ICINCO 2026

23rd International Conference on Informatics in Control, Automation and Robotics

PROCEEDINGS

Volume 1

Angers, France

26 - 28 October, 2026

EDITORS

Giuseppina Carla Gini
Antonella Ferrara
Dimitar Filev

https://icinco.scitevents.org

SPONSORED BY

INSTICC

PAPERS AVAILABLE AT

SCITEPRESS DIGITAL LIBRARY

ICINCO 2026

Proceedings of the 23rd International Conference on Informatics in Control, Automation and Robotics

Volume 1

Angers - France

October 26 - 28, 2026

Sponsored by
INSTICC - Institute for Systems and Technologies of Information, Control and Communication

Locally Organized and Hosted by
ESEO

IEEE Technically co-sponsored by
IEEE SMC - TC on Evolving Intelligent Systems

Technically Co-sponsored by
IFAC - International Federation of Automatic Control

ACM In Cooperation
SIGAI - ACM Special Interest Group on Artificial Intelligence

In Cooperation with
AAAI - Association for the Advancement of Artificial Intelligence
INNS - International Neural Network Society

Copyright © 2026 by SCITEPRESS – Science and Technology Publications, Lda.

Edited by Giuseppina Gini, Antonella Ferrara and Dimitar Filev

Printed in Portugal
ISSN: 2184-2809
ISBN: 978-989-758-870-9
DOI: 10.5220/0000224600004979
Depósito Legal: 570354/26

https://icinco.scitevents.org
icinco.secretariat@insticc.org

BRIEF CONTENTS

INVITED SPEAKERS IV

SPECIAL SESSIONS CHAIRS IV

ORGANIZING COMMITTEES V

PROGRAM COMMITTEE VI

AUXILIARY REVIEWERS IX

SPECIAL SESSIONS PROGRAM COMMITTEE IX

SELECTED PAPERS BOOK X

FOREWORD XI

CONTENTS XIII

INVITED SPEAKERS

Johan Suykens
KU Leuven
Belgium

Bogdan Epureanu
University of Michigan
USA

Maria Domenica Di Benedetto
University of L'Aquila
Italy

SPECIAL SESSIONS CHAIRS

SPECIAL SESSION ON BRIDGING THE GAP IN COLLABORATIVE ROBOTICS: FROM THEORY TO REAL APPLICATIONS

Juan Corrales Ramón, Universidade de Santiago de Compostela, Spain

Chedli Bouzgarrou, Sigma-Clermont, France

Pedro Gaspar, University of Beira Interior, Faculty of Engineering, Portugal

Daniel Sánchez, Technological Institute for Children's Products & Leisure (AIJU), Spain

SPECIAL SESSION ON HYBRID INTELLIGENCE: INTEGRATING PHYSICS-BASED PRINCIPLES WITH DATA-DRIVEN AI

Vittorio Lippi, Calejo Hybrid Intelligence AB, Germany

Ndivhuwo Makondo, IBM Research, University of the Witwatersrand, South Africa

Leonard Johard, Calejo Hybrid Intelligence AB, Sweden

ORGANIZING COMMITTEES

CONFERENCE CHAIR

Dimitar Filev, Ford Research / Texas A&M University, United States

PROGRAM CO-CHAIRS

Giuseppina Gini, Politecnico di Milano, Italy
Antonella Ferrara, University of Pavia, Italy

SECRETARIAT

Ana Rita Paciência, INSTICC, Portugal

GRAPHICS PRODUCTION AND WEBDESIGNER

Inês Teles, INSTICC, Portugal

WEBMASTER

João Francisco, INSTICC, Portugal
Carolina Ribeiro, INSTICC, Portugal

PROGRAM COMMITTEE

Sherif Abdelwahed, Virginia Commonwealth University, United States

Hussein Abdullah, University of Guelph, Canada

Khalid Abidi, Newcastle University Singapore, Singapore

Rahib Abiyev, Near East University, Türkiye

El-Houssaine Aghezzaf, Ghent University, Faculty of Engineering and Architecture, Belgium

Eugenio Aguirre, University of Granada, Spain

Rudy Agustriyanto, University of Surabaya, Indonesia

Mojtaba Ahmadieh Khanesar, University of Nottingham, United Kingdom

Manuel Aleixandre, Tokyo Institute of Technology, Japan

Yannick Aoustin, Laboratoire des Sciences du Numérique de Nantes, France

Rui Araujo, University of Coimbra, Portugal

Laurent Arcese, University of Reims Champagne-Ardenne, France

Filippo Arrichiello, University of Cassino and Southern Lazio, Italy

Gustavo Arroyo-Figueroa, Instituto Nacional de Electricidad y Energías Limpias, Mexico

Mohd Ashraf Ahmad, University Malaysia Pahang, Malaysia

T. Asokan, Indian Institute of Technology Madras, India

Wudhichai Assawinchaichote, King Mongkut's University of Technology Thonburi, Thailand

Ralf Bachmayer, University of Bremen, Germany

Ramiro Barbosa, Institute of Engineering – Polytechnic of Porto (ISEP/IPP), Portugal

Michele Basso, University of Florence, Italy

Pedro Batista, University of Lisbon, Portugal

Yazdan Batmani, University of Kurdistan, Iran

Mahmoud Belhocine, CDTA, Algeria

Juri Belikov, Tallinn University of Technology, Estonia

Lyes Benyoucef, Aix-Marseille University, France

Karsten Berns, University of Kaiserslautern-Landau, Germany

Zafer Bingul, Kocaeli University, Türkiye

Claudia-Adina Bojan-Dragos, Politehnica University of Timisoara, Romania

Mohammad Bozorg, Yazd University, Iran

Jean-François Brethe, GREAH, France

Tobias Bruckmann, University of Duisburg-Essen, Germany

Ivan Buzurovic, Harvard University, United States

Marco Castellani, University of Birmingham, United Kingdom

Ratchatin Chanchareon, Chulalongkorn University, Thailand

Amitava Chatterjee, Jadavpur University, India

Paul Christodoulides, Cyprus University of Technology, Cyprus

Paulo Jorge Coelho, Polytechnic University of Leiria, Portugal

Michel Dambrine, Polytechnic University of Hauts-de-France, France

Radu-Codrut David, Politehnica University of Timisoara, Romania

Mingcong Deng, Tokyo University of Agriculture and Technology, Japan

Pawel Drag, Wroclaw University of Science and Technology, Poland

Richard J. Duro, Universidade Da Coruña, Spain

Marc Ebner, Ernst-Moritz-Arndt-Universität Greifswald, Germany

Simon Fabri, University of Malta, Malta

Mohammed Fadali, UNR, United States

Baris Fidan, University of Waterloo, Canada

Alberto Finzi, University of Naples Federico II, Italy

Paolo Fiorini, University of Verona, Italy

Thierry Fraichard, INRIA, France

Giuseppe Franzè, University of Calabria, Italy

Eduardo Oliveira Freire, Federal University of Sergipe, Brazil

Iván García Daza, University of Alcalá, Spain

Péter Gáspár, HUN-REN Institute for Computer Science and Control (SZTAKI), Hungary

Lech Grzesiak, Warsaw University of Technology, Poland

Maki Habib, The American University in Cairo, Egypt

Kensuke Harada, Osaka University, Japan

Khalifa Harib, UAE University, United Arab Emirates

Sallehuddin Haris, Universiti Kebangsaan Malaysia, Malaysia

Ramdane Hedjar, King Saud University, Saudi Arabia

Randa Herzallah, University of Warwick, United Kingdom

Gianluca Ippoliti, Università Politecnica delle Marche, Italy

Sarangapani Jagannathan, Missouri University of Science and Technology, United States

Aleš Janota, University of Zilina, Slovakia

Michael Jenkin, York University, Canada

Isabel Jesus, ISEP/IPP - School of Engineering, Polytechnic Institute of Porto, Portugal

Fabrício Junqueira, University of São Paulo (USP), Brazil

Mete Kalyoncu, Department of Mechanical Engineering, Türkiye

Tohru Kawabe, University of Tsukuba, Japan

Mehrdad Kermani, University Western Ontario, Canada

Azfar Khalid, Nottingham Trent University, United Kingdom

Balint Kiss, Budapest University of Technology and Economics, Hungary

Bahare Kiumarsi, Michigan State University, United States

Alexandr Klimchik, University of Lincoln, United Kingdom

Moharram Korayem, Robotic Research Lab, Mechanical Engineering Department, Iran University of Science and Technology, Iran

