- Performance Evaluation: Guidance on the assessment of GAAs with relevant datasets.

- Ethical Considerations: A discussion on the societal impacts and ethical challenges of deploying Agent AI, highlighting responsible development practices.
- Future Challenges: A categorization of the latest developments in each domain and a discussion of future directions.

Led by experts from academia and industry, we expect the tutorial to be an interactive and enriching experience. This event will include talks, Q&A sessions, and a panel discussion, ensuring a comprehensive and engaging learning experience for all participants.

Tutorial: All You Need To Know About Point Cloud Understanding

Organizers: Xiaoyang Wu, Hengshuang Zhao, Fuxin Li, Zhijian Liu
Date: Tuesday, June 18
Time: 9:00 AM–12:15 PM
Location: Summit 444

Summary: Unstructured point clouds serve as a sparse representation of the 3D world, playing pivotal roles in 3D perception, generation, autonomous driving, virtual/augmented reality, and robotics. Despite their significance, there lacks a comprehensive resource covering state-of-the-art approaches and engineering nuances in deep point cloud networks. This tutorial aims to fill this gap by offering a comprehensive exploration of the subject. It features lectures that progress from classical point cloud backbones to state-of-the-art point transformers, large-scale 3D representation learning (including pre-training technologies), efficient libraries for sparse systems, and diverse applications for deep point cloud networks. Participants will acquire systematic and practical knowledge on managing and extracting robust deep feature representations from point cloud data. They'll also learn to make informed decisions regarding model architectures and data structures when dealing with point cloud data. Armed with these skills, attendees will be well-equipped to comprehend and leverage these models in real-world applications across various fields, including autonomous driving, embodied AI, and other domains grappling with sparse data in low-dimensional Euclidean spaces.

Tutorial: All You Need to Know about Self-Driving

Organizers: Raquel Urtasun, Ioan Andrei Bârsan, Sergio Casas, Abbas Sadat, Sivabalan Manivasagam
Date: Tuesday, June 18
Time: 9:00 AM–6:00 PM
Location: Summit 445

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.

Since last year's instance (https://waabi.ai/cvpr-2023), countless new and promising avenues of research have started gaining traction, and we have updated our tutorial accordingly. To name a few examples, this includes topics like occupancy forecasting, self-supervised learning, foundation models, the rise of Gaussian Splatting and diffusion models for simulation as well as the study of closed-loop vs. open-loop evaluation.

Tutorial: Computational Design of Diverse Morphologies and Sensors for Vision and Robotics

Organizers: Amir Zamir, Andrew Spielberg, Andrei Atanov
Date: Tuesday, June 18
Time: 9:00 AM–5:00 PM
Location: Summit 344

Summary: Animals exhibit a wide variety of morphologies and sensors, believed to have appeared through billions of years of evolution. Common examples relevant to vision include differences in pupil shapes, the positioning of eyes, various types of eyes, and a varying level of multimodality across animals. Such adaptations are hypothesized to be instances of the so-called Ecological Theory, which posits a strong connection between the specifics of vision and the environment surrounding the agent, its objectives, and its body. How can we replicate this diversity and achieve adaptive design in robotics and vision systems?

In this tutorial, we discuss I) alternative forms of visual sensors that can be useful for real-world robots and II) computational approaches to robot and vision design that can achieve the goal of adaptive design automatically, effectively, and efficiently. The tutorial covers topics in sensing, control, simulation, optimization, and learning-based design for various rigid and soft robots and visual sensors. The material is drawn from state-of-the-art breakthroughs in the field and insights from other disciplines.

This material is accessible to individuals of all backgrounds and levels of expertise.

Tutorial: Learning Deep Low-dimensional Models from High-Dimensional Data: From Theory to Practice

Organizers: Sam Buchanan, Yi Ma, Qing Qu, Yuqian Zhang, Zhihui Zhu
Date: Tuesday, June 18
Time: 9:00 AM–6:00 PM
Location: Summit 442

Summary: Over the past decade, the advent of machine learning and large-scale computing has immeasurably changed the ways we process, interpret, and predict with data in imaging and computer vision. The "traditional" approach to algorithm design, based around parametric models for specific structures of signals and measurements—say sparse and low-rank models—and the associated optimization toolkit, is now significantly enriched with data-driven learning-based techniques, where large-scale networks are pre-trained and then adapted to a variety of specific tasks. Nevertheless, the successes of both modern data-driven and classic model-based paradigms rely crucially on correctly identifying the low-dimensional structures present in real-world data, to the extent that we see the roles of learning and compression of data processing algorithms—whether explicit or implicit, as with deep networks—as inextricably linked. As such, this tutorial provides a timely tutorial that uniquely bridges low-dimensional models with deep learning in imaging and vision. This tutorial will show how: Low-dimensional models and principles provide a valuable lens for formulating problems and understanding the behavior of modern deep models in imaging and computer vision; and how Ideas from low-dimensional models can provide valuable guidance for designing new parameter efficient, robust, and interpretable deep learning models for computer vision problems in practice.

Tutorial: Towards Building AGI in Autonomy and Robotics

Organizers: Li Chen, Andreas Geiger, Huijie Wang, Jiajie Xu
Date: Tuesday, June 18
Time: 9:00 AM–12:00 PM
Location: Summit 447

Summary: In this tutorial, we explore the intersection of AGI technologies and the advancement of autonomous systems, specifically in the field of robotics. We invite participants to embark on an investigative journey that covers essential concepts, frameworks, and challenges. Through discussion, we aim to shed light on the crucial role of fundamental models in enhancing the cognitive abilities of autonomous agents. Through cooperation, we aim to chart a path for the future of robotics, where the integration of AGI enables autonomous systems to push the limits of their capabilities and intelligence, ushering in a new era of intelligent autonomy.

Tutorial: End-to-End Autonomy: A New Era of Self-Driving

Organizers: Long Chen, Oleg Sinavski, Fergal Cotter, Gianluca Corrado, Nikhil Mohan, Vassia Simaiaki, Elahe Arani, Jamie Shotton
Date: Tuesday, June 18
Time: 1:30 PM–6:00 PM
Location: Summit 444

Summary: A comprehensive half-day tutorial focused on End-to-End Autonomous Driving (E2EAD), reflecting the significant shift in focus towards this approach within both industry and academia. Traditional modular approaches in autonomous driving, while effective in specific contexts, often struggle with scalability, long-tail scenarios, and compounding errors from different modules, thereby paving the way for the end-to-end paradigm. This tutorial aims to dissect the complexities and nuances of end-to-end autonomy, covering theoretical foundations, practical implementations and validations, and future directions of this evolving technology.A comprehensive half-day tutorial focused on End-to-End Autonomous Driving (E2EAD), reflecting the significant shift in focus towards this approach within both industry and academia. Traditional modular approaches in autonomous driving, while effective in specific contexts, often struggle with scalability, long-tail scenarios, and compounding errors from different modules, thereby paving the way for the end-to-end paradigm. This tutorial aims to dissect the complexities and nuances of end-to-end autonomy, covering theoretical foundations, practical implementations and validations, and future directions of this evolving technology.

Tutorial: From Multimodal LLM to Human-level AI: Modality, Instruction, Reasoning and Beyond

Organizers: Hao Fei, Yuan Yao, Ao Zhang, Haotian Liu, Fuxiao Liu, Zhuosheng Zhang, Shuicheng Yan
Date: Tuesday, June 18
Time: 1:30 PM–6:00 PM
Location: Summit 446

Summary: Artificial intelligence (AI) encompasses knowledge acquisition and real-world grounding across various modalities. As a multidisciplinary research field, multimodal large language models (MLLMs) have recently garnered growing interest in both academia and industry, showing an unprecedented trend to achieve human-level AI via MLLMs. These large models offer an effective vehicle for understanding, reasoning, and planning by integrating and modeling diverse information modalities, including language, visual, auditory, and sensory data. This tutorial aims to deliver a comprehensive review of cutting-edge research in MLLMs, focusing on three key areas: MLLM architecture design, instructional learning, and multimodal reasoning of MLLMs. We will explore technical advancements, synthesize key challenges, and discuss potential avenues for future research.

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

Organizers: Maying Shen, Hongxu Yin, Jason Clemons, Pavlo Molchanov, Jan Kautz, Jose M Alvarez
Date: Tuesday, June 18
Time: 1:30 PM–5:00 AM
Location: Summit 447

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: Diffusion-based Video Generative Models

Organizers: Mike Zheng Shou, Jay Zhangjie Wu, Deepti Ghadiyaram

Date: Tuesday, June 18

Time: 2:00 PM–5:00 PM

Location: Summit 437-439

Summary: In the past year, the landscape of video generation has transformed dramatically, achieving remarkable strides from rudimentary outputs to strikingly realistic videos. Central to this evolution are diffusion models, which have become a cornerstone technology in pushing the boundaries of what's possible in video generation. This tutorial will delve into the critical role of diffusion models in video generation and modeling.

Participants will engage in a deep dive into the broad spectrum of topics related to video generative models. We will start with the foundational elements, including the core principles of video foundation models. The session will then extend to explore specific applications such as image-to-video animation, video editing, and motion customization. A significant focus will also be placed on the evaluation of video diffusion models, as well as on safety technologies to mitigate the potential risks of using these models.

Attendees will leave this tutorial with a comprehensive understanding of both fundamental techniques and the cutting-edge advancements in diffusion-based video modeling, fully equipped to navigate and contribute to the rapidly evolving field in the GenAI era.

