POSTER SESSION BOOKLET

http://www.dmi.unict.it/icvss

University of Catania - University of Cambridge

International Computer Vision Summer School 2025

Computer Vision for Spatial Intelligence

Sicily, 6 - 12 July 2025

International Computer Vision Summer School

Computer Vision is the science and technology of making machines that see. It is concerned with the theory, design and implementation of algorithms that can automatically process visual data to recognize objects, track and recover their shape and spatial layout.

The International Computer Vision Summer School - ICVSS was established in 2007 to provide both an objective and clear overview and an in-depth analysis of the state-of-the-art research in Computer Vision. The courses are delivered by world renowned experts in the field, from both academia and industry, and cover both theoretical and practical aspects of real Computer Vision problems.

The school is organized every year by University of Cambridge (Computer Vision and Robotics Group) and University of Catania (Image Processing Lab). The general entry point for past and future ICVSS editions is:

http://www.dmi.unict.it/icvss

ICVSS Poster Session

The International Computer Vision Summer School is especially aimed to provide a stimulating space for young researchers and Ph.D. Students. Participants have the possibility to present the results of their research, and to interact with their scientific peers, in a friendly and constructive environment.

This booklet contains the abstract of the posters accepted to ICVSS 2025.

Best Presentation Prize A subset of the submitted posters will be selected by the school committee for short oral presentation. A best presentation prize will be given to the best presentations selected by the school committee.

Scholarship A scholarship will be awarded to the best PhD student attending the school. The decision is made by the School Committee at the time of the School, taking into account candidates' CV, poster and oral presentation.

Sicily, June 2025

Roberto Cipolla
Sebastiano Battiato
Giovanni Maria Farinella

List of Posters [1]

1. LOOK&LEARN: BRIDGING PERCEPTION AND GROUNDING GAP IN VISION-LANGUAGE MODELS Abdelrahman E., Li Y., Iyer S. S., Zhao N., Shechtman E., Singh K. K., Elhoseiny M
2. COMMONLY INTERESTING IMAGES Abdullahu F., Grabner H.
3. MEASURING RELIABILITY AND GENERALIZATION BEYOND IMAGE CLASSIFICATION Agnihotri S., Keuper M.
4. CRACK SEGMENTATION FOR STRUCTURAL HEALTH MONITORING Ahmad T.
5. PRIVACY-PRESERVING PAIN ESTIMATION VIA FACIAL LANDMARK GRAPHS Fatemah Alhamdoosh, Pietro Pala, Stefano Berretti
6. TOWARDS REALISTIC TEST-TIME ADAPTATION: A TRACKLET-BASED BENCHMARK Alhuwaider S., Alfarra M., Perez J., Ramazanova M., Ghanem B.
7. UNCERTAINTY-AWARE KNOWLEDGE DISTILLATION FOR EFFICIENT 6DOF POSE ESTIMATION ALI OUSALAH N., KACEM A., GHORBEL E., KOUMANDAKIS M., AOUADA D.
8. PDISCOFORMER: RELAXING PART DISCOVERY CONSTRAINTS WITH VISION TRANSFORMERS Ananthu Aniraj, Cassio F. Dantas, Dino Ienco, Diego Marcos
9. B-COSIFICATION: TRANSFORMING NEURAL NETWORKS TO BE INHERENTLY INTERPRETABLE Arya S., Rao S., Böhle M., Schiele B.
10. TEXT-DRIVEN 3D HAND MOTION GENERATION Bensabath L, Petrovich M, Varol G
11. HIERASURG: WORLD MODELS FOR SURGICAL DATA SCIENCE Biagini D., Farshad A., Navab N.

[1] Posters are ordered by surname of the speaker. Each poster is identified by a number.

12. WHAT HAPPENS NEXT? ANTICIPATING FUTURE MOTION BY GENERATING POINT TRAJECTORIES Boduljak G., Karazija L., Laina I., Rupprecht C., Vedaldi A.

13. 3D FACE RECONSTRUCTION FROM RADAR IMAGES Braeutigam V., Wirth V., Ullmann I., Schuessler C., Vossiek M., Berking M., Egger B.