Peter Košťál, Slovenská Technická Univerzita v Bratislave, Slovakia

Mianowski Krzysztof, Politechnika Warszawska, Poland

Kolja Kühnlenz, Coburg University of Applied Sciences and Arts, Germany

Miroslav Kulich, Czech Technical University in Prague, Czechia

Reza Langari, Texas A&M University, United States

Fabian Andres Lara-Molina, Federal University of Triângulo Mineiro, Brazil

Nuno Lau, Universidade de Aveiro, Portugal

Egons Lavendelis, Riga Technical University, Latvia

Sukhan Lee, Sungkyunkwan University, Republic of Korea

Fátima Leite, University of Coimbra, Portugal

Kauko Leiviskä, Independent Researcher, Finland

Gordon Lightbody, University College Cork, Ireland

Anthony Maciejewski, Colorado State University, United States

Mohammad Javad Mahmoodabadi, Sirjan University of Technology, Iran

André Maitelli, Federal University of Rio Grande Do Norte, Brazil

Ezio Malis, Inria, France

Federico Mari, University of Rome Foro Italico, Italy

Philippe Martinet, INRIA, France

Ester Martinez-Martin, University of Alicante, Spain

Pavlo Maruschak, Ternopil National Ivan Puluj Technical University, Ukraine

Julien Marzat, DTIS, ONERA, Université Paris Saclay, France

Luca Mazzola, HSLU - Lucerne University of Applied Sciences, Switzerland

Seán McLoone, Queen's Unviersity Belfast, Ireland

Héctor Montes, Universidad Tecnologica de Panama, Panama

Rafael Morales, University of Castilla La Mancha, Spain

Fabio Morbidi, University of Picardie Jules Verne, France

Javier Moreno-Valenzuela, Instituto Politecnico Nacional-CITEDI, Mexico

Olfa Mosbahi, University of Carthage, Tunisia

Juan J. Nieto, University of Santiago de Compostela, Spain

Urbano Nunes, University of Coimbra, Portugal

Frank Ortmeier, Otto-von-Guericke University of Magdeburg, Germany

Fernando Osorio, University of São Paulo - USP, Brazil

Stamatios Papadakis, University of Crete, Greece

Ju H. Park, Yeungnam University, Republic of Korea

Igor Paromtchik, Robotic Technologies, France

Tejaswin Parthasarathy, NVIDIA Corporation, United States

Anatol Pashkevich, IMT Atlantique Nantes, France

Krzysztof Patan, University of Zielona Gora, Poland

Luis Payá, Miguel Hernandez University, Spain

Dariusz Pazderski, Poznan University of Technology, Poland

Adrian Peidro, Miguel Hernandez University, Spain

Ignacio Peñarrocha Alos, Universitat Jaume I, Spain

Qingjin Peng, University of Manitoba, Canada

Milica Petrovic, University of Belgrade, Serbia

Alexandre Philippot, University of Reims Champagen Ardenne, France

Raul Marin Prades, Jaume I University, Spain

Oscar Reinoso, Miguel Hernandez University, Spain

Bernard Riera, University of Reims Champagne-Ardenne, France

Gerasimos Rigatos, Industrial Systems Institute, Greece

Paolo Rocco, Politecnico di Milano, Italy

Olympia Roeva, Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, Bulgaria

Raul-Cristian Roman, Politehnica University of Timisoara, Romania

Julio Romero Pérez, Universitat Jaume I, Spain

Juha Röning, University of Oulu, Finland

Christophe Sabourin, Univ. Paris Est Creteil, LISSI, France

Antonio Sala, Universitat Politecnica de Valencia, Spain

Addisson Salazar, Universitat Politècnica de València, Spain

Jorge Sales, Jaume I University, Spain

Javier Sanchis, Universitat Politècnica de València, Spain

Guillaume Sandou, CentraleSupelec, France

Marcello Sanguineti, University of Genoa, Italy

Pedro Santana, ISCTE - University Institute of Lisbon, Portugal

Dieter Schramm, University of Duisburg-Essen, Germany

Eurico Seabra, University of Minho, Portugal

Amir Shafie, International Islamic University Malaysia, Malaysia

Michael Short, Teesside University, United Kingdom

Vasile Sima, Technical Sciences Academy of Romania, Romania

Armando Sousa, INESC TEC and FEUP, Portugal

J. Stigter, Wageningen University, Netherlands

Adrian-Mihail Stoica, Polytechnic University Bucharest, Romania

Mu-Chun Su, National Central University, Taiwan, Republic of China

Pavel Surynek, Czech Technical University in Prague, Czechia

József Tar, Óbuda University, Hungary

Tomasz Tarczewski, Nicolaus Copernicus University, Poland

Jerzy Tchórzewski, University of Siedlce, Poland

Hakan Temeltaş, Istanbul Technical University, Türkiye

Daniel Thalmann, Ecole Polytechnique Federale de Lausanne, Switzerland

Enrico Tronci, Universita' Di Roma "la Sapienza", Italy

José Valente de Oliveira, Universidade do Algarve, Portugal

Angel Valera, Universidad Politécnica de Valencia, Spain

Goran Vasiljević, University of Zagreb Faculty of Electrical Engineering and Computing, Croatia

Yolanda Vidal, Universitat Politecnica de Catalunya, Spain

Constantin Volosencu, Politehnica University Timisoara, Romania

Damir Vrancic, Jožef Stefan Institute, Slovenia

James Whidborne, Cranfield University, United Kingdom

Qinmin Yang, Zhejiang University, China

Zennir Youcef, Skikda University, Algeria, Algeria

Erkan Zergeroglu, Gebze Technical University, Türkiye

Jie Zhang, Newcastle University, United Kingdom

Cezary Zielinski, Warsaw University of Technology, Poland

AUXILIARY REVIEWERS

Carlos Aguilar-Avelar, Facultad de Ingeniería, Universidad Autónoma de Baja California, Mexicali, Mexico

Ahmed AlShakhs, University of Waterloo, Canada

Luís Garrote, ISR-UC, Portugal

Salma Ghafouri Varzaneh, Electrical and Computer Engineering, Virginia Commonwealth University, United States

Raúl Robles, Instituto Politécnico Nacional-CITEDI, Mexico

Rafal Sobanski, Poznan University of Technology, Poland

SPECIAL SESSIONS PROGRAM COMMITTEE

SPECIAL SESSION ON BRIDGING THE GAP IN COLLABORATIVE ROBOTICS: FROM THEORY TO REAL APPLICATIONS

Hernan Abaunza Gonzalez, Tecnologico de Monterrey, Mexico

Chedli Bouzgarrou, Sigma-Clermont, France

Frédéric Chapelle, Clermont Auvergne INP - SIGMA Clermont, France

Juan Corrales Ramón, Universidade de Santiago de Compostela, Spain

Mélodie Daniel, Université de Bordeaux, France

Carlos Jara, University of Alicante, Spain

Carlos Mateo Agulló, Université Bourgogne Europe, France

Artemiy Oleinikov, Universidade de Santiago de Compostela, Spain

Daniel Sánchez, Technological Institute for Children's Products & Leisure (AIJU), Spain

Saltanat Seitzhan, University of Santiago de Compostela, Spain

SPECIAL SESSION ON HYBRID INTELLIGENCE: INTEGRATING PHYSICS-BASED PRINCIPLES WITH DATA-DRIVEN AI

Elena Gramellini, Univ. of Manchester, United Kingdom

Leonard Johard, Calejo Hybrid Intelligence AB, Sweden

Vittorio Lippi, Calejo Hybrid Intelligence AB, Germany

Ines Schweigert, Jonas & Redmann, Germany

Riccardo di Sipio, Dayforce, ex-CERN, Canada

Claudio Zito, Heriot-Watt University, United Arab Emirates

SELECTED PAPERS BOOK

A number of selected papers presented at ICINCO 2026 will be published by Springer in a LNEE Series book. This selection will be done by the Conference Chair and Program Co-chairs, among the papers actually presented at the conference, based on a rigorous review by the ICINCO 2026 Program Committee members.

FOREWORD

This book contains the proceedings of the 23rd International Conference on Informatics in Control, Automation and Robotics. This year, ICINCO is held in collaboration with the ESEO, which hosts this event in Angers, France, on October 26-28, 2026. It was sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC) and technically co-sponsored by the IEEE SMC - TC on Evolving Intelligent Systems and the International Federation of Automatic Control. ICINCO 2026 was also organized in cooperation with the ACM Special Interest Group on Artificial Intelligence, the Association for the Advancement of Artificial Intelligence and the International Neural Network Society.

The purpose of the International Conference on Informatics in Control, Automation and Robotics (ICINCO) is to bring together researchers, engineers and practitioners interested in the application of informatics to Control, Automation and Robotics, covering topics related to Intelligent Control Systems, Optimization, Robotics, Automation, Signal Processing, Sensors, Systems Modelling and Control, and Industrial Informatics.

ICINCO 2026 received 154 paper submissions from 40 countries, of which 25% were accepted and published as full papers. A double-blind paper review was performed for each submission by at least 2, but usually 3 or more, members of the International Program Committee, which is composed of established researchers and domain experts.

The high quality of the ICINCO 2026 program is enhanced by the keynote lectures delivered by distinguished speakers who are renowned experts in their fields: Johan Suykens (KU Leuven, Belgium), Maria Domenica Di Benedetto (University of L'Aquila, Italy) and Bogdan Epureanu (University of Michigan, United States).

The conference is complemented by a Special Session on Hybrid Intelligence: Integrating Physics-Based Principles with Data-Driven AI, chaired by Ndivhuwo Makondo, Vittorio Lippi and Leonard Johard, and a Special Session on Bridging the Gap in COllaborative roBOtics: from Theory to real Applications, chaired by Pedro Dinis Gaspar, Chedli Bouzgarrou, Daniel Sánchez and Juan Antonio Corrales Ramón. Furthermore, a Tutorial on Designing an Industrial Robot within the ROS2 Environment, Forward Kinematics, Inverse Kinematics, and Motion Planning was offered.

All presented papers will be available at the SCITEPRESS Digital Library and will be submitted for evaluation for indexing by SCOPUS, Google Scholar, DBLP Computer Science Bibliography, Semantic Scholar, Engineering Index and Web of Science / Conference Proceedings Citation Index.

Several awards, based on the combined marks of paper reviewing, as assessed by the Program Committee, and the quality of the presentation, as assessed by session chairs at the conference venue, are conferred at the conference's closing session as recognition for the best contributions.

Authors of selected papers will be invited to submit extended versions for inclusion in a forthcoming book of ICINCO Selected Papers to be published by Springer, as part of the LNEE Series. Some papers will also be selected for publication of extended and revised versions in the special issue of the Springer Nature Computer Science Journal.

The program for this conference required the dedicated effort of many people. Firstly, we must thank the authors, whose research efforts are reported here. Next, we thank the members of the Program Committee and the auxiliary reviewers for their diligent and professional reviewing. We would also like to deeply thank the invited speakers for their invaluable contribution and for taking the time to prepare their talks. Finally, a word of appreciation for the hard work of the INSTICC team; organizing a conference of this level is a task that can only be achieved by the collaborative effort of a dedicated and highly competent team.

We wish you all an exciting and inspiring conference. We hope to have contributed to the development of our research community, and we look forward to having additional research results presented at the next edition of ICINCO, details of which are available at https://icinco.scitevents.org.