Tutorial: Unifying Spectral and Spatial Graph Neural Networks

Organizers: Zhiqian Chen, Lei Zhang, Liang Zhao

Date: Tuesday, June 18

Time: 2:00 PM–5:00 PM

Location: Summit 440-441

Summary: Over recent years, Graph Neural Networks (GNNs) have garnered significant attention. However, the proliferation of diverse GNN models, underpinned by various theoretical approaches, complicates the process of model selection, as they are not readily comprehensible within a uniform framework. Specifically, early GNNs were implemented using spectral theory, while others were developed based on spatial theory . This divergence between spectral and spatial methodologies renders direct comparisons challenging. Moreover, the multitude of models within each domain further complicates the evaluation of their respective strengths and weaknesses.

In this half-day tutorial, we examine the state-of-the-art in GNNs and introduce a comprehensive framework that bridges the spatial and spectral domains, elucidating their complex interrelationship. This emphasis on a comprehensive framework enhances our understanding of GNN operations. The tutorial's objective is to explore the interplay between key paradigms, such as spatial and spectral-based methods, through a synthesis of spectral graph theory and approximation theory. We provide an in-depth analysis of the latest research developments in GNNs in this tutorial, including discussions on emerging issues like over-smoothing. A range of well-established GNN models will be utilized to illustrate the universality of our proposed framework.

Notes:

WORKSHOPS

Transformers for Vision

Organizers: Gedas Bertasius, Rohit Girdhar, Zhiding Yu, Lucas Beyer, Gul Varol, Ce Zhang, Feng Cheng, Yan-Bo Lin, Md Mohaiminul Islam, Yi-Lin Sung, Jaemin Cho, Yue Yang, Xin Wang, Mohit Bansal, Alaaeldin El-Nouby, Tyler Zhu

Date: Tuesday, June 18

Time: 7:50 AM–6:00 PM

Location: Summit 347-348

Summary: Over the last few years, the field of natural language processing (NLP) has been revolutionized by the emergence of transformer models. These models have recently 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.

VizWiz Grand Challenge: Describing Images and Videos Taken by Blind People

Organizers: Danna Gurari, Jeffrey Bigham, Ed Cutrell, Daniela Massiceti, Abigale Stangl, Chongyan Chen, Everley Tseng, Josh Myers-Dean

Date: Tuesday, June 18

Time: 8:00 AM–12:05 PM

Location: Summit 435

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 six dataset challenges, where the tasks are to answer visual questions, ground answers to visual questions, recognize visual questions with multiple answer groundings, recognize objects in few-shot learning scenarios, locate objects in few-shot learning scenarios, and classify images in a zero-shot setting. 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.

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

Organizers: Chris Padwick, Humphrey Shi, Naira Hovakimyan, Pan Zhao, Jing Wu, Leandro Almeida, Jim Ostrowski, Ripudman Arora, Ignacio Ciampitti, Leonardo Bosche, Kai Wang

Date: Tuesday, June 18

Time: 8:00 AM–6:00 PM

Location: Arch 3B

Summary: With the recent success of computer vision and deep learning in various applications, there has been significantly increasing attention towards its use in agriculture. Agriculture-related vision problems are of great economic and social value. For example, effective monitoring and tracking of crops and anomalies in farmlands give higher yields and can benefit millions of farmers and consumers. At the same time, such problems are very challenging, both in terms of the underlying research tasks and their adaptation to the real-world scenarios. To encourage research in the integration of computer vision and agriculture, we propose to organize the 5th International Workshop and Challenge on Agriculture-Vision: Challenges and Opportunities for Computer Vision in Agriculture. With our team's strong dedication and expertise, and given the great success of its first three editions at CVPR 2020-2023, we are confident that this workshop will become a highly visible venue and valuable addition to CVPR 2024. The organization team features a strong diversity in experiences, gender, race, and hands-on levels. Importantly, for generating significant challenge participation, the organizing team has secured a total of at least $5,000 in sponsorship for challenge winners confirmed by Blue River Technology. Also, several world-renowned/well-positioned invited speakers plan to speak at the workshop from diverse backgrounds: not only those from academic communities such as computer vision, robotics, and agriculture, but also pioneer practitioners from related top industry sectors. This workshop provides a great opportunity to both demonstrate current progress in such interdisciplinary areas and encourage further research and applications on The Future Of Agriculture.

Face Recognition Challenge in the Era of Synthetic Data

Organizers: Ruben Tolosana, Ivan DeAndres-Tame, Pietro Melzi, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia

Date: Tuesday, June 18

Time: 8:10 AM–12:00 PM

Location: Arch 212

Summary: Synthetic data is gaining increasing relevance for training machine learning models. This is mainly motivated due to several factors such as the lack of real data and intra-class variability, time and errors produced in manual labeling, and in some cases privacy concerns, among others. To promote and advance the use of synthetic data for face recognition, we organize the Second Edition of the Face Recognition Challenge in the Era of Synthetic Data (FRCSyn). This challenge aims to investigate the use of synthetic data in face recognition to address current technological limitations, including data privacy concerns, demographic biases, generalization to novel scenarios, and performance constraints in challenging situations such as aging, pose variations, and occlusions. Unlike the first edition, in which synthetic data from DCFace and GANDiffFace methods was only allowed to train

face recognition systems, in this second edition we propose new subtasks that allow participants to explore novel face generative methods. The outcomes of the Second Edition FRCSyn Challenge, along with the proposed experimental protocol and benchmarking contribute significantly to the application of synthetic data to face recognition.

Computer Vision for Materials Science

Organizers: Alexei Skurikhin, Christopher Eberl, Kari Sentz, Katherine Sytwu

Date: Tuesday, June 18

Time: 8:10 AM–12:50 PM

Location: Arch 214

Summary: Computer vision and machine learning are emerging as critical tools for materials characterization. In materials science, a variety of microscopy image data (e.g., optical microscopy, electron microscopy, x-ray microscopy) are used to understand material properties. Many of these microscopy techniques make it easy and inexpensive to collect large, complex image data sets that can overwhelm the available time of the subject matter experts required to interpret it. Other microscopy techniques, such as analytical spectroscopic imaging, are much more time consuming, but have the potential to be accelerated through joint collection and modeling with more scalable imaging modalities. In addition, large video and 3D datasets (e.g., produced by new high-speed electron detectors and X-ray computed tomography) are nearly impossible to manually quantify. The aim of the Computer Vision for Materials Science (CV4MS) workshop is to bring together cross-disciplinary researchers to demonstrate recent advancements in machine learning, computer vision, and materials microscopy, and discuss open problems such as explainability, uncertainty quantification, and representation learning in materials microscopy analysis. There will be a focus on the unique challenges and advantages in materials characterization including limited ground-truth labels, multimodal datasets, and physics-based constraints, which could lead to new computer vision frameworks for the broader scientific imaging community.

The Future of Generative Visual Art

Organizers: Anyi Rao, Aleksander Holynski, Jon Barron, Fabian Caba, Ruihang Zhang, Mia Tang, Yuwei Guo, Victor Escorcia, Linning Xu, Jean-Peic Chou, Elia Peruzzo, Yu Xiong, Alejandro Pardo, Ali Thabet, Dong Liu, Dahua Lin, Bernard Ghanem, Angjoo Kanazawa, Alexei A. Efros, Maneesh Agrawala

Date: Tuesday, June 18

Time: 8:20 AM–5:40 PM

Location: Summit 343

Summary: The generative models that have been developed and publicly released over these past couple of years have developed a great following online, with artists regularly posting new creations on Twitter and other social media platforms. Despite the fact that these generative tools are used so widely to create art, very little connection exists between the artists who create the art and the researchers who create the tools. In this workshop, we intend to bring these two groups together, in the interest of developing ideas and directions for the future of generative visual art.

Omnidirectional Computer Vision Workshop

Organizers: Kaavya Rekanar, Priyadarshi Sweta Singh, Ciarán Eising, Varun Ravi Kumar, Li Guan, Jonathan Horgan, Stefan Milz, Pierre Moulon, Senthil Yogamani, Fatih Porikli

Date: Tuesday, June 18

Time: 8:30 AM–12:00 PM

Location: Arch 205

Summary: From immersive experiences in autonomous driving to medical imaging and more, we see from expansive research that maximizing a camera's field of view can solve real-world problems. In the past few years, omnidirectional imaging has grown in interest, driven in part by the desire to maximize the amount of content and context encapsulated by a single image. Fisheye cameras installed in modern vehicles and commodity omnidirectional cameras from such companies as Ricoh and Insta360 have helped increase the popularity of this imaging modality. They have opened the door for more consumer-facing applications, such as automotive, home services, and real estate industries. Our workshop seeks to link 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.

Human Motion Generation

Organizers: Peizhuo Li, Rishabh Dabral, Sigal Raab, Guy Tevet, Chuan Guo, Ikhsanul Habibie, Amit Haim Bermano, Christian Theobalt

Date: Tuesday, June 18

Time: 8:30 AM–12:00 PM

Location: Summit 430

Summary: Motion is one of the fundamental attributes of human (and animal) life and underlies our actions, gestures as well as our behavior. The capture and synthesis of human motion have been among the core areas of interest for the CVPR community and facilitate a variety of applications such as avatar creation, 3D character animations, AR/VR, crowd simulation, sports analytics and many more. The prime goal of the workshop is to bring the human motion synthesis community together and foster discussions about the existing challenges and future direction. To enable this, we feature invited talks presented by a diverse group of leading experts spanning a variety of sub-domains. With this workshop, we hope to encourage cross-pollination of ideas coming from different vantage points as well as discuss the gap between the academic and the industrial perspectives of the topic.