14. CROSS-SPECTRAL GATED-RGB STEREO DEPTH ESTIMATION Brucker S., Walz S., Bijelic M., Heide F.

15. SUPEREVENT: CROSS-MODAL LEARNING OF EVENT-BASED KEYPOINT DETECTION Burkhardt Y., Schaefer S., Leutenegger S.

16. CONDITIONAL DDPMS FOR LOW-LIGHT ENHANCEMENT Cabassa G.

17. RECURRENCE-ENHANCED TRANSFORMERS FOR ROBUST MULTIMODAL DOCUMENT RETRIEVAL Caffagni D., Sarto S., Cornia M., Baraldi L., Cucchiara R.

18. INCREMENTAL AND DECREMENTAL CONTINUAL LEARNING FOR PRIVACY-PRESERVING VIDEO RECOGNITION Caselli L., Magistri S., Bianconcini T., Benericetti A., de Andrade DC., and Bagdanov AD.

19. SAEMNESIA: ERASING CONCEPTS IN DIFFUSION MODELS WITH SPARSE AUTOENCODERS Cassano E., Renzulli R., Grangetto M.

20. GROCO: GROUND CONSTRAINT FOR METRIC SELF-SUPERVISED MONOCULAR DEPTH Aurélien Cecille, Stefan Duffner, Franck Davoine, Thibault Neveu, Rémi Agier

21. FORCE-AWARE 3D CONTACT MODELING FOR STABLE GRASP GENERATION Chen Z., Zhang Z., Cheng Y., Leonardis A., Chang H.

22. SPLATFORMER: POINT TRANSFORMER FOR ROBUST 3D GAUSSIAN SPLATTING (ICLR 2025 SPOTLIGHT) Chen Yutong, Mihajlovic Marko, Chen Xiyi, Wang Yiming, Prokudin Sergey, Tang Siyu

23. BRIDGING DOMAIN GAP IN 6-DOF POSE ESTIMATION VIA CONTRASTIVE ALIGNMENT AND PSEUDO-LABEL REFINEMENT Chenni N., Rathinam A., Aouada D.

24. ARTIFICIAL INTELLIGENCE FOR ROBOTIC SURGERY Chiesa G., Renzulli R., Grangetto M.

25. GRAMIAN MULTIMODAL LEARNING AND ALIGNMENT Cicchetti Giordano

26. SHEDDING LIGHT ON DEPTH: EXPLAINABILITY ASSESSMENT IN MONOCULAR DEPTH ESTIMATION Cirillo L., Schiavella C., Papa L., Russo P., Amerini I.

27. JOINT OPTIMIZATION OF FILTER ATTACHMENTS AND SUPER-RESOLUTION FOR SPECTRAL IMAGING WITH STEREO RGB CAMERAS Cogo L., Buzzelli M., Bianco S., Schettini R.

28. MULTIMODAL SAM-ADAPTER FOR SEMANTIC SEGMENTATION Curti Iacopo, Zama Ramirez Pierluigi, Petrelli Alioscia, Di Stefano Luigi

29. JUST DANCE WITH π ! A POLY-MODAL INDUCTOR FOR WEAKLY-SUPERVISED VIDEO ANOMALY DETECTION Mahji, S., D'Amicantonio, G., Dantcheva, A., Kong, Q., Garattoni, L., Francesca, G., Bondarev, E., Bremond, F.,

30. NESYLAD: A NEURO-SYMBOLIC APPROACH FOR UNSUPERVISED LOGICAL ANOMALY DETECTION Dahmardeh M., Setti F.

31. SUPERVISING 3D TALKING HEAD AVATARS WITH ANALYSIS-BY-AUDIO-SYNTHESIS Daněček R., Schmitt C., Polikovsky S., Black M.

32. VIDEO OBJECT DETECTION IN MARITIME SCENARIOS Denk F., Moser D., Rankl C., Sablatnig R.

33. RESOLUTION WHERE IT COUNTS: HASH-BASED GPU-ACCELERATED 3D RECONSTRUCTION VIA VARIANCE-ADAPTIVE VOXEL GRIDS De Rebotti L., Giacomini E., Grisetti G., Di Giammarino L.