Giuseppina Gini
Politecnico di Milano, Italy

Antonella Ferrara
University of Pavia, Italy

Dimitar Filev
Ford Research / Texas A&M University, United States

CONTENTS

INVITED SPEAKERS

KEYNOTE SPEAKERS

From Shallow to Deep Kernel Machines
Johan Suykens 5

Human Autonomy Teaming Strategies for Physical Orchestrated Intelligence
Bogdan Epureanu 7

Observability and Diagnosability of Hybrid Dynamical Systems: Theory and Applications
Maria Domenica Di Benedetto 9

INTELLIGENT CONTROL SYSTEMS AND OPTIMIZATION

FULL PAPERS

Learning Changes the World and Then What Was Learned Has Changed: Training and Deploying Neural Networks in the Control Loop
Vittorio Lippi 15

Planning Mobile Project Resources as Integrated Systems: A Research Roadmap for the Assignment, Scheduling, and Routing Problem
Ayman R. Mohammed, Majed Hadid and Roberto Baldacci 27

Adaptive Stochastic Model Predictive Control for Grid-Connected Hybrid Energy Storage Management
Taranjitsingh Singh, Boulaid Boulkroune, Maxime Monsieur and Chris Ganseman 39

An Explainable Neuro-Fuzzy Adaptive Graph KAN–TCN Framework for Short-Term Urban Parking Occupancy Forecasting
Mohammad Reza Mohebbi, Farzaneh Keighobadi and Mario Döller 51

An Evaluation of Deep Learning Models for DBH Estimation Using UAV-LiDAR Data in Eucalyptus Inventory
Felipe M. Salles, Adimara B. Colturato, Cristiano R. Reis, Luis Pelin, Marcos G. Quiles and Kalinka Castelo Branco 63

A Symmetry-Reduced Optimal Control Formulation for Quadrotor Dynamics Based on Noether's Theorem
Tanzeela Tasneem, Enzo Avelino, George Kamel, Joseph Chazbani and Daniela Nikolova 74

Q-Learning in Action: Explainable and Verified Renewable Energy Source Selection
Afef Gafsi and Imen Ben Hafaiedh 84

Combining Large Language Models with Symbolic Reasoning for Robust Aspect-Based Sentiment Analysis
Moemen Said, Nadia Smairi, Houda Abadlia and Olfa Belkahla Driss 96

SHORT PAPERS

MIMO and Multi-Group Comparison Problem in Process Control: A Multivariate Statistical Framework for Systems Represented with Frequency Response Functions 109
Vittorio Lippi, Leonard Johard and Gabriele Landucci

A Biped Walking Consisting of Single-Support and Double-Support Phases 117
Yannick Aoustin and Alexander Formal'Skii

Analysis of Adhesion Effect on Grasp Stability with Robotic Gripper 125
Caroline Pascal and Mathieu Grossard

Robust Bell-State Preparation and Observation in Two-Qubit Systems Using Sliding Mode Methods 133
Kaydian Quintero, Bereket Berhane and Sergey V. Drakunov

Design of Neurocontrollers for Nonlinear Dynamic Systems with Summation and Difference Neurons 141
Ivan Sekaj, Aleš Melichar, Danil Zlobin, Miroslav Halás, Jaromír Skirkanič and Ivan Kenický

An Epipolar Geometry-Based Approach for Distributed Formation Control Using Vision 149
Patricia M. Tavares, Edgar I. Chavez-Aparicio, Jean-Bernard Hayet and Hector M. Becerra

Integrated Multi-Objective Optimization for AMR Fleet Operations in Automotive Manufacturing 157
Ahmed Tolba

STRIDE: Investigating Efficient Sampling Strategies for Neural Approximations of Optimal Control 166
Andrea Giusti, Maxime Monsieur and Branimir Mrak

Runway-Referenced Deep Reinforcement Learning Control for Autonomous Landing of a 6-DoF Aircraft Model 174
Melisa Nisa Aşlayan Üzer and Barış Başpınar

Evolutionary Fixed-Structure H_∞ Controller Synthesis with Performance Robustness against IQCs 182
Philippe Feyel

Spiking Neural Networks for Object Detection: A Review of 2024/2025 Advancements 191
Mikołaj Waśniewski, Jakub Caputa, Wojciech Chmiel and Maciej Wielgosz

A Genetic Algorithm Approach for LPG Transportation Scheduling under Loading and Unloading Constraints 199
Yadira Bravo, Pablo Proaño, Danilo Chavez and Alan Cuenca-Sánchez

Smooth Optimal Tracking-Control Approximating Maximum Control Energy Usage for a Class of Linear Systems with a Nonlinear Output Function and Control Limitations 207
Mir Mamunuzzaman and Jörg Mareczek

Reinforcement Learning Based Distributed Tracking Control of Multiple Uncertain Linear Systems 215
S. Hernandez and Wenjie Dong

Feature-Level Explainability for Hyperdimensional Computing-Based Anomaly Detection 222
Víctor Ortega, Soledad Escolar, Fernando Rincón, Jesús Barba, Julián Caba and Juan Carlos López

ROBOTICS AND AUTOMATION

FULL PAPERS

Model-Free Dynamic Compensation Based on Ultra-Local Modeling for Trajectory Tracking of an Aerial Manipulator 235
Gabriela M. Andaluz, Paulo Leica, Oscar Camacho and Angélica Quito

Enhanced Indoor and Parking Navigation for UGV: A Multi-Layered Algorithmic Overlay for Patchwork++ 243
Kevin Helvig, Julien Marzat and Martial Sanfourche

Visual Servoing under Scalar Depth Uncertainty: Classical, Predictive, and Takagi–Sugeno Control
Boucherika Mohamed Abdelmalek, Jason Chevrie, Laurent Vermeiren, Benyamine Allouche and Antoine Dequidt 252

A Dual-Stream Transformer Architecture for Illumination-Invariant TIR-LiDAR Person Tracking 262
Yuki Minase, Kanji Tanaka, Changjun Wang and Tomoe Hiroki

Model-Based Switching Manifold Embedded Terminal Second-Order Sliding Mode Control for Rehabilitation Exoskeletons 270
Brahim Brahmi, Wissem Dilmi, Jawhar Ghommam and Mohammad H. Rahman

Practical Feature-Based 3D Object Recognition for Complex-Shaped Sand Cores in Casting 278
Cong Huy Nguyen and Yoshiyuki Noda

LoRa-Based Large-Scale Indoor Localization Using Probabilistic Models of Strongest Received Signals 288
J. Ranz-Abad, S. Maldonado-Bascón, S. Lafuente-Arrroyo and R. Flor-Rabadan

Design and Experimental Validation of a 3-DOF Delta Robot with Trajectory Tracking Control
Pablo Rivadeneira, Viviana Moya, Juan Pablo Vásconez, Luis A. Córdova-Tandazo, Angélica Quito and William Chamorro 298

Comparative Evaluation of Path Planning Algorithms and Control Strategies for Omnidirectional Mobile Robot Navigation 308
Mateo García, Viviana Moya, Andrea Pilco, Juan Pablo Vásconez, Álvaro Prado-Romo and William Chamorro

Comparative Analysis of Bio-Inspired Optimization Algorithms for Nonlinear PID Tuning of a Mobile Manipulator 319
Pablo Proaño, Paulo Leica, Gabriela M. Andaluz, William Chamorro and Luis Morales

Continuous-Time Set-Transformers for Asynchronous WiFi-IMU Indoor Localization 331
Mohamed Bachar, Ilyass Abouelaziz and Yuehua Ding

Data-Efficient Pedestrian Multi-Object Tracking via Zero-Shot Detection and Lightweight DINOv3 Re-Identification 343
Manu Gupta, Oliver Zeilerbauer, Mohammad Reza Mohebbi and Mario Doeller

Safe Trajectory Planning and Control via Flatness-Based MPC with Convex Polytopic Constraints
Solal Wargon, Harendra Rangaradjou, Christophe Louembet, Martin Mujica, Adrien Durand-Petiteville and Viviane Cadenat 355

Indoor SLAM and Waypoint Navigation for a Low-Cost Mobile Robot: A Heterogeneous FPGA and Embedded GPU Architecture with BLE-Assisted Localization 365
Tudor Mihai Crihălmeanu, Mircea Paul Mureşan and Sergiu Nedevschi

SHORT PAPERS

A Novel Piecewise Autonomous Multi-Model Odometry Approach 377
Andrea Righetti, Giuseppe Spampinato, Arcangelo Bruna and Giorgio Battistelli

Liquids Detection Using SWIR Data and Deep Learning 383
Antonio Buemi, Andrea Accettola, Arcangelo Ranieri Bruna, Cosimo Ieracitano, Nadia Mammone, Massimo Merenda and Valeria Tomaselli

Deep Global Point Cloud Registration for Long-Term Robot Localization 390
Míriam Máximo, Enrique Heredia-Aguado, Esther González-Amorós, Francisco José Martínez-Peral, Luis Payá, David Valiente and Mónica Ballesta

A Deep Learning Framework for Underwater Perception and Navigation Using Super-Resolution Enhanced YOLOv8 and Reinforcement Learning 398
Khaled Benkhoud, Saliha Sebbane and Sarra Bouzayane

Traversability-Aware Frontier Selection for Goal-Directed Exploration: An Ablation Study 406
Carlos Tampier Cotoras, Frank Höller and Dirk Schulz

Workspace Computation via Constrained Forward Kinematics: Application to Robotic Manipulators 414
Francisco José Martínez-Peral, Míriam Máximo, Mónica Ballesta, José Vicente Segura-Heras, Arturo Gil and Carlos Perez-Vidal

Influence of FEA Modeling Parameters on the Accuracy of Robot-Link Compliance Identification 422
David Eskoundos, Anatol Pashkevich, Kevin Subrin and Stéphane Caro

A Certification-Based Predefined Time for Dynamic Average Consensus for Multi-Agent Systems 430
Muzammil Ali, Muhammad Sabih, Khubaib Ahmad and Aamer Iqbal Bhatti

Automatic Labeling for Low-Light Pedestrian Detection 438
Dimitrios Bouzoulas, Eerik Alamikkotervo and Risto Ojala

KGAT-RL: Kernelized Graph Attention Networks with Drift-Triggered PPO Residual Correction for Robust Multi-Horizon Traffic Forecasting 447
Mohammad Reza Mohebbi, Mohammad Javad Hassanzada, Erfan Ranjbar, Mohammad Soltanirad and Mario Döller

A Bounded Circle Covering Strategy for Mobile Robotics Inspired by Nesting Optimization 455
Stépane Binon, Lucas Si Larbi and Eric Lucet

Redundant Safety Mechanisms of Artificial Intelligence Models for Autonomous Driving 463
Florence Wagner, Heinrich Litagin and Rolf Jung

Multi-Arm Motion Planning Based on Conflict Elimination 471
Matouš Kulhan and Pavel Surynek

Liquid Recurrent Reinforcement Learning for Constrained Multi-UAV 3D Coverage Path Planning 479
Mohammad Reza Mohebbi, Mario Döller and Nauman Ahmed