Mobile Intelligent Photography & Imaging

Organizers: Xiaoming Li, Zongsheng Yue, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Yuekun Dai, Peiqing Yang, Chunle Guo, Xin Jin, Yaqi Wu, Dafeng Zhang, Jimmy S. Ren, Chen Change Loy

Date: Tuesday, June 18

Time: 8:30 AM–12:20 PM

Location: Arch 213

Summary: This workshop focuses on Mobile Intelligent and Photography Imaging (MIPI). It is closely connected to the impressive advancements of computational photography and imaging on mobile platforms (e.g., phones, AR/VR devices, and automatic cars), especially with the explosive growth of new image sensors and camera systems. Currently, the demand for developing and perfecting advanced image sensors and camera systems is rising rapidly. Meanwhile, new sensors and camera systems present interesting and novel research problems to the community. Moreover, the limited computing resources on mobile devices further compound the challenges, as it requires developing lightweight and efficient algorithms. However, the lack of high-quality data for research and the rare opportunity for an in-depth exchange of views from industry and academia constrain the development of mobile intelligent photography and imaging. With the consecutive success of the 1st MIPI Workshop@ECCV 2022 and the 2nd MIPI Workshop@CVPR 2023, we will continue to arrange new sensors and imaging systems-related competition with industry-level data, and invite keynote speakers from both industry and academia to fuse the synergy. In this MIPI workshop, the competition will include three tracks: few-shot raw image denoising, demosaic with defect pixels for hybridevs camera, and Nighttime Flare Removal. MIPI wishes to gather researchers and engineers together, encompassing the challenging issues and shaping future technologies in the related research directions.

Gaze Estimation and Prediction in the Wild

Organizers: Hyung Jin Chang, Xucong Zhang, Shalini De Mello, Thabo Beeler, Seonwook Park, Jean-Marc Odobez, Yihua Cheng, Xi Wang, Otmar Hilliges, Aleš Leonardis

Date: Tuesday, June 18

Time: 8:30 AM–12:30 PM

Location: Arch 309

Summary: Intelligent computer systems should be able to anticipate human intentions in order to present the most appropriate information or enable efficient interactions. The clearest single indicator for human attention is eye gaze and the eye movement patterns of the user. The many applications of gaze tracking include crowd-sourced attention studies, adaptive user interfaces, AR/VR, and driver monitoring in vehicles. Many such applications are expected to be performed in environments beyond the laboratory where low image quality and non-ideal lighting conditions can pose significant challenges to existing algorithms. Unlike many other areas in computer vision, state-of-the-art deep learning methodologies have been introduced relatively slowly to address these challenges in the gaze estimation task because of its high complexity, lack of available diverse large datasets, and relatively small community size. By organizing the successful GAZE workshop series, GAZE2019 (at ICCV2019), GAZE2020 (at ECCV2020), GAZE2021 (at CVPR2021), GAZE2022 (at CVPR2022), and GAZE2023 (at CVPR2023) workshops, we have played a pivotal role in bringing together researchers from both academia and industry related to eye gaze, and have successfully set up a venue to share current research achievements and discuss future research directions together. In this new edition, the 6th GAZE workshop at CVPR2024, we especially aim to encourage and highlight novel strategies for eye gaze estimation and prediction with a focus on synthetic eye gaze dataset generations and various potential applications including VR/AR and driver monitoring, etc.

Responsible Generative AI

Organizers: Adriana Romero-Soriano, Michal Drozdzal, Melissa Hall, Agata Lapedriza, Ye Zhu, Negar Rostamzadeh, Golnoosh Farnadi, Utsav Prabhu, Raesetje Sefala

Date: Tuesday, June 18

Time: 8:30 AM–12:30 PM

Location: Summit 433

Summary: In the recent year, we have witnessed remarkable progress in both visual and multi-modal generative models. The impressive results achieved by these models has propelled an arms race towards their widespread use in content creation applications, positioning generative models at the core of the current AI revolution. Yet, works studying the biases in vision and multi-modal systems trained on large scale Internet crawled data have mainly focused on discriminative and representation learning models. With the unprecedented success of recently released high performing multi-modal generative systems, there is an urgent need to build deep understanding and open the discussion around the responsible use of these high performing models, and their impact on our society. Thus, this workshop aims to bring together researchers, practitioners, and industry leaders working at the intersection of generative AI, vision, data, ethics, privacy and regulation, with the goal of discussing existing concerns, and brainstorming possible avenues forward to ensure the responsible progress of generative AI. We hope that the topics addressed in this workshop will constitute a crucial step towards ensuring a positive experience with generative AI for everyone.

Advances in Radiance Fields for the Metaverse

Organizers: Aayush Prakash, Daeil Kim, Peter Vajda, Fernando De la Torre, Angela Dai

Date: Tuesday, June 18

Time: 8:30 AM–12:30 PM

Location: Summit 332

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 the content inside it can be costly and time consuming. In 2020, new techniques were developed in neural volume rendering, also known as NeRF (Neural Radiance Fields), which brought an explosion of new work that has direct applicability to the future metaverse. In CVPR 2023, there were more than 100+ published papers aimed towards improving the fidelity, efficiency, scalability and application of RFs. With the recent development of Gaussian Splatting, we are also seeing alternative approaches to neural networks driving advancements as well. We believe that these techniques represent some of the most viable solutions to address the growing content needs of Metaverse. There have been many recent advances in Radiance Fields (RF- NeRF, Gaussian Splatting, etc.) that have enabled them 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 on mobile devices, c) generative RF through natural language, d) spatio-temporal RF of videos, and e) stylizing and editing RF. This workshop is an opportunity to showcase work on RF 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 general state of this technology. Specifically, we would like to cover recent advances in radiance fields that focus on three areas that need significant gains for the metaverse: scale, efficiency, and fidelity.

Test-Time Adaptation: Model, Adapt Thyself!

Organizers: Evan Shelhamer, Evgenia Rusak, Steffen Schneider, Motasem Alfarra, Francesco Croce, Xiaolong Wang, Dequan Wang

Date: Tuesday, June 18

Time: 8:30 AM–12:45 PM

Location: Summit 324

Summary: Will every use case of deep learning for vision be solved in the limit of more data, parameters, and computation for training? We propose that learning during training alone is not enough, and argue that learning during testing by test-time training / test-time adaptation (TTT/TTA) is helpful and complementary. Our workshop aims to unite researchers on adaptation and robustness to push the boundaries between training and testing by enabling systems to keep updating during deployment. Our focus is on updating to maintain or improve accuracy, calibration, and fairness on changing data in diverse settings. Our scope encompasses data, evaluation, algorithms, and unresolved challenges for test-time updates while emphasizing unsupervised adaptation with minimal computational overhead. Join us to keep learning and never stop updating!

"What is Next in Multimodal Foundation Models?"

Organizers: Rogerio Feris, Leonid Karlinsky, Nina Shvetsova, Sivan Doveh, Muhammad Jehanzeb Mirza, Eli Schwartz, Wei Lin, Paola Cascante-Bonilla, Roei Herzig, Bo Wu, Irene Huang, Amit Alfassy, Sivan Harary, Assaf Arbelle, Hilde Kuehne, Aude Oliva

Date: Tuesday, June 18

Time: 8:30 AM–1:00 PM

Location: Summit 437-439

Summary: The term "Foundation Models" has been generally used to denote large-scale models (e.g., with billions of parameters) pre-trained on massive-scale datasets, which can be further adapted to a variety of downstream tasks with little or no supervision. In recent years, these big models have taken the AI world by storm, significantly advancing the state of the art in computer vision, natural language processing, speech analysis, and other fields. In particular, multimodal foundation models, which are trained with multiple modalities simultaneously, have shown remarkable success in a wide range of applications, including text-to-image/video/3D generation, zero-shot classification, cross-modal retrieval, and many others. The purpose of this workshop is to create a forum for discussion on what is next in multimodal foundation models, i.e. what are the paths forward and the fundamental problems that still need to be addressed in this emerging research area. We will bring a diverse set of leaders in the field to deliver talks, present their views, and engage in a discussion with our community on the various facets of multimodal foundation models, including, but not limited to, the models' design, generalization properties, efficiency, ethics, fairness, scale, and open availability. We will also have a poster session, a panel discussion, and will host a challenge to study multimodal document analysis and beyond.

Notes:

What is Next in Video Understanding?

Organizers: Davide Moltisanti, Hazel Doughty, Michael Wray, Bernard Ghanem, Lorenzo Torresani

Date: Tuesday, June 18

Time: 8:30 AM–1:00 PM

Location: Summit 335-336

Summary: Video understanding has seen tremendous progress over the past decades, with models attaining excellent results on classic benchmarks and widespread applications in industry. However, there are still many open challenges and the path forward is not clear. Image-text models have pushed progress on current video benchmarks, but we are still lacking in video foundation models. Recent works have shown several popular benchmarks can be solved using single video frames. New benchmark tasks are frequently being proposed, however their adoption is often limited and researchers default to saturated tasks such as classification and localisation. The ever-growing computational requirements are also particularly demanding for this field, which curbs the accessibility of video understanding. This raises the questions of how much deeper our understanding of video needs to be, whether the current benchmarks we have are enough, how to work in this field with limited resources, and what comes next beyond the task of action recognition. The purpose of this workshop is to create a forum for discussion on what is next in video understanding, i.e. what are the paths forward and the fundamental problems that still need to be addressed in this research area. We will engage in a discussion with our community on the various facets of video understanding, including, but not limited to, tasks, model design, video length, multi-modality, generalisation properties, efficiency, ethics, fairness, and scale.

Women in Computer Vision

Organizers: Sachini Herath, Asra Aslam, Ziqi Huang, Estefania Talavera, Vanessa Staderini, Azade Farshad, Himangi Mittal, Deblina Bhattacharjee, Mengwei Ren

Date: Tuesday, June 18

Time: 8:30 AM–1:30 PM

Location: Arch 201

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 female faculty members and researchers both in academia and in industry is still relatively low. As a result, many female researchers working 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 the visibility of female computer vision researchers through invited keynotes by leading women researchers; provide opportunities for junior female researchers to present their work via oral/ poster sessions and travel awards; share experience and career advice for female students and professionals via mentoring/ networking event. The half-day Women in Computer Vision workshop is a gathering for researchers of all genders and career stages. Travel grants will be offered to selected female presenters of oral and poster sessions.