34. PARAMETRIC SHAPE MODELS FOR VESSELS LEARNED FROM SEGMENTATIONS VIA A DIFFERENTIABLE VOXELIZATION LOSS DIMA Alina, SHIT Suprosanna, QIU Huaqi, HOLLAND Robbie, MUELLER Tamara, MUSIO Fabio, YANG Kaiyuan, MENZE Bjoern, BRAREN Rickmer, MAKOWSKI Marcus, RUECKERT Daniel

35. MACHINE LEARNING FOR HUMAN ANALYSIS AND BIOMETRICS Di Domenico N.

36. SELF-SUPERVISED PRE-TRAINING WITH DIFFUSION MODEL FOR FEW-SHOT LANDMARK DETECTION IN X-RAY IMAGES Di Via R., Odone F., Pastore V. P.

37. SAIL: SELF-SUPERVISED ALBEDO ESTIMATION FROM REAL IMAGES WITH A LATENT DIFFUSION MODEL Djeghim.H, Piasco.N, Roldao.L, Bennehar.M, Tsishkou.D, Loscos.C, Sidibé.D

38. TOWARDS ROBUST MULTIMODAL OUT-OF-DISTRIBUTION GENERALIZATION AND DETECTION FOR REAL-WORLD SYSTEMS Dong H., Chatzi E., Fink O.

39. METPOSE: TESTING POSE ESTIMATION ON UNLABELLED DATA Duran M.

40. INTERACTVLM: 3D INTERACTION REASONING FROM 2D FOUNDATIONAL MODEL Dwivedi S.K., Antić D., Tripathi S., Taheri O., Schmid C., Black M.J., Tzionas D.

41. HD-EPIC: A HIGHLY-DETAILED EGOCENTRIC VIDEO DATASET Perrett T, Darkhalil A, Sinha S, Emara O, Pollard S, Parida K, Liu K, Gatti P, Bansal S, Flanagan K, Chalk J, Zhu Z, Guerrier R, Abdelazim F, Zhu B, Moltisanti D, Wray M, Doughty H, Damen D

42. VIDEO UNLEARNING VIA LOW-RANK REFUSAL VECTOR Facchiano S., Saravalle S., Migliarini M., De Matteis E., Sampieri A., Pilzer A., Rodolà E., Spinelli I., Franco L., Galasso F.

43. INTRAOPERATIVE REGISTRATION BY CROSS-MODAL INVERSE NEURAL RENDERING Fehrentz M., Azampour M., Dorent R., Rasheed H., Galvin C., Golby A., Wells W., Frisken S., Navab N., Haouchine N.

44. G-SOLVER: GAUSSIAN BELIEF PROPAGATION AND GAUSSIAN PROCESSES FOR CONTINUOUS-TIME SLAM Ceriola D., Ferrari S., Di Giammarino L., Brizi L., Grisetti G.

45. PATCH SIZE CURRICULUM IN 3D PATCH-BASED SEGMENTATION Fischer S., Kiechle J., Peeken J. Schnabel J.

46. LEARN YOUR SCALES Forghani F, Yu J, Aumentado-Armstrong T, Derpanis K, Brubaker M

47. ONLINE EPISODIC MEMORY VISUAL QUERY LOCALIZATION WITH EGOCENTRIC STREAMING OBJECT MEMORY Zaira Manigrasso, Matteo Dunnhofer, Antonino Furnari, Moritz Nottebaum, Antonio Finocchiaro, Davide Marana, Rosario Forte, Giovanni Maria Farinella, Christian Micheloni

48. DENSE 3D MAPPING FOR SPATIAL AI Fry N., Kelly P., Davison A.

49. SKELETON-BASED ACTION RECOGNITION FOR BIOMECHANICAL RISK ASSESSMENT Gennarelli I., Ranavolo A., Micheloni C., Martinel N.