Robust Fractional-Order Sliding Mode Control for Trajectory Tracking of a Mobile Manipulator 487
Gabriela M. Andaluz, Paulo Leica, Luis Morales, Pablo Proaño and Oscar Camacho

Hamiltonian-Based Human Intention Estimation for Compliant Exoskeleton Robot 495
Brahim Brahmi, Abderraouf Maoudj, Jawhar Ghommam and Mohammad H. Rahman

A Pose-Based AI Coach for Exercise Recognition and Quality Assessment 503
Ahmed Aly, Al Amir Eldomiatty, Zeyad Elsayed and Walid Gomaa

From Reactive to Map-Based AI: Tuned Local LLMs for Semantic Zone Inference in Object-Goal Navigation 511
Yudai Noda, Kanji Tanaka, Sosi Tanaka, Changjun Wang and Tomoe Hiroki

Self-Supervised Appearance-Based in-Plane Rotation Estimation 519
Sreejani Chatterjee, Nichols Taylor and Daniel Nikovski

Vision-Based Excavator Boom Pose Estimation with Sensor-Fused Uncertainty-Aware Tracking 527
Robin Vetsch, Lucas Falch, Christoph Würsch and Sebastien Gros

Optimization of a Higher Order Sliding Mode Controller for Tactical Missile Guidance Applications 538
Mirza Hodžić, Naser Prljača and Jakub Osmić

A Retrofit State Monitoring Framework for Safe Robot-Instrument Interaction in Legacy Analytical Laboratories 547
Mohamed Ali Tlili, Thomas Roddelkopf, Heidi Fleischer and Kerstin Thurow

Monitoring Endangered Animals Using Blockchain and Artificial Vision 555
Ludwin Lope Cala and Roseli Aparecida Francelin Romero

Robotic Manipulator Motion Planning and Obstacle Avoidance Using Physics-Informed Reinforcement Learning via Artificial Potential Fields 563
Edmond Samia, Alexandre Parant, Laurent Arcese and Yuehua Ding

GARL: Grasp-Aware Reinforcement Learning for Robust and Dexterous in-Hand Rotation 571
Guillaume Thomas, Jaonary Rabarisoa, David Filliat, Quoc-Cuong Pham and Mathieu Grossard

Optimization-Based Inverse Kinematics for a 4-DOF Robotic Arm: A Comparative Analysis with Real-Time ROS Implementation 582
Jofre Suárez, Viviana Moya, Juan Pablo Vásconez, Álvaro Prado-Romo, Andrea Pilco and William Chamorro

Towards Single-Contact-Point Textile Unfolding with Hierarchical Point Selection and Trajectory Generation 590
Santiago T. Puente and Carlos Mateo-Agulló

Scene Description Generation for Traffic Scenes Using Segmentation and Visual Language Models 598
Boitor Diana, Sergiu Nedevschi and Raluca Didona Brehar

Fault-Tolerant Multi-Sensor Fusion via Factor Graph Optimisation for Mobile Robot Localisation 606
Clovis Adeler, Maan El Badaoui El Najjar, Cindy Cappelle and Denis Pomorski

TASC: Fast Skill Chaining for Complex Sequential Robotic Tasks via Pure Imitation Learning 614
Ryan James Hayward and Kasim Terzić

Electro-Mechanical Modelling of Generatively Designed Robotic Arm Segments with Integrated Wire Routing 622
Johannes Lamm and Thomas M. Wendt

Mapless Rover Navigation Leveraging Obstacle Perception with k-NN Vector-Space Clustering 629
Giovanni De Gasperis, Sante Dino Facchini and Luigi Iaconi

Testing Video Anomaly Detection through State Space Models on Challenging Real-World Surveillance Datasets 639
Esther González-Amorós, Alberto Chueco, Míriam Máximo, Carlos Viegas, Adrián Peidró and Luis Paya

Comparative Analysis of Eye-in-Hand and Eye-to-Hand Motion Estimation Techniques for Dynamic Target Tracking in Visual Servoing 649
Mario Talarico, Ibrahim Ben-Othman, Andrea Monguzzi, Navvab Kashiri, Andrea Maria Zanchettin, Elena De Momi and Paolo Rocco

Point Cloud Streaming from Mobile Robots under Bandwidth and Scene Complexity Constraints 657
Mohamad El Dika, Jalal Possik, Petra Bilane, Jad Nassar, Roger Achkar and Charles Yaacoub

Tether-Length-Constrained Reference Governance for Cooperative UGV-UAV Agricultural Inspection 664
Ammar Selim, Zhengjia Xu, Mohsen Zahmatkesh and James E. Pickering

Energy-Aware Adaptive Communication for Efficient Mobile Robots 675
Giorgio Vivoli and Navvab Kashiri

Proposal of an Intuitive Robot-Summoning Method for Agricultural Mobile Robots Using a Retroreflective Handheld Device 683
Atsuhiro Ohno, Sho Yokota and Akihiro Matsumoto

AIVFusion+3D Features: An Approach for Features Extraction Aiming to Implement Sensors Data Fusion from Monocular Camera and LIDAR 691
Fernando S. Osorio, Felipe Freitas Maia, Alexandre M. Ishi Jr, Marcos Vinicius da Rocha, Vitor Augusto Pereira and Leticia de Carvalho Pereira

A Real-Time Framework for Multi-Person Reciprocal Line-of-Sight Modeling and Recovery in Human-Robot Interaction 699
Jesus Eduardo Hermosilla-Diaz, Jeeangh Jennessi Reyes-Montiel, Yaser Aguas-Garcia and Antonio Marin-Hernandez

Implementation of an Autonomous Liquid Pouring System Based on Source Container Modeling for Different Target Containers 707
Jeeangh Jennessi Reyes-Montiel, Jesus Eduardo Hermosilla-Diaz, Yaser Aguas-Garcia, Yahir Hernandez-Mella and Antonio Marin-Hernandez

Kalman Filter Fusion of Forward Kinematics and Vision for Robot Pose Estimation in Construction Environments 714
Juan Pablo Cahueque Perez, Christine Chevallereau, Corentin Monet, Loïc Buatois and Andrea Cherubini

SIGNAL PROCESSING, SENSORS, SYSTEMS MODELLING AND CONTROL

FULL PAPERS

Extended Neural Ordinary Differential Equations 727
Leonard Johard and Vittorio Lippi

Robotic Sorting of Dirty Laundry Using a Multispectral Vision-Based Grasping Strategy 736
Ferréol Gagey, Gabriel Prudhomme and Mathieu Grossard

Sliding Mode Control of Single-Mode Photon by Nonlinear Quadrature Actuation 744
Sergey V. Drakunov and Bereket Berhane

A Biquaternion-Based State Observer and Its Application to a Two-Level Quantum System 751
Andrei A. Markov and Sergey V. Drakunov

Beyond Human Captions: Semantic Supervision and PCA-Regularized Decoding for Stable-Diffusion-Based fMRI Reconstruction 760
Abdallah Mohammed AboHassan, Ehab Elshazly, Tetsuji Ogawa and Ahmed Fares

Negative Binomial Mixtures with Data-Driven Initialization for Streaming Overdispersed Count Data 772
Hana Sedláčková and Evženie Uglickich

EJUST-GYM-3: A Multi-View Exercise Dataset for Human Activity Recognition and Motion Analysis 783
Mohamed Abouelella, Ahmed Ashraf and Walid Gomaa

Effects of Seasonality in Dengue Spread Equilibria and Transient Behaviors 795
Flavia Altamura, Paolo Di Giamberardino and Daniela Iacoviello

Implicit Predecessor-Based Region of Attraction Estimation and Robust Invariance Analysis for a Two-Wheeled Inverted Pendulum 805
Lorenzo Fici and Guillaume Ducard

Data-Driven Modeling of Complex MIMO LPV Dynamics: A Scalable Iterative Approach Using Khatri–Rao Matrix Augmentation and Kalman-Enhanced State Estimation 813
Jorge Puerto, Edwar Jacinto and Fredy Martínez

SHORT PAPER

Interpretable Human Activity Recognition via Phase-Space Trajectories 827
Walid Gomaa and Ahmed Gomaa

SPECIAL SESSION ON HYBRID INTELLIGENCE: INTEGRATING PHYSICS-BASED PRINCIPLES WITH DATA-DRIVEN AI

FULL PAPERS

Markerless Human Body Volume and Center of Mass Estimation from Dual RGB-D Cameras with Biomechanical Preliminary Assessment 839
Usman Asghar, Shengjie He, Giulia Bassani, Josep Rueda Collell, Carlo Alberto Avizzano and Alessandro Filippeschi

A Gated Temporal Causality Scheme for Physics-Informed Neural Networks 850
Peng Shu Ng, Elisa Y. M. Ang, Peng Cheng Wang and Teng Yong Ng

When Does Physics Help? A Persistence-Aware Benchmark of Hybrid pH Prediction in Geothermal Steam Washing 857
Massimo Satler and Carlo Alberto Avizzano

SHORT PAPERS

Hybrid Intelligence for Industrial Process Optimization 869
Vittorio Lippi and Leonard Johard

A Simple Gripper Interface for Simulator-Agnostic Cloth Manipulation 877
Abhilash Nayak, Franco Coltraro, Maria Alberich-Carramiñana and Carme Torras

Experimental Comparison of Linear MPC and Adaptive MPC for Transparency in a 1-DOF Elbow Exoskeleton: A Discussion 885
Josep Rueda Collell, Alberto Maria Nobili, Shengjie He, Usman Asghar, Fabio Salsedo and Alessandro Filippeschi

SPECIAL SESSION ON BRIDGING THE GAP IN COLLABORATIVE ROBOTICS: FROM THEORY TO REAL APPLICATIONS

FULL PAPERS

Cobot-Based Operator Assistance for Manual Meat Processing: Experimental Validation on Porcine Cuts 897
Alexis Babut, Chedli Bouzgarrou, Laurent Sabourin and Nicolas Bouton

Facial Emotion Recognition Based on FAUs and DINOv3 Embeddings with Reliability-Gated Fusion 906
Giuseppe Lorenzo Di Prima, Francesco Termine, Ankit Kumar, Giuseppe Sutera, Filippo Sanfilippo and Mario Collotta

Keeping the Franka Emika Panda Alive: A ROS 2 Stack with a Reliable Position Interface 918
Antonio Langella, Davide Risi, Vincenzo Petrone, Enrico Ferrentino and Pasquale Chiacchio