Visual Odometry and Computer Vision Applications Based on Location Clues

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

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 330

Summary: Visual odometry and localization have maintained an increasing interest in recent years, especially with the extensive applications for autonomous driving, augmented reality, and mobile computing. With the location information obtained through odometry, services based on location clues are also rapidly emerging. Particularly, in this workshop, we focus on mobile and robot platform applications. This workshop invites papers in the areas including advances in visual odometry and computer vision applications based on location context.

Compositional 3D Vision

Organizers: Habib Slim, Wolfgang Heidrich, Peter Vajda, Natalia Neverova, Mohamed Elhoseiny

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 327

Summary: The C3DV workshop focuses on compositional approaches for understanding and generating 3D visual data. Topics include 3D scene understanding through object detection and segmentation, part-aware shape analysis and modeling, and data-driven techniques for creating novel 3D content such as part-based shape generation and scene synthesis. The workshop features an exciting line of keynote speakers and data challenges aimed at driving research in this field forward.

Computer Vision in Sports

Organizers: Rikke Gade, Thomas Moeslund, Graham Thomas, Adrian Hilton, James Little, Michele Merler, Silvio Giancola, Anthony Cioppa

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Arch 3A

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, a significant quantity of resources has been channeled into sports in order to improve understanding, performance, and presentation. For example, areas like performance assessment, which were previously mainly of interest to coaches and sports scientists are now finding applications in broadcast and other media, driven by the increasing use of on-line sports viewing which provides a way of making 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. Computer vision algorithms have a huge potential in many aspects of sports ranging from automatic annotation of broadcast footage, through to better understand of sport injuries, coaching, 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.

Data-Driven Autonomous Driving Simulation

Organizers: Maximilian Igl, Zan Gojcic, Azadeh Dinparastdjadid, Maximilian Naumann, Or Litany, Thomas Gilles, Katie Luo, Anqi Joyce Yang, Peter Karkus, Yiren Lu, Jonah Philion, Yue Wang, Xinshuo Weng, Jiawei Yang, Shimon Whiteson, Marco Pavone, Sanja Fidler

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 342

Summary: Real-world on-road testing of autonomous vehicles can be expensive or dangerous, making simulation a crucial tool to accelerate the development of safe autonomous driving (AD), a technology with enormous real-world impact. However, to minimise the sim-to-real gap, good agent behaviour models and sensor/perception imitation are paramount. A recent surge in published papers in this fast-growing field has led to a lot of progress, but several fundamental questions remain unanswered, for example regarding the fidelity and diversity of generative behaviour and perception models, generation of realistic controllable scenes at scale and the safety assessment of the simulation toolchain. In this workshop, our goal is to bring together practitioners and researchers from all areas of AD simulation and to discuss pressing challenges, recent breakthroughs and future directions.

Efficient Deep Learning for Computer Vision

Organizers: Hongxu Yin, Bichen Wu, Pavlo Molchanov, Peizhao Zhang, Andrew Howard, Chas Leichner, Xiaoliang Dai, Ji Hou, Peter Vajda, Dilin Wang, Jan Kautz

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 420-422

Summary: Computer Vision has a long history of academic research, and recent advances in deep learning have provided significant improvements in the ability to understand visual content. As a result of these research advances on problems such as object classification, object detection, and image segmentation, there has been a rapid increase in the adoption of Computer Vision in industry; however, mainstream Computer vision research has given little consideration to speed or computation time, and even less to constraints such as power/energy, memory footprint and model size, or carbon emission. Nevertheless, addressing all of these metrics is essential if advances in Computer Vision are going to be widely available on mobile and AR/VR devices.

Continual Learning in Computer Vision

Organizers: Marc Masana, Gido M. van de Ven, Pau Rodriguez, Dhireesha Kudithipudi, Tyler Hayes, Andrea Cossu, James Seale Smith, Gianluca Guglielmo, Benedikt Tscheschner, Hamed Hemati, Lama Alssum

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 325

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 have to 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 challenge 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.

Equivariant Vision: From Theory to Practice

Organizers: Congyue Deng, Jiahui Lei, Yinshuang Xu, Li Yi, Christine Allen-Blanchette, Vitor Guizilini, Ameesh Makadia, Kostas Daniilidis

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 321

Summary: Exploiting symmetry in structured data is a powerful way to improve the generalization ability, data efficiency, and robustness of AI systems, which leads to the research direction of equivariant deep learning. Showing its effectiveness, it has been widely adopted in a large variety of subareas of computer vision, from 2D image analysis to 3D perception, as well as further applications such as medical imaging and robotics. Our topics include but are not limited to: Theoretical foundations of equivariant deep learning with symmetry and group theory, Equivariance by design: Neural network architectures and mathematical guarantees, Equivariance from data: Learning equivariant and invariant features, Applications in 2D and 3D computer vision and robotics, Applications in broader science: computational biology, medicine, natural science, etc, Equivariance in the large-model era and potential future directions.

Generative Models for Computer Vision

Organizers: Adam Kortylewski, Fangneng Zhan, Lingjie Liu, Michael Niemeyer, Michael Oechsle, Alan Yuille, Christian Theobalt

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 432

Summary: Recent advances in generative modeling have enabled the synthesis of near photorealistic images, thus 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 2nd edition of the workshop, we again 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 computer vision benefit from the advances in generative image modeling?".

Visual Perception via Learning in an Open World

Organizers: Shu Kong, Yanan Li, Neehar Peri, Yu-Xiong Wang, Andrew Owens, Deepak Pathak, Carl Vondrick, Abhinav Shrivastava

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 328

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, langauges 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.

Explainable AI for Computer Vision

Organizers: Indu Panigrahi, Sunnie S. Y. Kim, Vikram V. Ramaswamy, Sukrut Rao, Stefan Kolek, Lenka Tětková, Jawad Tayyub, Katelyn Morrison, Pushkar Shukla, Deepti Ghadiyaram

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Arch 2A

Summary: Explainability of computer vision systems is critical for people to effectively use and interact with them. The 3rd 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.

Responsible Data

Organizers: Candice Schumann, Caner Hazirbas, Olga Russakovsky, Vikram Ramaswamy, Jerone Andrews, Alice Xiang, Susanna Ricco, Courtney Heldreth, Biao Wang, Christian Canton Ferrer, Jess Holbrook

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Arch 303

Summary: The development of large-scale datasets has been essential to the progress of machine learning and artificial intelligence. However, many of these datasets are not inclusive or diverse–particularly computer vision datasets, which can lead to biased models and algorithms. This workshop will bring together practitioners and researchers to discuss the challenges and opportunities of building more responsible datasets. This workshop will cover a range of topics, including:–Moving beyond pragmatism and implementation of context and consent-driven procedures in dataset development–What are the main themes when it comes to responsible datasets? Are there specific benchmarks currently utilized?–Challenges, risks and benefits of collecting gender, race, skin tone, physical attributes, accessibility data, and other person attributes.–What are the best practices when training individuals for data collection and annotators. To what extent does diversity matter when it comes to data collection and annotators? How the organizational structures of these businesses and the ecosystem of stakeholders contribute to the responsible dimension of the datasets?–What are the new considerations in a world of pretrained models and synthetic data?–How should we build responsible datasets for generative AI models and applications?–How do we quantitatively measure how responsible a dataset is?–What does Transparency translate to in the context of dataset development?–How do notions of Data Privacy like those articulated in proposals such as the Blueprint for an Bill of Rights translate to building towards responsible datasets?–How do we build a framework for Dataset Accountability?–How should we best engage the open source community when building, updating, and maintaining datasets?–State of Affairs: a summary of progress to date–how responsible datasets have evolved. What best practices can be leveraged more broadly?"

Synthetic Data for Computer Vision

Organizers: Jieyu Zhang, Cheng-Yu Hsieh, Zixian Ma, Shobhita Sundaram, Wei-Chiu Ma, Phillip Isola, Ranjay Krishna

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 423-425

Summary: The workshop aims to explore the use of synthetic data in training and evaluating computer vision models, as well as in other related domains. During the last decade, advancements in computer vision were catalyzed by the release of painstakingly curated human-labeled datasets. Recently, people have increasingly resorted to synthetic data as an alternative to laborintensive human-labeled datasets for its scalability, customizability, and costeffectiveness. Synthetic data offers the potential to generate large volumes of diverse and high-quality vision data, tailored to specific scenarios and edge cases that are hard to capture in real-world data. However, challenges such as the domain gap between synthetic and real-world data, potential biases in synthetic generation, and ensuring the generalizability of models trained on synthetic data remain. We hope the workshop can provide a forum to discuss and encourage further exploration in these areas.

Computer Vision with Humans in the Loop

Organizers: Lei Zhang, Gang Hua, Nicu Sebe, Kristen Grauman, Yasuyuki Matsushita, Aniruddha Kembhavi, Ailing Zeng, Jianwei Yang, Xi Yin, Heung-Yeung Shum

Date: Tuesday, June 18

Time: 8:30 AM–5:45 PM

Location: Summit 329

Summary: This workshop aims to explore the pivotal role of human interaction in advancing computer vision technologies. Despite significant progress over the past two decades, computer vision systems often fall short of human capabilities, especially in specialized or complex scenarios. Featuring a series of invited talks and panels, this workshop will highlight innovations like interactive segmentation, advancements in language and visual prompt integration, and the development of self-aware systems capable of recognizing and compensating for their limitations. Distinguished speakers from both academia and industry will share their insights, offering a rich dialogue on the integration of human cognitive skills with machine precision to tackle the challenges of computer vision. Participants will gain an understanding of the evolving landscape of human-in-the-loop methodologies and their impact on practical applications and foundational models in the field. Join us to discuss historical perspectives, current research, and future directions in this dynamic area.