50. PIXEL3DMM: VERSATILE SCREEN-SPACE PRIORS FOR SINGLE-IMAGE 3D FACE RECONSTRUCTION Giebenhain Simon, Kirschstein Tobias, Rünz Martin, Agapito Lourdes, Nießner Matthias

51. TOWARDS A PERCEPTUAL EVALUATION FRAMEWORK FOR LIGHTING ESTIMATION Giroux J., Dastjerdi M., Hold-Geoffroy Y., Vazquez-Corral J., Lalonde J.-F.

52. UNDERLOC: IMAGE BASED RELOCALIZATION AND ALIGNMENT FOR DYNAMIC UNDERWATER ENVIRONMENTS Gorry B., Fischer T., Milford M., Fontan A.

53. EVENT-BASED PHOTOMETRIC BUNDLE ADJUSTMENT Guo S., Gallego G.

54. VID2AVATAR-PRO: AUTHENTIC AVATAR FROM VIDEOS IN THE WILD VIA UNIVERSAL PRIOR Guo C., Li J., Kant Y., Sheikh Y., Saito S., Cao C.

55. ETAP: EVENT-BASED TRACKING OF ANY POINT Hamann F., Gehrig D., Febryanto F., Daniilidis K., Gallego G.

56. THE INVISIBLE EGOHAND: 3D HAND FORECASTING THROUGH EGOBODY POSE ESTIMATION Hatano M., Zhu Z., Saito H., Damen D.

57. LIMO: LIFELIKE HUMAN MOTION GENERATION WITH CONTINUOUS-SPACE GENERATIVE MODELS He Y., Tiwari G., Zhang X., Bora P., Birdal T., Lenssen J., Pons-Moll G.

58. ADVANCING PERSONAL HUMAN-CENTRIC AI WITH GENERATIVE MODELING Ho H., Kaufman M., Zhang L., Hilliges O.

59. NEURAL RENDERING FOR SENSOR ADAPTATION IN 3D OBJECT DETECTION Embacher F., Holtz D., Uhrig J., Cordts M., Enzweiler M.

60. MONOCULAR CAMERA-BASED SIDEWALK: WIDTH ESTIMATION FOR WHEELCHAIR ACCESSIBILITY Houshyar Yazdian S.H., Jacquet W., Stiens J

61. LITEREALITY: GRAPHIC-READY 3D SCENE RECONSTRUCTION FROM RGB-D SCANS Huang Z., Wu X., Zhong F., Zhao H., Niessner M., Lasenby J.

62. HOW TO PREDICT SOCIO-ECONOMIC DEVELOPMENT FROM SPACE? Janisiów Ł., Wójcik P.

63. ADVANCING FOREST TYPOLOGY FROM SPACE Jiang Y., Neumann M.

64. GEO4D: LEVERAGING VIDEO GENERATORS FOR GEOMETRIC 4D SCENE RECONSTRUCTION Jiang Zeren, Zheng Chuanxia, Laina Iro, Larlus Diane, Vedaldi Andrea

65. SELF-SUPERVISED COLLABORATIVE DISTILLATION FOR LIGHTNING-ROBUST AND 3D-AWARE 2D REPRESENTATIONS Jo W., Ha H., Kim J.-Y., Jeong H., Oh T.-H.

66. ADVANCING GENERALIZABILITY AND FAIRNESS IN BREAST CANCER: MAMA-MIA CHALLENGE Garrucho Lidia, Joshi Smriti, Kushibar Kaisar, Bobowicz Maciej, Bargalló Xavier, Jaruševičius Paulius, Lekadir Karim

67. UNFOLDING CLOTH: NEURAL DEFORMATION FIELDS FOR SIMULATION AND MONOCULAR TRACKING Kairanda N., Habermann M., Naik S., Theobalt C., Golyanik V.

68. PRIMEDEPTH: EFFICIENT MONOCULAR DEPTH ESTIMATION WITH A STABLE DIFFUSION PREIMAGE Zavadski D., Kalšan D., Rother C.

69. SELF-SUPERVISED PRETRAINING FOR FINE-GRAINED PLANKTON RECOGNITION Kareinen J., Eerola T., Kraft K., Lensu L., Suikkanen S., Kälviäinen H.