Robust Learning from Demonstration Pipeline Based on Inertial Motion Capture and Force-Synchronised Demonstration Capture for Industrial Demoulding Tasks by Collaborative Robots 930
Alejandro Solar Iglesias, Gabriel Novas Domínguez, Saltanat Seitzhan and Juan Antonio Corrales Ramón

SHORT PAPERS

Automated IMU-Based Ergonomic Evaluation and Postural Risk Assessment for Industrial Automation Frameworks 943
Jorge Galindo, Alejandro Solar, Juan Antonio Corrales Ramón and Hernan Abaunza

Transformer-Based Ego-Motion Filtering for Tracking in Eye-in-Hand Event-Driven Visual Servoing 949
Gabriel Barrigón-Gómez, Adrián Fernández-Casas, Carlos Mateo-Agulló, Juan Antonio Corrales Ramón and Cédric Demonceaux

Slice More, Push Less: Admittance-Controlled Force Amplification for Collaborative Meat Cutting 957
Saltanat Seitzhan and Juan Antonio Corrales Ramon

Design of a Low-Cost Planar Delta Robot for Agri-Food Sorting Applications 964
Simão Lopes, Martim Aguiar, Yebo Lu and Pedro Dinis Gaspar

AUTHOR INDEX 973

Testing Video Anomaly Detection through State Space Models on Challenging Real-World Surveillance Datasets

Esther González-Amorós[1][a], Alberto Chueco[1], Míriam Máximo[1][b], Carlos Viegas[3][c], Adrián Peidró[1][d] and Luis Paya[1,2][e]

[1]*Instituto de Investigación en Ingeniería (I3E), Universidad Miguel Hernández, Elche 03202, Spain*

[2]*Valencian Graduate School and Research Network of Artificial Intelligence, Valencia, Spain*

[3]*ADAI, Department of Mechanical Engineering, University of Coimbra, Coimbra, Portugal*

{*esther.gonzaleza, mmaximo, apeidro, lpaya*}*@umh.es, alberto.chueco@goumh.umh.es, carlos.viegas@uc.pt*

Keywords: Video Anomaly Detection, State Space Models, Model Efficiency, Surveillance, Computational Cost.

Abstract: This work presents the first comprehensive evaluation of the VADMamba framework on the UCF-Crime dataset, a challenging benchmark for video anomaly detection (VAD) in real-world surveillance environments. While VADMamba has demonstrated strong performance on controlled-environment benchmarks, its applicability to semantically complex surveillance scenarios remains unexplored. Through systematic experimentation, we demonstrate that a lightweight, prediction-only configuration utilizing a depth profile of $[3,3,3,3]$ achieves an AUC (Area Under Curve) of 72.23% and 1.8 times fewer parameters and 2.6 times fewer FLOPs than the full original dual-task pipeline, by eliminating the optical flow reconstruction bottleneck. This performance surpasses the previous state-of-the-art one-class classification methods on UCF-Crime, despite training on only 260 videos of the available 800 normal videos. Cross-dataset transfer experiments reveal substantial domain gaps when models trained on controlled benchmarks are evaluated on UCF-Crime, highlighting the limited representativeness of widely-used datasets for real-world deployment.

1 INTRODUCTION

Video Anomaly Detection (VAD) has become a fundamental component of modern security infrastructure due to the increase of video surveillance systems. With the deployment of cameras in public or private spaces, the volume of video data generated daily has grown to a scale that exceeds the capacity of human operators to monitor effectively (Sultani et al., 2018). Manual inspection is not only time-consuming and labor-intensive, but also prone to errors due to fatigue and limited attention span, particularly when anomalous events represent a small fraction of the recorded data (Nawaratne et al., 2019). This has driven the development of automated VAD systems, which aim to identify unusual events in video sequences automatically.

The predominant approach to VAD follows an unsupervised learning paradigm, where models are trained exclusively on normal video data and learn to identify anomalies as deviations from the normal activities learned. Following this paradigm, two main strategies have been developed: frame prediction and frame reconstruction. On the one hand, frame prediction methods receive a sequence of consecutive frames as input and learn to generate the next frame in the sequence. Since the model has only been exposed to normal patterns during training, it produces accurate predictions for regular events but fails to anticipate anomalous situations, resulting in a high prediction error that serves as an anomaly score (Liu et al., 2018). On the other hand, frame reconstruction methods receive an input frame or its corresponding optical flow and attempt to reproduce it through an encoder-decoder architecture. The encoder compresses the input into a compact latent representation that captures the essential characteristics of normal patterns, while the decoder attempts to reconstruct the original input from this compressed representation. During training, the model learns to efficiently encode and decode the regularities present in normal video data. At in-

[a] https://orcid.org/0009-0002-0353-6727
[b] https://orcid.org/0009-0005-5297-9059
[c] https://orcid.org/0000-0002-9242-3195
[d] https://orcid.org/0000-0002-4565-496X
[e] https://orcid.org/0000-0002-3045-4316

González-Amorós, E., Chueco, A., Máximo, M., Viegas, C., Peidró, A. and Paya, L.
Testing Video Anomaly Detection through State Space Models on Challenging Real-World Surveillance Datasets.
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 23rd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2026) - Volume 1*, pages 639-648
ISBN: 978-989-758-870-9; ISSN: 2184-2809
Proceedings Copyright © 2026 by SCITEPRESS – Science and Technology Publications, Lda.

ference time, when the model encounters anomalous content that was not represented during training, the latent space lacks adequate representations to faithfully reproduce it, causing the reconstruction quality to degrade and producing higher errors that indicate the presence of an anomaly (Gong et al., 2019).

In terms of methodology, widely adopted CNN-based (Convolutional Neural Networks) methods such as (Gong et al., 2019) and (Park et al., 2020) have demonstrated strong spatial feature extraction capabilities, while Transformer-based methods have shown improved temporal modeling through self-attention mechanisms (Dosovitskiy et al., 2020). However, both families of architectures present computational limitations: CNNs find it challenging to capture long-range temporal dependencies due to their limited receptive fields, and Transformers suffer from computational costs related to sequence length, making them too expensive for processing long video sequences in surveillance environments.

The emergence of State Space Models (SSMs), particularly the Mamba architecture (Gu and Dao, 2023), has introduced an alternative which combines the long-range modeling capability of Transformers with linear computational cost. This kind of architectures has been adapted to visual tasks through VMamba (Liu et al., 2024) and Vision Mamba (Zhu et al., 2024), demonstrating competitive performance with significantly reduced computational cost. VADMamba (Lyu et al., 2025) represents the first application of Mamba regarding the VAD task, proposing a framework that combines Vector Quantization (VQ) with Mamba-based blocks in a dual-branch architecture for frame prediction and optical flow reconstruction, VQ-Mamba UNet (VQ-MaU).

VADMamba has demonstrated strong performance on established benchmarks such as UCSD Ped2 (Li et al., 2013), CUHK Avenue (Lu et al., 2013), and ShanghaiTech (Luo et al., 2017). These datasets have been instrumental in the VAD area, mainly focusing on detecting deviations in pedestrian behavior patterns within structured environments such as identifying unusual trajectories, unexpected objects or vehicles, or abnormal movement speeds in street scenes. Nevertheless, the scope of anomalies addressed by these benchmarks differs considerably from the challenges encountered in surveillance systems, where the objective extends to detecting other abnormal events like criminal activities, violent incidents or traffic across heterogeneous scenarios. UCF-Crime (Sultani et al., 2018) includes 1900 untrimmed videos divided into 13 categories of real-world anomalous events, offering a benchmark that reflects the conditions and requirements of practical video surveillance applications with greater precision.

Consequently, this work evaluates the VADMamba framework on the UCF-Crime dataset, introducing the first implementation of this architecture on this benchmark. By utilizing exclusively the prediction branch and training on a restricted subset of 260 out of the 800 normal videos available in this dataset, the studied model achieves an Area Under the Curve (AUC) of 72.23% when using the most complete variant of those proposed in Lyu et al. (2025). This performance establishes a new state-of-the-art result among unsupervised methods on the UCF-Crime dataset, surpassing the previous best method.

The remainder of this manuscript is structured as follows. Section 2 reviews the state of the art on VAD. Section 3 presents the experimental methodology. Section 4 describes the experiments and analyzes the results. Finally, conclusions and future work are discussed in Section 5.

2 RELATED WORK

Early deep learning approaches to VAD employed convolutional autoencoders to learn normal spatiotemporal regularities, detecting anomalies through reconstruction errors (Hasan et al., 2016). MemAE (Gong et al., 2019) improved this paradigm by incorporating a memory module that stores prototypical normal representations, while MNAD (Park et al., 2020) introduced mechanisms for managing the diversity of normal behaviours. In parallel, prediction-based methods formulated VAD as a future frame prediction task, where high prediction errors indicate anomalies. Liu et al. (2018) established this approach combining a U-Net generator with optical flow constraints. Several works have combined both strategies, leveraging prediction for appearance modelling and reconstruction of optical flow estimated by networks such as FlowNet 2.0 (Ilg et al., 2017) for motion modelling.

However, CNNs are inherently limited in capturing long-range temporal dependencies due to their local receptive fields. Vision Transformers addressed this through attention mechanisms (Dosovitskiy et al., 2020), but their quadratic computational complexity limits applicability in resource-constrained environments. More recently, SSMs, particularly Mamba (Gu and Dao, 2023), have achieved Transformer-like long-range modelling with linear complexity, motivating adaptations to visual tasks through VMamba (Liu et al., 2024) and Vision Mamba (Zhu et al., 2024). In the domain of VAD, VADMamba (Lyu

et al., 2025) was the first to integrate SSMs, proposing a VQ-MaU that combines VQ with NVSS (Non-negative Visual State Space) blocks in a dual-task architecture for frame prediction and optical flow reconstruction.

Regarding benchmarks, the original evaluation of the VADMamba framework was conducted across three distinct benchmarks: UCSD Ped2 (Li et al., 2013), CUHK Avenue (Lu et al., 2013), and ShanghaiTech (Luo et al., 2017). These datasets focus on pedestrian behaviour in structured environments with fixed cameras, relatively homogeneous scenes, and anomalies defined as deviations from normal pedestrian patterns.