RetailVision – Field Overview and Amazon Deep Dive

Organizers: Ehud Barnea, Yosi Keller, Marina Paolanti, Bruno Artacho, Austen Groener, Yin Wang, Weijian Li, Sean Ma, Ananth Sadanand, Mohsen Malmir

Date: Tuesday, June 18

Time: 8:30 AM–6:00 PM

Location: Arch 310

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 online shopping, traditional markets also quickly embraced AI-related technology solutions at the physical store level. Following the introduction of computer vision to the world of retail, a new set of challenges emerged. These challenges were further expanded with the introduction of image and video generation capabilities. The physical domain exhibits challenges such as the detection of shopper and product interactions, fine-grained recognition of visually similar products, as well as new products that are introduced on a daily basis. The online domain contains similar challenges, but with their own twist. Product search and recognition is performed on more than 100,000 classes, each including images, textual captions, and text by users during their search. In addition to discriminative machine learning, image generation has also started being used for the generation of product images and virtual try-on. All of these challenges are shared by different companies in the field, and are also at the heart of the computer vision community. This workshop aims to present the progress in these challenges and encourage the forming of a community for retail computer vision.

LatinX in Computer Vision Research Workshop

Organizers: Nils Murrugarra, Rodolfo Valiente, Hoover Rueda, Juan Gonzales, Willams de Lima, Andrés Villa, Francisco López, Jorge Bacca, Gilberto Ochoa, Dustin Carrión, Iván Reyes, Eduardo Moya, Daniela Herrera, Lidia Talavera, Ariana Villegas, Miguel Arevalo, Pedro Braga, Jose Sosa

Date: Tuesday, June 18

Time: 8:30 AM–6:00 PM

Location: Arch 203

Summary: Computer vision has made a lot of progress in many different areas such as image recognition, object detection, image segmentation, and visual search. Despite its crucial importance in our digital life, there is a relatively low representation of LatinX researchers, engineers, and educators. Hence, our goal is to promote and increase the representation of the LatinX community in computer vision. Despite a lack of resources, mentors, role models, and support networks, we want to encourage members of the LatinX community to participate in the broad field of computer vision as well as participate in the discussion of important and relevant issues of societal impact and in the pursuit of their solutions. By having a focused venue to highlight research from the LatinX community, we aim to foster an increased representation in computer vision and the development of more inclusive technology. We consider that our community has unique points of view that can enhance future research directions in the fascinating field of computer vision.

Computer Vision for Microscopy Image Analysis

Organizers: Mei Chen, Cassiano Carromeu, Dimitris Metaxas, Steven Finkbeiner

Date: Tuesday, June 18

Time: 8:30 AM–6:00 PM

Location: Summit 431

Summary: Advances in imaging technologies have enabled the acquisition of large volumes of microscopy images and made it possible to conduct large-scale, image-based experiments for biomedical discovery. Computer vision and machine learning-based methodologies coupled with domain knowledge have immense potential for automating the analysis and understanding of such data. There is a strong need for biologists and computer vision experts to collaborate to overcome the limits of human visual perception and leverage state-of-the-art machine learning methods to enable quantitative, interpretable, explainable, and high-content analysis of phenotypic traits. The Workshop on Computer Vision for Microscopy Image Analysis aims to create a comprehensive forum to encourage research and in-depth discussion of technical, application, and community issues. The emphasis of the 2024 workshop will be leveraging the advances in large foundational models (LFMs) to improve the multimodal analysis of microscopy and omics data with enhanced explainability. We aim to promote discussions on methods that offer explainability, interpretability, and usefulness for disease classification, prediction, and treatment, while integrating domain knowledge with LFMs. The goal is to strengthen and nurture the community through interdisciplinary collaboration that pushes the envelope of what is possible when the worlds of microscopy, omics, and computer vision come together to disrupt biomedical discovery and improve the quality of life for all.

Fine-grained Visual Categorization

Organizers: Nico Lang, Elijah Cole, Suzanne Stathatos, Abby Stylianou, Srishti Yadav, Jennifer Sun, Lukas Picek, Kimberly Wilber, Omiros Pantazis, Justin Kay, Jong-Chyi Su, Xiangteng He, Serge Belongie, Oisin Mac Aodha, Subhransu Maji, Sara Beery, Grant Van Horn

Date: Tuesday, June 18

Time: 8:45 AM–4:45 PM

Location: Summit 326

Summary: It may be tempting to think that image classification is a solved problem. However, one only needs to look at the poor performance of existing techniques in domains with limited training data and highly similar categories to see that this is not the case. In particular, fine-grained categorization, e.g., 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 performance. In these domains, expert knowledge is typically required, and the question that must be addressed is how we can develop artificial systems that can efficiently discriminate between large numbers of highly similar visual concepts. The 11th Workshop on Fine-Grained Visual Categorization (FGVC11) will explore topics related to supervised learning, self-supervised learning, semi-supervised learning, vision and language, 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, anomaly detection, open-set recognition, human-in-the-loop learning, and taxonomic prediction, all through the lens of fine-grained understanding. Hence, the relevant topics are neither restricted to vision nor categorization. Our workshop is structured around five main components: (i) invited talks from world-renowned computer vision experts, (ii) invited talks from experts in application domains (e.g., medical science and ecology), (iii) interactive discussions during poster and panel sessions, (iv) novel fine-grained challenges that are hosted as part of the workshop, and (v) peer-reviewed extended abstract paper submissions. We aim to stimulate debate and to expose the wider computer vision community to new and challenging problems in areas that have the potential for large societal impact but do not traditionally receive a significant amount of exposure at other CVPR workshops.

Computational Cameras and Displays

Organizers: Salman Asif, Yi Xue, Mark Sheinin, Kristina Monakhov

Date: Tuesday, June 18

Time: 8:45 AM–4:55 PM

Location: Arch 204

Summary: Computational photography has become an increasingly active area of research within the computer vision community. Within the last few 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 features keynote and invited talks from leading experts, poster presentations from a diverse set of researchers, and a panel discussion.

Media Forensics

Organizers: Shruti Agarwal, Cristian Canton, Hany Farid, Luisa Verdoliva

Date: Tuesday, June 18

Time: 8:45 AM–5:00 PM

Location: Arch 2B

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). 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. This workshop aims to bring together a heterogeneous group of specialists from academia and industry to discuss emerging threats, technologies, and mitigation strategies.

ScanNet++ Novel View Synthesis and 3D Semantic Understanding Challenge

Organizers: Angela Dai, Yueh-Cheng Liu, Chandan Yeshwanth, Ben Mildenhall, Peter Kontschieder, Matthias Niessner

Date: Tuesday, June 18

Time: 8:50 AM–12:30 PM

Location: Arch 211

Summary: Recent advances in generative modeling and semantic understanding have spurred significant interest in synthesis and understanding of 3D scenes. In 3D, there is significant potential in application areas, for instance augmented and virtual reality, computational photography, interior design, and autonomous mobile robots all require a deep understanding of 3D scene spaces. We propose to offer the first benchmark challenge for novel view synthesis in large-scale 3D scenes, along with high-fidelity, large-vocabulary 3D semantic scene understanding -- where very complete, high-fidelity ground truth scene data is available. This is enabled through the new ScanNet++ dataset, which offers 1mm resolution laser scan geometry, high-quality DSLR image capture, and dense semantic annotations over 1000 class categories. In particular, existing view synthesis leverages data captured from a single continuous trajectory, where evaluation of novel views outside of the original trajectory capture is impossible. In contrast, our novel view synthesis challenge leverages test images captured intentionally outside of the train image trajectory, allowing for comprehensive evaluation of methods to test new, challenging scenarios for state-of-the-art methods.

Towards 3D Foundation Models: Progress and Prospects

Organizers: Sai Bi, Angela Dai, Leo Guibas, Hao Su, Kalyan Sunkavalli, Hao Tan, Gordon Wetzstein, Peter Wonka, Yinghao Xu, Zexiang Xu, Kai Zhang

Date: Tuesday, June 18

Time: 8:50 AM–5:00 PM

Location: Summit 434

Summary: Textual and 2D visual foundation models, like GPT, BERT, DINO, and StableDiffusion, have revolutionized the fields of natural language processing and 2D computer vision with broad applications across academia and industry. However, the community still awaits the establishment of a 3D foundation model, one that could lead to diverse downstream applications in areas ranging from 3D content creation to AR/VR, robotics, and autonomous driving. This workshop aims to bring together researchers from various domains of 3D computer vision, facilitating discussions around the development of 3D foundation models. In particular, we investigate the questions including, but not limited to: 1) What types of datasets are necessary for 3D foundation models?, 2) On which 3D tasks should such models be trained?, 3) What are the consensus on the basic architecture and 3D representations for understanding and synthesis tasks? and 4) What applications and capabilities can they unlock?. We invite a diverse group of top experts in the field to share their recent research outcomes and future visions in this regard, with a focus on scalable, generalizable, and adaptable 3D AI frameworks. We hope this workshop will inspire and accelerate the future innovations in 3D foundation models and their broad applications.

Notes:

Annual Embodied AI Workshop

Organizers: Anthony Francis, Claudia Pérez D'Arpino, Luca Weihs, Lamberto Ballan, Yontatan Bisk, Angel Chang, Changan Chen, Matt Deitke, David Hall, Devon Hjelm, Eric Li, Oleksandr Maksymets, Rin Metcalfe, Chris Paxton, Soren Pirk, Ram Ramrakhya, Mike Roberts, Naoki Yokoyama

Date: Tuesday, June 18

Time: 8:50 AM–5:30 PM

Location: Summit 428

Summary: The goal of the Fifth Annual 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. The overarching theme of this year's workshop is "Open World Embodied AI": truly effective embodied AI agents should be able to deal with tasks, objects, and situations markedly different from those that they have been trained on. This umbrella theme is divided into three topics: "Embodied Mobile Manipulation" builds on embodied navigation and manipulation topics from previous years and makes them more challenging, "Generative AI for Embodied AI" is an important tool researchers are using to support embodied artificial intelligence research, and "Language Model Planning" uses large language models (LLMs), vision-language models (VLMs), and multimodal foundation models to turn arbitrary language commands into plans and sequences for action–a key feature needed to make embodied artificial intelligence systems useful for performing the tasks in open worlds.

Multimodal Learning and Applications

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

Date: Tuesday, June 18

Time: 9:00 AM–6:00 PM

Location: Summit 320

Summary: This workshop combines cutting-edge research at the intersection of computer vision, machine learning, multimedia, remote sensing, and robotics, focusing on the challenges and opportunities in multimodal data processing. Recent advancements have highlighted the significant benefits of leveraging multiple data modalities, such as audio, visual, thermal, and textual information, to enhance the performance and robustness of machine learning applications. The 16 accepted papers showcase innovative approaches in multimodal dataset handling, synchronization, annotation, and exploring self-supervised learning methodologies that facilitate learning without manual labeling. Topics range from cross-modal fusion and attention mechanisms for video anomaly detection, leveraging generative language models, 3D object detection, and trajectory prediction for autonomous driving. This workshop aims to foster interdisciplinary interaction and collaboration, highlighting the growing interest from academia and industry in the fusion of information from multiple sensors for applications in automotive, drones, surveillance, robotics, and beyond.

New frontiers for zero-shot Image Captioning Evaluation

Organizers: Taehoon Kim, Pyunghwan Ahn, Gwangmo Song, Emily Webber, Youngjoon Choi, KyoungMu Lee, SeungHwan Kim, Bohyung Han, Larry Davis

Date: Tuesday, June 18

Time: 9:00 AM–5:00 PM

Location: Summit 323

Summary: The purpose of this workshop is 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 (i.e. mitigating societal biases). Both of these issues must be addressed fully before image captioning technology can be reliably deployed in a large-scale setting. The workshop is going to cover various topics related to image captioning, which includes but are not limited to: hallucination of image captioning models, societal bias of vision-language datasets, and controllable image captioning models. It aims to challenge the models to generate reliable captions which explains the image with the expected level of details, while not presenting any wrong information. To accomplish this, the models need to understand the image accurately and thoroughly, and also explain it with the words that accurately correspond to each part of the image. In advance of the workshop, we plan to organize two challenges: one on zero-shot image captioning and the other one on caption re-ranking, both of which will focus on reducing caption hallucination and estimating how much each caption resembles human-generated ones. Various approaches are expected to be presented during the challenge, and through the workshop, we will share the novel and reliable methods of zero-shot image captioning and caption re-ranking, in order to assist the vision-language research community advancing to the next level. Throughout the workshop and challenge, we will cover a broad range of topics on image understanding and text generation, with deeper consideration on common problems of vision-language models, such as caption hallucination and societal biases. Therefore, we plan to invite researchers to provide talks on various topics under the range of combination of language and vision. The list of related topics include: Zero-shot Image Captioning, Caption Hallucination, Caption Evaluation, Multimodal Learning, Fairness.

Deep Learning for Geometric Computing

Organizers: Dena Bazazian, Ilke Demir, Adarsh Krishnamurthy, Géraldine Morin, Kathryn Leonard, Silvia Sellán, Aditya Balu, Sainan Liu

Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

Location: Summit 448

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, compu-

tational geometry, computer graphics, and machine learning in this sixth edition of "Deep Learning for Geometric Computing" workshop at CVPR 2024. The workshop encapsulates competitions with prizes, proceedings, keynotes, paper presentations, and a fair and diverse environment for brainstorming about future research collaborations.

Safe Artificial Intelligence for All Domains

Organizers: Timo Sämann, Oliver Wasenmüller, Thomas Stauner, Markus Enzweiler, Oliver Grau, Peter Schlicht, Joachim Sicking, Stefan Milz, Claus Bahlmann

Date: Tuesday, June 18

Time: 9:00 AM–5:00 PM

Location: Arch 304

Summary: The Workshop on Safe Artificial Intelligence for All Domains (SAIAD) focuses on the critical aspects of AI safety across multiple fields. As AI technologies increasingly integrate into various sectors, ensuring their safety, robustness, and ethical considerations becomes essential. This workshop will delve into the following key areas: Automated driving, robotics, aerospace, ethics and medical applications. The workshop gathers experts to discuss innovative solutions and best practices for developing safe, robust and responsible AI systems. Join us to explore the intersection of AI safety with diverse domains and learn how to make AI safer.

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

Organizers: Iro Armeni, Fuxin Li, Michael Olsen, Yelda Turkan, Erzuo Che, Martin Fischer, Daniel Hall, Jaehoon Jung, Marc Pollefeys, Heidar Rastiveis

Date: Tuesday, June 18

Time: 9:00 AM–5:00 PM

Location: Summit 443

Summary: The workshop connects the domains of Architecture, Engineering, and Construction (AEC) with that of Computer Vision by establishing a common ground of interaction and identifying shared research interests. It focuses on the semantic understanding, analysis, and generation of built environments, with a focus on key sustainability issues such as reuse, resilience, and performance-based design. Through keynote talks, presentations of accepted short papers, posters, discussions, and the workshop's challenge, the topics will be presented to attendees from the dual lens of Computer Vision and AEC, highlighting the requirements, limitations, and bottlenecks related to developing and adopting robust computer vision applications for this domain that actually work in the real-world.

Vision and Language for Autonomous Driving and Robotics

Organizers: Boyi Li, Yue Wang, Hang Zhao, Jiawei Yang, Jiageng Mao, Serge Belongie, Sanja Fidler, Marco Pavone

Date: Tuesday, June 18

Time: 9:00 AM–6:00 PM

Location: Summit 345-346

Summary: The contemporary discourse in technological advancement underscores the increasingly intertwined roles of vision and language processing, especially within the realms of autonomous driving and robotics. The necessity for this symbiosis is apparent when considering the multifaceted dynamics of real-world environments. An autonomous vehicle, for instance, operating within the framework of urban locales should not merely rely on its visual sensors for pedestrian detection, but must also interpret and act upon auditory signals, like vocalized warnings. Similarly, robots that integrate visual data with linguistic context promise more adaptive functionalities, particularly in diverse settings. This workshop is expected to spotlight the intricate arena of data-centric autonomous driving, emphasizing vision-based techniques. Central to our discussions will be topics like vision and language for autonomous driving, language-driven perception, and simulation. We will delve into the nuanced realms of vision and language representation learning and explore the future of multimodal motion prediction and planning in robotics. Recognizing the rapid expansion of this field, the introduction of new datasets and metrics for multimodal learning will also be on our agenda. Equally paramount are the discussions on privacy concerns associated with multimodal data. Moreover, our emphasis will firmly rest on safety, ensuring that systems are adept at correctly interpreting and acting on both visual and linguistic inputs, thereby preventing potential mishaps in real-world scenarios. Through a comprehensive examination of these topics, this workshop seeks to foster a deeper academic understanding of the intersection between vision and language in autonomous systems. By convening experts from interdisciplinary fields, our objective is to decipher current state-of-the-art methodologies, address challenges, and chart avenues for future endeavors, ensuring our findings resonate within both academic and industrial communities.

Learning with Limited Labelled Data for Image and Video Understanding

Organizers: Mennatullah Siam, Issam Laradji, Leonid Sigal, Katerina Fragkiadaki, Xin Wang, Raghav Goyal, Valentina Zantedeschi, Kosta Derpanis, He Zhao, Junshi Xia, Clifford Broni-Bediako, Mai Gamal

Date: Tuesday, June 18

Time: 9:00 AM–6:10 PM

Location: Summit 322

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 third L3D-IVU workshop in conjunction with CVPR, where it had a great success and wide interest from multiple researchers as it explores the intersection of learning with limited labelled data and video understanding. This year, our workshop's theme will be around learning with limited labelled data and AI for social good, where we discuss assistive technologies and remote sensing amongst others.

Social Presence with Codec Avatars

Organizers: Julieta Martinez, Javier Romero, Yaser Sheikh, Michael Zollhoefer

Date: Tuesday, June 18

Time: 9:30 AM–5:30 PM

Location: Summit 333

Summary: A workshop devoted to telepresence: the task of generating and driving realistic human representations. Regarding generation, we plan to cover the learning of efficient 3d representations of faces, hands, and bodies, and the particular challenges of each modality. For driving, we will focus on using headsets to drive faces and hands, as well as external cameras for full-body tracking. The invited speakers will provide context on the state of the art both in industry and academia for efficient 3d representations and their applications to human modelling.

Implicit Neural Representation for Vision

Organizers: Matthew Gwilliam, Sihyun Yu, Subin Kim, Shishira Maiya, Max Ehrlich, Hao Chen, Jinwoo Shin, Abhinav Shrivastava

Date: Tuesday, June 18

Time: 1:00 PM–6:30 PM

Location: Summit 335-336

Summary: An emerging area within deep learning, implicit neural representation (INR), also known as neural fields, offers a powerful new mechanism and paradigm for processing and representing visual data. In contrast with the dominant big data setting, INR focuses on neural networks which parameterize a field, often in a coordinate-based manner. The most well-known of this class of models is NeRF, which has been wildly successful for 3D modeling, especially for novel view synthesis. INR for 2D images and videos have many compelling properties as well, particularly for visual data compression. By treating the weights of the networks as the data itself and leveraging their implicit reconstruction ability, several multimodal compression techniques have been developed. This is a relatively new area in vision, with many opportunities to propose new algorithms, extend existing applications, and innovate entirely new systems. Since working with INRs often requires less resources than many areas, this sort of research is especially accessible in the academic setting. Additionally, while there are many workshops for NeRF, there are often none for the incredibly broad spectrum of other INR work. Therefore, we propose this workshop as an avenue to build up the fledgling INR community, and help disseminate knowledge in the field about this exciting area. We thus invite researchers to the Workshop for Implicit Neural Representation for Vision (INRV) where we investigate multiple directions, challenges, and opportunities related to implicit neural representation.