UCF-Crime (Sultani et al., 2018) offers a substantially more challenging setting, comprising 1,900 untrimmed real-world surveillance videos across 13 anomaly categories. While Sultani et al. (2018) formulated anomaly detection on this dataset as a Weakly Supervised (WS) Multiple Instance Learning (MIL) problem, and subsequent works such as RTFM (Tian et al., 2021) refined this paradigm, the application of unsupervised frame-prediction methods remains underexplored. In this paper, we address this gap by providing the first evaluation of VADMamba on UCF-Crime, analysing both detection performance and computational efficiency in this more demanding surveillance scenario.

3 METHODOLOGY

VADMamba (Lyu et al., 2025) employs a dual-branch VQ-MaU architecture comprising a Frame Prediction (FP) model and an Optical Flow Reconstruction (FR) model. Both models feature a symmetric four-stage encoder-decoder structure denoted as $[n_1, n_2, n_3, n_4]$, where n_i represents the number of NVSS blocks in the i-th stage—adjustable to scale representation capacity to dataset complexity. The encoder progressively reduces spatial resolution while increasing channel depth through patch merging layers, with each stage i containing n_i NVSS blocks that apply Mamba's selective scan mechanism in four spatial directions to capture long-range dependencies. At the bottleneck, the latent features are processed through a VQ layer, which discretizes them by finding the nearest vector in its codebook using the Euclidean (L_2) distance. The decoder mirrors the encoder with patch expanding layers and skip connections. The anomaly score is derived from the Peak Signal-to-Noise Ratio (PSNR) between the predicted and actual frames, where a lower PSNR indicates a higher prediction error and, consequently, a higher likelihood of an anomaly.

3.1 Adaptation to UCF-Crime

VADMamba was originally validated on UCSD Ped2, CUHK Avenue, and ShanghaiTech benchmarks, where videos are captured from fixed cameras in controlled environments, anomalies involve pedestrian deviations, and the datasets are distributed in formats readily compatible with the framework's input pipeline. On the contrary, applying the framework to UCF-Crime introduced challenges.

The first challenge is related to frame extraction and storage constraints. VADMamba operates on individual frames stored as image files on disk, not on raw video. The UCF-Crime dataset is structurally partitioned into a training set containing 800 normal and 810 anomalous videos, and a testing set comprising the remaining 150 normal and 140 anomalous videos. Decoding the full UCF-Crime dataset produced approximately 4.5 million normal training frames and over 1.1 million testing frames. The complete extraction of all 800 normal training videos would require approximately 151 GB of disk space, exceeding typical embedded platform storage budgets (32–64 GB). Consequently, this constraint forced a reduction to 260 normal videos, which produced 1,482,044 raw frames occupying 49 GB of storage, from which 1,477,884 overlapping 16-frame input sequences were generated using a single-frame sliding window. The selection followed a systematic criterion: the first 300 normal videos in sequential order from the dataset were chosen, excluding the 40 clips that overlap with the official test partition to prevent data leakage.

The frame-level binary annotations were constructed by mapping the dataset's video-level temporal annotations to individual frame indices across all 290 test videos, accounting for variable frame rates and video durations. The resulting label vector contains 1,102,595 entries, each classifying a single frame as normal (0) or anomalous (1).

Beyond these engineering challenges, UCF-Crime presents fundamental differences from the original benchmarks. The videos span diverse camera viewpoints, variable lighting conditions, and heterogeneous scene compositions ranging from indoor to outdoor environments. The anomalous events exhibit high semantic variability and represent only 7.65% of the total test frames (84,338 anomalous out of 1,102,595 total test frames), producing a highly imbalanced distribution that reflects the conditions of real-world surveillance scenarios.

3.2 Experimental Design

Throughout this work, One-Class Classification (OCC) paradigm is adopted. The model is trained exclusively on normal video data without access to any anomalous samples or labels during training. Unlike fully unsupervised (US) methods, which train on mixed normal and anomalous data without labels (Zaheer et al., 2022), and WS methods, which require video-level anomaly annotations (Sultani et al., 2018), OCC represents a highly restrictive and deployment-oriented supervision setting. To systematically evaluate the practical deployment of SSM-based VAD in real-world surveillance, four complementary experimental axes are defined:

- **Task Ablation:** The objective of this task is to compare the two configurations of the pipeline illustrated in Figure 1. The prediction-only path (Task 0) employs a single VQ-MaU encoder-decoder (Model 1) that receives 16 concatenated RGB frames and predicts the subsequent frame. The anomaly score is derived from the PSNR between the predicted and actual frame. The dual-task path (Task 2) additionally incorporates FlowNet2SD, a pretrained optical flow estimator called twice per sequence, and a second encoder-decoder (Model 2) that reconstructs the resulting flow fields. Both branch scores are fused through per-video selection. This experiment tests whether the reconstruction branch provides sufficient empirical value to justify this computational cost within the context of UCF-Crime, or whether the prediction branch alone is sufficient.
- **Depth Scaling:** In this experiment, a comparison between three depth configurations, trained for up to 50 epochs and evaluated at 10-epoch intervals. These configurations define the number of NVSS blocks assigned to each of the four stages in both the encoder and decoder of the VQ-MaU architecture: $[1,1,1,1]$ uses a single block per stage, offering high efficiency but limited representational capacity; $[2,2,2,2]$ doubles the blocks, providing a broader receptive field through additional four-directional selective scan operations at each resolution level; and $[3,3,3,3]$ assigns three blocks per stage, reaching the highest representational capacity at the cost of increased computation and memory. This experiment determines the optimal training horizon under the reduced 260-video dataset to prevent early overfitting, while assessing whether scaling model capacity addresses architectural or data-driven limits on UCF-Crime.
- **Cross-Dataset Transfer:** The pretrained VADMamba models released by the original authors trained on UCSD Ped2, CUHK Avenue, and ShanghaiTech, respectively, are evaluated directly on UCF-Crime without any retraining or fine-tuning. This experiment does not appear in the original VADMamba publication and addresses a critical question often overlooked in the VAD literature: whether the high AUC values reported on controlled benchmarks possess enough predictive value for operational performance in real-world surveillance conditions. The three source datasets span a spectrum of complexity, from the single-camera setup of Ped2 to the multi-camera environment of ShanghaiTech, enabling an analysis of how source dataset diversity correlates with transfer performance.
- **Efficiency Profiling:** Parameters, FLOPs, inference throughput, latency per sequence, training throughput, and GPU memory consumption are rigorously measured across all configurations. This analysis extends beyond the metrics reported in the original work by providing a detailed breakdown of the computational cost attributable to each pipeline component, with particular emphasis on FlowNet2SD as the dominant bottleneck. The efficiency metrics are contextualized against the real-time processing requirements of standard surveillance cameras, establishing a practical reference for deployment feasibility.

4 EXPERIMENTS AND RESULTS

4.1 Implementation Details

Experiments are conducted on an NVIDIA RTX 5090 GPU (32 CPU cores, 64 GB RAM) using PyTorch 2.11. Models are trained with Adam ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay 10^{-4}), CosineAnnealingLR, batch sizes of 32 (for $[1,1,1,1]$ and $[2,2,2,2]$) or 16 ($[3,3,3,3]$), and initial learning rate 2×10^{-4}. Input sequences consist of $t = 16$ RGB frames at 256×256 resolution. The VQ codebook sizes are $K = 50$ for prediction and $K = 10$ for reconstruction. A fixed seed (1111) is used throughout. Frame-level AUC serves as the primary metric.

The experimental evaluation is structured along the four axes defined in Section 3.2. First, the task ablation study compares the prediction-only configuration (Task 0) against the full dual-task pipeline (Task 2) using a depth configuration of $[1,1,1,1]$. This setup trains the prediction branch for 10 and 20 epochs in Task 0, and combines 5 and 10 prediction epochs with 5 and 10 reconstruction epochs in Task 2.

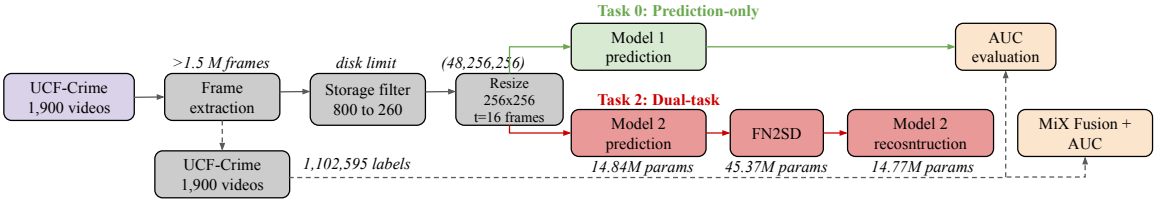


Figure 1: Overview of the experimental pipeline. The preprocessing stage is common to both configurations. Task 0 (prediction-only) eliminates the reconstruction branch, while task 2 (dual configuration) includes both the prediction and the reconstruction, using FlowNet2SD to calculate the optical flow. Dashed lines indicate the ground-truth labels distributed to the evaluation blocks.

Second, the convergence and depth analysis extends the prediction-only mode across three model capacities: $[1,1,1,1]$ with 14.84M parameters, $[2,2,2,2]$ with 28.21M parameters and $[3,3,3,3]$ with 41.57M parameters. These variants are trained for up to 50 epochs, with AUC evaluation conducted at 10-epoch intervals, producing fifty discrete checkpoints that characterize the convergence trajectory and the onset of overfitting for each capacity level.

Third, the cross-dataset generalisation study loads the pretrained weights released by the original VAD-Mamba authors (Lyu et al., 2025) for UCSD Ped2, CUHK Avenue, and ShanghaiTech, evaluating them directly on the UCF-Crime test partition without any retraining, fine-tuning, or adaptation of the input pipeline.

Finally, the computational efficiency profiling collects the number of parameters, FLOPs, inference throughput, latency per sequence, training throughput measured in samples per second, and peak GPU memory consumption for every model configuration. Inference metrics are computed as the average over 100 independent forward passes with a single input sequence to ensure statistical stability, while training throughput is recorded directly from the epoch-level logs generated during the convergence experiments.

4.2 Detection Performance on UCF-Crime

4.2.1 Task Ablation: Prediction vs. Dual-Task

Table 1 compares the detection performance of the prediction-only (P) configuration (Task 0) against the full dual-task pipeline (Task 2), which also includes reconstruction (R), both using depth $[1,1,1,1]$ and 260 training videos.

The results reveal two key findings. First, the prediction-only configuration achieves a higher AUC (68.25%) than the best dual-task result (60.78% MiX AUC), despite using only one of the two branches. The MiX fusion strategy, which selects the best-performing branch per video, is counterproductive when the reconstruction branch is insufficiently trained: it selects reconstruction scores for certain videos where they appear locally superior but are in fact unreliable, degrading the global AUC. Second, the prediction branch converges to a similar AUC in both configurations (67.54% in Task 2 vs. 67.44% in Task 0 at 10 epochs), confirming that the prediction capacity is not affected by the presence or absence of the reconstruction branch during training.

Table 1: Task ablation on UCF-Crime AUC (%).

Config.	AUC_P	AUC_R	AUC_MiX
Task 2 (ep.5 P)	65.25	52.82	56.33
Task 2 (ep.10 P)	67.54	52.82	60.65
Task 2 (ep.5 R)	67.54	55.85	56.90
Task 2 (ep.10 R)	67.54	54.28	60.78
Task 0 (ep.10)	67.44	—	—
Task 0 (ep.20)	**68.25**	—	—

The low AUC of the reconstruction branch (52%–55%) warrants further analysis, as several compounding factors may contribute to this underperformance. First, a significant domain shift affects the optical flow estimation, given that FlowNet2SD was pretrained on synthetic datasets with highly controlled motion patterns; in contrast, UCF-Crime contains heterogeneous camera movements, variable frame rates, and complex scene dynamics that are expected to produce noisy or unreliable flow estimates during inference. Second, the model may suffer from insufficient training horizons; due to the high computational overhead of FlowNet2SD, which requires approximately 7.6 hours per epoch, the reconstruction branch was trained for 5 to 10 epochs, a constrained timeline that may prevent the network from learning robust and generalizable flow representations in such a challenging domain. These limitations compound through the MiX fusion strategy, which currently lacks a dynamic confidence mechanism to discount or down-weight unreliable branch scores when the underlying optical flow signal degrades.

4.2.2 Convergence and Overfitting Analysis

Table 2 presents the AUC evolution across training epochs for the prediction-only configuration at three depth levels.

Table 2: AUC (%) vs. training epochs for prediction-only mode at three model depths.

Epoch	$[1,1,1,1]$	$[2,2,2,2]$	$[3,3,3,3]$
10	67.44	68.76	68.82
20	**68.25**	69.88	70.40
30	68.13 ↓	**69.90**	**72.23**
40	67.72 ↓	68.90 ↓	69.77 ↓
50	66.86 ↓	68.43 ↓	68.57 ↓

All three configurations exhibit the same convergence pattern: the AUC increases during the early epochs, reaches an optimum, and then progressively decreases due to overfitting. The lightweight model $[1,1,1,1]$ peaks at epoch 20 with 68.25% AUC, the intermediate model $[2,2,2,2]$ peaks at epoch 30 with 69.90%, and the deepest model $[3,3,3,3]$ also peaks at epoch 30 with 72.23%. Notably, while $[2,2,2,2]$ delays the onset of overfitting by 10 epochs relative to $[1,1,1,1]$, the deepest model $[3,3,3,3]$ does not delay it further, indicating that the training horizon is constrained by data diversity rather than model capacity. Furthermore, the overfitting effect is more severe for deeper models: the AUC drop from peak to epoch 40 is 0.12 percentage points (pp) for $[1,1,1,1]$, 1.00 pp for $[2,2,2,2]$, and 2.46 pp for $[3,3,3,3]$. This behavior suggests that larger models memorize the limited training patterns more aggressively once they exhaust the available normal variability.

The overfitting behaviour can be attributed to the limited diversity of the training set: with 260 normal videos, the model learns the regularities of the available scenes quickly and subsequently begins memorising video-specific patterns rather than generalising. Consequently, extended training schedules are counterproductive, making validation-based early stopping essential.

It should be noted that these results were obtained with only 260 of the 800 available normal training videos due to storage constraints as argued in Section 3.1. Training with the complete set would likely improve both the peak AUC and the resistance to overfitting by exposing the model to a broader range of normal patterns, potentially shifting the optimal epoch further and reducing the performance gap with the original VADMamba results. As a compelling direction for future studies, training the architecture with the entire 800 normal videos could offer invaluable insights into the absolute theoretical performance bounds of the model.

An analysis of the frame-level anomaly score trajectories in Fig. 2 highlights the specific operational strengths and limitations of the framework across different video event types:

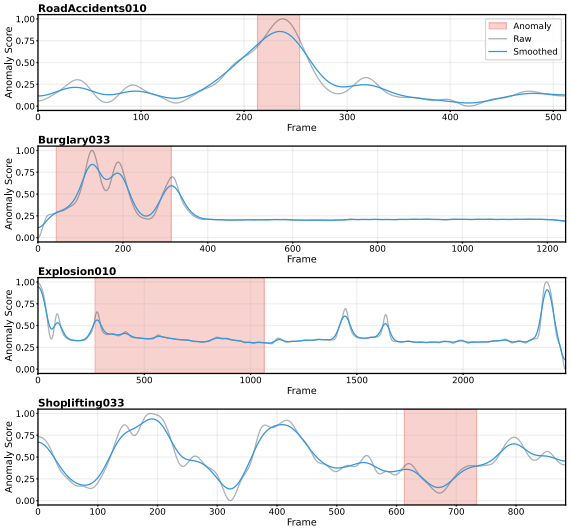


Figure 2: Frame-level anomaly score trajectories. Red shaded regions indicate ground-truth anomalous segments. The blue curve shows the smoothed anomaly score.

First, in the sequence *RoadAccidents010* (Fig. 2, top), the framework demonstrates excellent alignment, as the smoothed score maps a clear, localized trajectory that peaks precisely within the ground-truth interval. Second, the sequence *Burglary033* (Fig. 2, upper-middle) showcases a highly responsive and well-aligned detection behavior, where the anomaly score accurately rises and remains elevated almost exclusively within the ground-truth interval. Third, the sequence *Explosion010* (Fig. 2, lower-middle) showcases that, while the model reacts sharply at the beginning of the video, it fails to capture the core ground-truth interval, where the anomaly score remains low. Lastly, the sequence *Shoplifting033* (Fig. 2, bottom) exemplifies a high false-positive rate under nominal conditions. The anomaly score shows cyclic fluctuations, peaking aggressively in regions where no anomaly is present. This reveals that fine-grained anomalies are easily masked by background dynamics, causing false positives on routine scenes while missing covert events.

4.3 Cross-Dataset Generalisation

The objetive of this experiment is to quantify the domain gap between controlled-environment benchmarks and real-world surveillance. To this aim, three different VADMamba models, individually pretrained with the datasets Ped2, Avenue or ShanghaiTech are

evaluated directly on UCF-Crime without any retraining or fine-tuning. The results are shown in Table 3, which presents the results of this cross-dataset evaluation. The *UCF AUC* column reports the frame-level AUC obtained when evaluating each pretrained model directly on the UCF-Crime test set, whereas the *Orig. AUC* column denotes the baseline AUC achieved by the same architecture on its native test set. The *Gap* column quantifies the performance degradation, calculated as the difference between these two metrics.

Table 3: Cross-dataset evaluation on UCF-Crime.

Pretrained on	UCF AUC (%)	Orig. AUC (%)	Gap (%)
Ped2	52.27	98.45	-46.18
Avenue	53.11	91.49	-38.38
ShanghaiTech	53.96	77.03	-23.07

The results demonstrate that models trained on controlled-environment datasets perform poorly ($\sim$50%) on UCF-Crime, regardless of their original performance. The domain gap is inversely proportional to the diversity of the source dataset: Ped2, the most constrained benchmark with a single fixed camera and simple pedestrian anomalies, exhibits the largest gap (46.18 percent points), while ShanghaiTech, which includes 13 camera locations and more varied anomalies, shows the smallest gap (23.07 points). These findings highlight that high AUC values on controlled benchmarks do not guarantee effective performance in real-world surveillance and underscore the importance of evaluating VAD methods on datasets that reflect operational conditions.

Table 4: Frame-level AUC (%) on UCF-Crime. OCC methods train exclusively on normal videos without access to any anomalous data. US methods train on all videos without labels. WS methods use video-level anomaly annotations.

Paradigm	Method	Feature	AUC
OCC	Conv-AE (Hasan et al., 2016)	–	50.60
	Lu et al. (Lu et al., 2013)	C3D	65.51
	BODS (Zaheer et al., 2022)	I3D	68.26
	OGNet (Zaheer et al., 2021)	ResNeXt	69.47
	GODS (Zaheer et al., 2022)	I3D	70.46
	Our tests with config. $[1,1,1,1]$	Mamba	68.25
	Our tests with config. $[2,2,2,2]$	Mamba	69.90
	Our tests with config. $[3,3,3,3]$	Mamba	**72.23**
US	GCL (Zaheer et al., 2022)	ResNeXt	74.20
	LANP (Shi et al., 2024)	ResNeXt	**76.64**
WS	Sultani et al. (Sultani et al., 2018)	C3D	75.41
	RTFM (Tian et al., 2021)	I3D	84.30
	DMU (Zhou et al., 2023)	I3D	86.97
	VadCLIP (Wu et al., 2024)	CLIP	88.02
	π-VAD (Pu et al., 2025)	CLIP	**90.33**

Table 4 contextualises the results obtained in this work within the broader landscape of anomaly detection methods evaluated on UCF-Crime. Among unsupervised approaches, the best VADMamba configuration that we tested, $[3,3,3,3]$, achieving 72.23% AUC, surpasses OCC previous methods. Notably, we achieve this result while training the $[3,3,3,3]$ configuration on only 32.5% of the available normal data, using a fundamentally different architectural paradigm based on SSMs without requiring any pretrained video backbone. However, a substantial gap remains between unsupervised methods and the WS state of the art: methods such as RTFM (84.30% AUC) and DMU (86.97% AUC) benefit from video-level anomaly labels during training, which provide explicit guidance for learning discriminative anomaly representations. This comparison accross paradigms should be interpreted with care, as the supervision signals differ fundamentally, but it establishes the current positioning of unsupervised SSM-based methods and highlights the margin for improvement that future architectural innovations or semi-supervised extensions could address. It should be noted that US methods such as GCL (74.20% AUC) and LANP (76.64% AUC), which train on both normal and anomalous videos without labels, achieve higher AUC but operate under a less restrictive paradigm with access to anomalous training data.