The Evaluation of Generative Foundation Models

Organizers: Maria Zontak, Xu Zhang, Mehmet Saygin Seyfioglu, Erran Li, Bahar Erar Hood, Suren Kumar, Karim Bouyarmane

Date: Tuesday, June 18

Time: 1:00 PM–6:30 PM

Location: Summit 433

Summary: The landscape of artificial intelligence is being transformed by the advent of Generative Foundation Models (GenFMs), such as Large Language Models (LLMs) and diffusion models. GenFMs offer unprecedented opportunities to enrich human lives and transform industries. However, they also pose significant challenges, including the generation of factually incorrect or biased information, which might be potentially harmful or misleading. With the emergence of multimodal GenFMs, which leverage and generate content in an increasing number of modalities, these challenges are set to become even more complex. This emphasizes the urgent need for rigorous and effective evaluation methodologies. The 1st Workshop on the Evaluation of Generative Foundation Models at CVPR 2024 aims to build a forum to discuss ongoing efforts in industry and academia, share best practices, and engage the community in working towards more reliable and scalable approaches for GenFMs evaluation. Our speakers and panelists include leading researchers from industry and academia: Ece Kamar (Managing Director at Microsoft Research), Tal Hassner (Applied Research Lead at Facebook AI), Bo Li (Associate Professor at University of Chicago), Ranjay Krishna (Associate Professor at University of Washington), Hanwang Zhang (Associate Professor at Nanyang Technological University), Leonid Karlinsky (Principal Research Scientist in

the MIT-IBM lab), Jungo Kasai (Co-founder & CTO at Kotoba Technologies, Inc.), Sadeep Jayasumana (Staff Research Scientist at Google Research), Besmira Nushi (Researcher at Microsoft Research). The workshop will also include a poster session featuring some of the most recent research papers on the Evaluation of GenFMs.

OpenSUN3D: Open-Vocabulary 3D Scene Understanding

Organizers: Francis Engelmann, Ayca Takmaz, Jonas Schult, Elisabetta Fedele, Alex Delitzas, Johanna Wald, Songyou Peng, Xi Wang, Or Litany, Despoina Paschalidou, Federico Tombari, Marc Pollefeys

Date: Tuesday, June 18

Time: 1:30 PM–5:30 PM

Location: Arch 211

Summary: The ability to perceive, understand and interact with arbitrary 3D environments is a long-standing goal in both academia and industry with applications in AR/VR as well as robotics. Current 3D scene understanding models are largely limited to recognizing a closed set of pre-defined object classes. Recently, large visual-language models, such as CLIP, have demonstrated impressive capabilities trained solely on internet-scale image-language pairs. Some initial works have shown that these models have the potential to extend 3D scene understanding not only to open set recognition, but also offer additional applications such as affordances, materials, activities, and properties of unseen environments. The goal of this workshop is to bundle these initial siloed efforts and to discuss and establish clear task definitions, evaluation metrics, and benchmark datasets.

Monocular Depth Estimation Challenge

Organizers: Ripudaman Singh Arora, Jaime Spencer, Fabio Tosi, Matteo Poggi, Chris Russell, Simon Hadfield, Richard Bowden

Date: Tuesday, June 18

Time: 1:30 PM–5:30 PM

Location: Summit 331

Summary: Monocular depth estimation (MDE) is an important low-level vision task, with application in fields such as augmented reality, robotics and autonomous vehicles. Recently, there has been an increased interest in self-supervised systems capable of predicting the 3D scene structure without requiring ground-truth LiDAR training data. Automotive data has accelerated the development of these systems, thanks to the vast quantities of data, the ubiquity of stereo camera rigs and the mostly-static world. However, the evaluation process has also remained focused on only the automotive domain and has been largely unchanged since its inception, relying on simple metrics and sparse LiDAR data. This workshop seeks to answer the following questions: How well do networks generalize beyond their training distribution relative to humans? What metrics provide the most insight into the model's performance? What is the relative weight of simple cues, e.g. height in the image, in networks and humans? How do the predictions made by the models differ from how humans perceive depth? Are the failure modes the same? The workshop will therefore consist of two parts: invited keynote talks discussing current developments in MDE and a challenge organized around a novel benchmarking procedure using the SYNS dataset.

AVA: Accessibility, Vision and Autonomy Meet

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

Date: Tuesday, June 18

Time: 1:30 PM–5:30 PM

Location: Summit 435

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 largescale 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.

Representation Learning with Very Limited Images: Zero-shot, Unsupervised, and Synthetic Learning in the Era of Big Models

Organizers: Hirokatsu Kataoka, Yuki M. Asano, Christian Rupprecht, Rio Yokota, Nakamasa Inoue, Dan Hendrycks, Xavier Boix, Manel Baradad, Connor Anderson, Ryo Nakamura, Ryosuke Yamada, Risa Shinoda, Ryu Tadokoro, Erika Mori

Date: Tuesday, June 18

Time: 1:30 PM–5:30 PM

Location: Summit 324

Summary: At this very moment, the era of 'foundation models' heavily relies on a huge amount (>100M-order data) of samples inside of a training dataset. We have witnessed that this kind of large-scale dataset tends to incur ethical issues such as societal bias, copyright, and privacy, due to the uncontrollable big data. On the other hand, the setting of very limited data such as self-supervised learning with a single image or synthetic pre-training with generated images are free of the typical issues. Efforts to train visual/multi-modal models on very limited data resources have emerged independently from various academic and industry communities around the world. This workshop aims to bring together these various communities to form a collaborative effort and find brave new ideas.

Perception Beyond the Visible Spectrum

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

Date: Tuesday, June 18

Time: 1:30 PM–5:30 PM

Location: Arch 201

Summary: The Perception Beyond the Visible Spectrum workshop series (IEEE PBVS), established over 2 decades ago in 2004, has consistently been a key event within the computer vision and pattern recognition (CVPR) community. This prestigious series has continually highlighted cutting-edge sensing technologies, and exploitation algorithms operating in the non-visible spectrum, including infrared, thermal, SAR, radio frequency, hyper-spectral, and x-ray. In 2024, we're hosting 3 challenges on thermal, EO/IR and SAR data fusion and recognition.

Precognition: Seeing through the Future

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

Date: Tuesday, June 18

Time: 1:30 PM–5:30 PM

Location: Summit Elliott Bay

Summary: Despite its potential and relevance for real-world applications, visual forecasting or precognition has not been in the focus of new theoretical studies and practical applications as much as detection and recognition problems. Through the organization of this workshop we aim to facilitate further discussion and interest within the research community regarding this nascent topic. The workshop 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, health-care, human face aging prediction, early event prediction, autonomous driving forecasting, and so on. In addition, this workshop will give an opportunity for the community in both academia and industry to meet and discuss future work and research directions. It will bring together researchers from different fields and viewpoints to discuss existing major research problems and identify opportunities in further research directions in both research topics and industrial applications.

Embedded Vision

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

Date: Tuesday, June 18

Time: 1:30 PM–5:30 PM

Location: Arch 205

Summary: 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.

Robot Visual Perception in Human Crowded Environments

Organizers: Hamid Rezatofighi, Alexandre Alahi, Ian Reid, Duy Tho Le, Hengcan Shi, Chenhui Gou

Date: Tuesday, June 18

Time: 1:30 PM–5:30 PM

Location: Arch 210

Summary: In the recent past, the computer vision and robotics communities have proposed several centralized benchmarks to evaluate and compare different machine visual perception solutions. With the rise of the popularity of 3D sensory data systems based on LiDAR, some benchmarks have begun to provide both 2D and 3D sensor data and to define new scene understanding tasks on this geometric information. Nonetheless, their targeted domain application is autonomous driving. In this workshop, similar to our previous workshops, we target a unique visual domain tailored to the perceptual tasks related to robot visual perception in human environments, both indoors and outdoors. In previous workshops in the JRDB series (ICCV 2019, CVPR 2021, ECCV 2022 and ICCV 2023) , we aim at 2D-3D human detection, tracking and forecasting, body skeleton pose estimation, human social grouping and activity recognition. In the CVPR 2024 workshop, we will broaden the horizon to human-centered environment analysis, particularly in 2D-3D scene panoptic segmentation, multiple object tracking and open-word recognition. New annotated data and challenges will also be provided. We believe these novel data, application domains and high-quality annotations will attract large attention from CVPR and other relevant communities, especially for 3D, segmentation, tracking, machine learning and robotics fields.

Neural Volumetric Video

Organizers: Sida Peng, Yiyi Liao, Xiaowei Zhou, Andreas Geiger

Date: Tuesday, June 18

Time: 1:30 PM–5:50 PM

Location: Summit 332

Summary: This workshop aims to bring researchers from the field of dynamic view synthesis together to discuss the latest progress and important technical challenges of this task. More concretely, we have the following question to talk about: 1) Assuming that the input observations are sufficient, e.g., videos recorded by the dense camera array, what are the right 4D scene representations that can faithfully capture complex dynamic scenes while being storage-efficient for easy transmission. 2) For the immersive telepresence, will the image-based blending technique be the ultimate solution? What are the limitations of this technique? 3) How to expand the view synthesis range when the input is monocular video? 4) What is the product form of the camera for common users to produce their own volumetric videos of daily life? More broadly, given the rapid advancements in AR/VR glasses, we hope that this workshop will facilitate discussions on how dynamic view synthesis techniques can be optimized to enhance content capture and display on these devices.

Competition on Affective Behavior Analysis in-the-wild

Organizers: Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Panagiotis Tzirakis, Alan Cowen

Date: Tuesday, June 18

Time: 1:30 PM–6:00 PM

Location: Arch 212

Summary: The ABAW Workshop and Competition has a unique aspect of fostering cross-pollination of different disciplines, bringing together experts (from academia, industry, and government) and researchers of mobile and ubiquitous computing, computer vision and pattern recognition, artificial intelligence and machine learning, multimedia, robotics, HCI, ambient intelligence 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 events provide a timely and relevant discussion and dissemination platform. The ABAW Workshop and Competition is a continuation of the respective Workshops and Competitions held at IEEE CVPR 2023 & 2022, ECCV 2022, ICCV 2021, IEEE FG 2020 and IEEE CVPR 2017 Conferences.

EgoMotion: Egocentric Body Motion Tracking, Synthesis and Action Recognition

Organizers: Lingni Ma, Richard Newcombe, Yuting Ye, Ziwei Liu, Yifeng Jiang, Vladimir Guzov, Edward Miller

Date: Tuesday, June 18

Time: 1:30 PM–6:00 PM

Location: Summit 429

Summary: The EgoMotion workshop focuses on in-the-wild egocentric full-body motion understanding from data captured by wearable devices. We address three major problems for full-body motion – tracking, synthesis and action recognition. The scope combines several topics and sub-fields that have grown in importance over the last years, and aims to bring these fields together in a common forum. In addition to algorithms, we will promote multiple recent datasets and their associated challenges to accelerate research in this field.

New Trends in Multimodal Human Action Perception, Understanding and Generation

Organizers: Yansong Tang, Siyuan Huang, Song Bai, Baoxiong Jia, Yufei Ye, Xudong Lin, Song-Chun Zhu, Philip H.S. Torr

Date: Tuesday, June 18

Time: 1:30 PM–6:00 PM

Location: Summit 430

Summary: Over the last decade, tremendous interests have been attracted to this field and great success has been achieved for various video-centric tasks (e.g., action recognition, detection and segmentation) based on conventional RGB videos. In recent years, with the explosion of videos and various application demands (e.g., video editing, AR/VR, human-robot interaction, etc.), significantly more efforts are required to enable an intelligent system to perceive, understand and generate human action under different scenarios within multimodal inputs. Moreover, with the development of recent large language models (LLMs)/large multimodal models (LMMs), there are growing new trends and challenges to be discussed and addressed. The goal of this workshop is to foster interdisciplinary communication of researchers so that more attention of the broader community can be drawn to this field. Through this workshop, current progress and future directions will be discussed, and new ideas and discoveries in related fields are expected to emerge.

Learning from Procedural Videos and Language: What is Next?

Organizers: Effrosyni Mavroudi, Triantafyllos Afouras, Yale Song, Tengda Han, Anna Kukleva, Ehsan Elhamifar, Kristen Grauman, Lorenzo Torresani

Date: Tuesday, June 18

Time: 1:30 PM–6:00 PM

Location: Summit 427

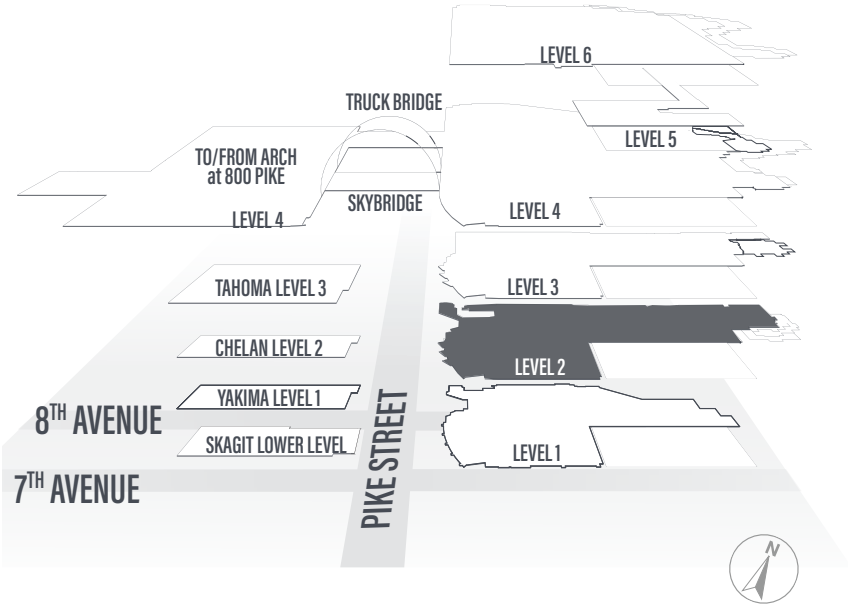
Summary: Procedural videos that illustrate multi-step activities have gained popularity as powerful tools for skill acquisition, covering a wide range of domains such as cooking, home improvement, and crafting. In addition to being effective educational materials for humans, procedural videos paired with language are also a promising medium for multimodal machine learning, as they combine visual demonstrations with detailed verbal descriptions. The recent introduction of relevant datasets (e.g., HowTo100M, HT-Step and Ego4D Goal-Step) has sparked a growing interest in procedural video understanding, to support applications in augmented reality, robotics, and more. This workshop aims to foster discussion on the future of language-based procedural video understanding. We'll explore paths to integrating diverse language sources, leveraging Large Language Models (LLMs) for structured task knowledge, and merging language with other information streams (visual, audio, IMU, etc.) to enhance procedural video understanding. This includes recognizing key steps, identifying mistakes, understanding hand-object interactions, and more.

Notes:

Notes:

The Arch building is at 705 Pike Street.

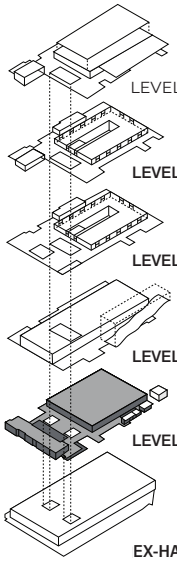
The main pedestrian entrance to Arch is on the corner of 7th Avenue and Pike Street, and the Arch drop-off points are 725 Pike (private/rideshare) and 800 Convention Place (bus).



The Summit building is at 900 Pine Street, just over a block away from Arch.

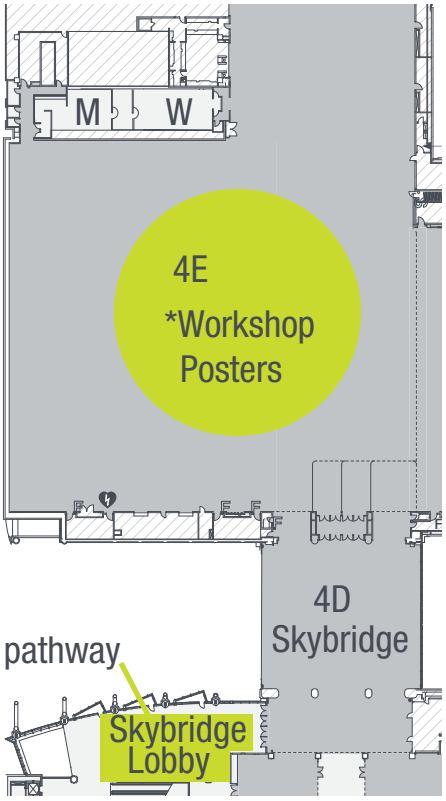
The main pedestrian entrance is on the corner of 9th Avenue and Pine Street, and the drop-off location is on 9th Avenue between Pine Street and Olive Way (private/rideshare and bus).

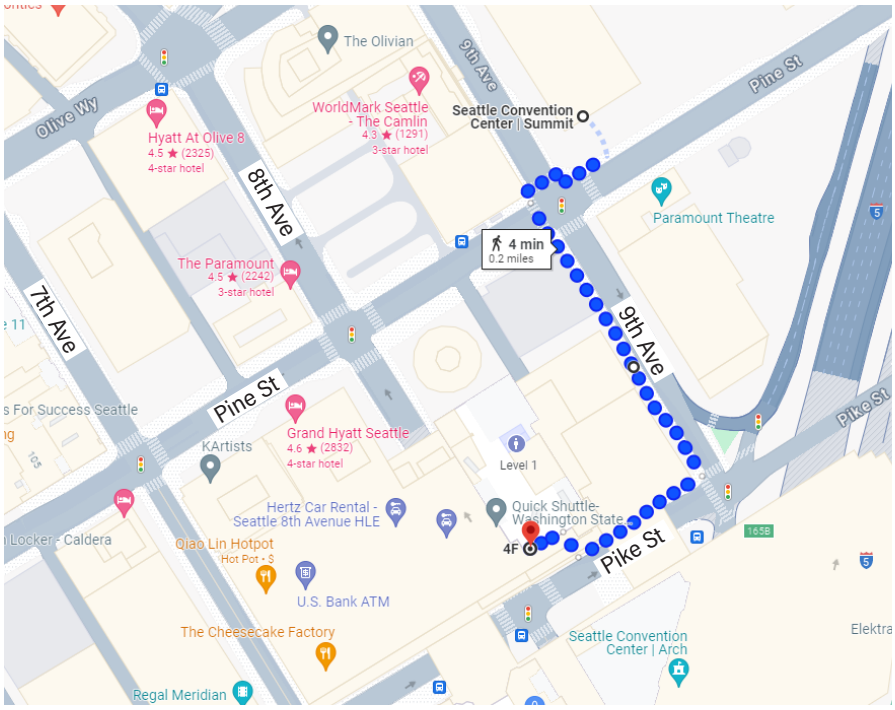
SUMMIT FLOOR LEVELS



** Elliott Bay & Terrace Suite on level 4

* FLEX HALL on level 2






SEATTLE, WA
CVPR
JUNE 17-21, 2024