4.4 Computational Efficiency

Table 5 reports the inference efficiency of the prediction-only configuration (Task 0) across the three depth levels. All three configurations share the same input format: $48 \times 256 \times 256$, corresponding to 16 concatenated RGB frames. The lightest model ($[1,1,1,1]$) achieves 285.5 samples/s (3.5 ms per sample), while even the deepest ($[3,3,3,3]$) maintains 98.9 samples/s. While surveillance cameras operate at 25–30 FPS, end-to-end evaluation on UCF-Crime, including data loading and score computation, yields 126.7, 92.2, and 71.2 FPS for depths $[1,1,1,1]$, $[2,2,2,2]$, and $[3,3,3,3]$, respectively, confirming that all three configurations are capable of real-time processing.

Table 5: Prediction-only inference efficiency across depth configurations. Throughput and latency measured per forward pass on RTX 5090 with input $48 \times 256 \times 256$.

Depth	Params (M)	FLOPs (G)	Throughput (samples/s)	Latency (ms)
$[1,1,1,1]$	14.84	2.80	285.5	3.5
$[2,2,2,2]$	28.21	5.13	148.4	6.7
$[3,3,3,3]$	41.57	8.61	98.9	10.1

Table 6 decomposes the full dual-task pipeline (Task 2) into its three components, reporting the cost that each contributes to processing one complete input sequence. Model 1 refers to the prediction branch

(Task 0), which receives 16 concatenated RGB frames as a $48 \times 256 \times 256$ tensor and predicts the subsequent frame. Model 2 refers to the reconstruction branch, which receives optical flow fields as a $2 \times 256 \times 256$ tensor and reconstructs them. FlowNet2SD is a pretrained optical flow estimator that computes the input for Model 2; it is called twice per sequence to compute the optical flow from frame $t-1$ to frame t, and from frame t to frame $t+1$. FlowNet2SD dominates the computational cost, accounting for 75.8% of the total FLOPs (16.96 G out of 22.37 G). In terms of latency, however, the three components contribute similarly (3.4–3.7 ms each). Eliminating the reconstruction branch removes both FlowNet2SD and Model 2 (Task 2), reducing the parameter count by 80% (from 74.98 M to 14.84 M), the computational cost by 87% (from 22.37 G to 2.80 G FLOPs), and the inference latency by 55% (from 7.7 to 3.5 ms per sample), while simultaneously improving detection performance (68.25% vs. 60.78% AUC).

Table 6: Full dual-task pipeline cost breakdown for depths $[1,1,1,1]$. All values represent the cost per input sequence processed through the complete pipeline.

Component	Params (M)	Calls	FLOPs (G)	Latency (ms)	FLOPs (%)
Model 1	14.84	1	2.80	3.5	12.5
FlowNet2SD	45.37	2	16.96	3.4	75.8
Model 2	14.77	1	2.61	3.7	11.7
Full pipeline	74.98	–	22.37	7.7	100

The analysis of training throughput (Table 7) further reveals the computational impact of FlowNet2SD. During training, the prediction phase processes 207 samples/s using 8.11 GB of GPU memory, while the reconstruction phase, which requires FlowNet2SD for optical flow computation, processes only 54 samples/s using 15.94 GB, representing a 3.85× slowdown and nearly double the memory consumption. This bottleneck is directly attributable to FlowNet2SD, which alone contains 45.37M parameters, more than both VADMamba models combined (29.61M).

Table 7: Training phase comparison for depths $[1,1,1,1]$.

Phase	Throughput (samples/s)	GPU Mem (GB)	Time/epoch (h)
Prediction	207	8.11	2.0
Reconstruction	54	15.94	7.6

Increasing model depth yields consistent improvements in peak AUC: $[1,1,1,1]$ achieves 68.25%, $[2,2,2,2]$ achieves 69.90% (+1.65 pp), and $[3,3,3,3]$ achieves 72.23% (+2.33 pp). The marginal gain increases with depth, suggesting that the performance ceiling on the UCF-Crime dataset with 260 training videos has not been reached, and that further capacity scaling could yield additional improvements if paired with a larger training set. However, this accuracy progression comes at the cost of 1.9× more parameters and 1.92× slower inference per depth increment, introducing a critical practical trade-off between detection performance and deployment efficiency.

These considerations suggest that, while the prediction-only VADMamba pipeline is architecturally viable for edge environments regarding parameter count and theoretical FLOPs, bridging the gap to real-time execution requires specialized advances in runtime compilation for SSM operators alongside systematic empirical benchmarking on target hardware platforms, a validation that constitutes a necessary direction for future work.

5 CONCLUSIONS AND FUTURE WORK

This work presents the first evaluation of the VADMamba framework on the UCF-Crime dataset, a large-scale real-world surveillance benchmark. The experimental results yield the next main conclusions.

First, the prediction branch alone establishes a new state-of-the-art among unsupervised methods on UCF-Crime, achieving 72.23% AUC with a depth configuration of $[3,3,3,3]$ and surpassing the previous best results. This result is obtained using only 32.5% of the available training data, demonstrating the high data efficiency of the VQ-MaU architecture. The prediction-only configuration outperforms the full dual-task pipeline (60.78% MiX AUC). In its lightest form $[1,1,1,1]$, it reduces the parameter count from 74.98M to 14.84M, increases inference speed from 129.4 to 285.5 throughput, and requires fewer FLOPs (2.80G vs. 22.37G). Even the best-performing $[3,3,3,3]$ configuration (41.57M parameters, 8.61G FLOPs, 98.9 throughput) remains substantially more efficient than the full pipeline while achieving the highest AUC.

Second, the cross-dataset transfer experiments reveal a substantial domain gap between controlled-environment benchmarks and real-world surveillance. Models pretrained on Ped2, Avenue, and ShanghaiTech achieve poor performance on UCF-Crime (52%–54% AUC), with degradations ranging from 23 to 46 pp. This finding highlights the limited representativeness of widely-used benchmarks for evaluating VAD systems intended for operational deployment.

Third, the depth scaling analysis reveals that increasing model capacity yields consistent improvements in peak AUC: +1.65 pp from $[1,1,1,1]$ to $[2,2,2,2]$, and +2.33 pp from $[2,2,2,2]$ to $[3,3,3,3]$,

with marginal gains that increase rather than diminish. However, all configurations exhibit overfitting after 20–30 epochs, with deeper models overfitting more aggressively (a 2.46 pp drop for $[3,3,3,3]$ vs. 0.12 pp for $[1,1,1,1]$), confirming that the performance ceiling is primarily constrained by training data diversity.

Fourth, while the prediction-only pipeline shows promising computational efficiency, physical validation on embedded platforms remains necessary to confirm its practical edge deployment capabilities.

Finally, a gap of approximately 3 pp remains between our best OCC result (72.23% AUC) and the WS baseline of Sultani et al. (75.41% AUC), which benefits from video-level anomaly labels during training.

Several directions for future work emerge from this study. First, exploring alternative temporal sequence lengths (t) and their impact on detection performance could reveal whether the current 16-frame context is optimal for the diverse anomaly types present in UCF-Crime. Second, to overcome the critical local storage limitations that forced the reduction of the training set, future efforts will be focused on in-memory video decoding schemes and temporal subsampling techniques. Instead of extracting and storing millions of uncompressed, redundant raw frames on disk, the original compressed video files can be maintained and dynamically decoded in RAM during training, or only one out of every N frames can be extracted. This drastic reduction in disk footprint will allow the training phase to be scaled to incorporate the full volume and semantic diversity of all 800 normal videos, mapping the definitive theoretical upper bounds of the proposed architecture without exceeding the strict storage budget. Third, the demonstrated inefficiency of FlowNet2SD motivates the exploration of flow-free architectures, such as the recently proposed VADMamba++, which eliminates optical flow computation entirely through a Gray-to-RGB paradigm (Lyu et al., 2026). (Note that the VADMamba++ architecture was published while this paper was being written, so it could not be included in the presented analysis, but will be analyzed in future work). Finally, validating the prediction-only pipeline in a real-time deployment scenario with live camera streams would provide practical evidence of the system's applicability in operational surveillance environments.

ACKNOWLEDGEMENTS

The Ministry of Science, Innovation and Universities (Spain) has funded this work through PREP2023-001630 (E. González-Amorós). This work is part of the project PID2023-149575OB-I00, funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE. It is also part of the project CIPROM/2024/8, funded by Generalitat Valenciana, Conselleria de Educación, Cultura, Universidades y Empleo (program PROMETEO 2025).

REFERENCES

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and Hengel, A. v. d. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., and Davis, L. S. (2016). Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470.

Li, W., Mahadevan, V., and Vasconcelos, N. (2013). Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32.

Liu, W., Luo, W., Lian, D., and Gao, S. (2018). Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., and Liu, Y. (2024). Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063.

Lu, C., Shi, J., and Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727.

Luo, W., Liu, W., and Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349.

Lyu, J., Zhao, M., Hu, J., Chen, Y., Du, S., and Shi, C. (2026). Vadmamba++: Efficient video anomaly detection via hybrid modeling in grayscale space.

Lyu, J., Zhao, M., Hu, J., Huang, X., Chen, Y., and Du, S. (2025). Vadmamba: Exploring state space models for fast video anomaly detection. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Nawaratne, R., Alahakoon, D., De Silva, D., and Yu, X. (2019). Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics*, 16(1):393–402.

Park, H., Noh, J., and Ham, B. (2020). Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381.

Pu, Z. et al. (2025). π-vad: Prompt-driven inference for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shi, H., Wang, L., Zhou, S., Hua, G., and Tang, W. (2024). Learning anomalies with normality prior for unsupervised video anomaly detection. In *European Conference on Computer Vision (ECCV)*, pages 163–180.

Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., and Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986.

Wu, P., Zhou, X., Pang, G., Sun, Y., Liu, J., Wang, P., and Zhang, Y. (2024). Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6074–6082.

Zaheer, M. Z., Mahmood, A., Khan, M. H., Segu, M., Yu, F., and Lee, S.-I. (2021). Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing*, 15(12):2826–2839.

Zaheer, M. Z., Mahmood, A., Khan, M. H., Segu, M., Yu, F., and Lee, S.-I. (2022). Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14744–14754.

Zhou, H., Yu, J., and Yang, W. (2023). Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3769–3777.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., and Wang, X. (2024). Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.