70. SEED4D: A SYNTHETIC EGO-EXO DYNAMIC 4D DATA GENERATOR, DRIVING DATASET AND BENCHMARK Kästingschäfer M., Gieruc T., Bernhard S., Campbell D., Insafutdinov E., Najafli E., Brox T.

71. DUALPM: DUAL POSED-CANONICAL POINT MAPS FOR 3D SHAPE AND POSE RECONSTRUCTION Kaye, B., Jakab, T., Wu, S., Rupprecht, C., Vedaldi, A.

72. INCORPORATING PROPERTIES OF HUMAN VISION INTO IMAGE GENERATION Kergaßner S., Tariq T., Didyk P.

73. YOUR VIT IS SECRETLY AN IMAGE SEGMENTATION MODEL Kerssies T., Cavagnero N., Hermans A., Norouzi N., Averta G., Leibe B., Dubbelman G., de Geus D.

74. KAIROSAD: A SAM-BASED MODEL FOR INDUSTRIAL ANOMALY DETECTION ON EMBEDDED DEVICES Khan U., Fummi F., Capogrosso L

75. DISENTANGLING MODALITY RELATIONS: A TWO-STAGE GRAPH FOR EMOTION RECOGNITION Khan Mohammad Mohammed Rahman Sherif., Kumar Swagat., Behera Ardhendu

76. MEDICAL MULTI-VIEW GNN: ADVANCING TUMOR MALIGNANCY PREDICTION USING SPATIAL-AWARE DINOV2 REPRESENTATIONS Kiechle J., Fischer S.M., Peeken J.C., Schnabel J.A.

77. ON-SENSOR OPTICAL FLOW FOR ALWAYS-ON ROBOT VISION Kim S., Kelly P., Davison A.

78. UNI-DVPS: UNIFIED MODEL FOR DEPTH-AWARE VIDEO PANOPTIC SEGMENTATION Ji-Yeon K., Hyun-Bin O., Byung-Ki K., Kim D., Kwon Y., Oh T.-H.

79. NERSEMBLE: MULTI-VIEW RADIANCE FIELD RECONSTRUCTION OF HUMAN HEADS KIRSCHSTEIN T., QIAN S., GIEBENHAIN S., WALTER T., NIESSNER M.

80. TACKLING DATA CHALLENGES IN DEEP LEARNING FOR ELECTRON MICROSCOPY Kniesel H.

81. TOWARDS AUTONOMOUS MULTI-ROBOT EXPLORATION IN UNSTRUCTURED ENVIRONMENTS Lasheras Hernández B., Giubilato R., Schuster M., Triebel R., Civera J.

82. ENHANCING THE OLDEST ICE CLIMATE SIGNALS THROUGH SUPER-RESOLUTION IMAGING TECHNIQUES Latif Hasan, Larkman Piers, Bohleber Pascal, Vascon Sebastiano

83. ENIGMA-360: A MULTI-VIEW DATASET FOR HUMAN BEHAVIOR UNDERSTANDING IN INDUSTRIAL SCENARIOS Ragusa F., Leonardi R., Mazzamuto M., Di Mauro D., Quattrocchi C., Passanisi A., D'Ambra I., Furnari A., Farinella G.M.

84. PROBABILISTIC CONTRASTIVE LEARNING VIA REGULARIZED VON MISES-FISHER DISTRIBUTIONS Li H. B., Ouyang C., Amiranashvili T., Rosen M., Menze B., Iglesias J. E.

85. UNIMOTION: UNIFYING 3D HUMAN MOTION SYNTHESIS AND UNDERSTANDING Li C., Chibane J., He Y., Pearl N., Geiger A., Pons-Moll G.

86. MEDBRIDGE: BRIDGING FOUNDATION VISION-LANGUAGE MODELS TO MEDICAL IMAGE DIAGNOSIS Li Yitong, Ghahremani Morteza, Wachinger Christian

87. FROG: FIELD ROBOTICS GROUP FOR REAL-WORLD 3D PERCEPTION SYSTEM Li W., Fusaro D., Mosco S., Pretto A.

88. GCE-POSE: GLOBAL CONTEXT ENHANCEMENT FOR CATEGORY-LEVEL OBJECT POSE ESTIMATION Li Weihang., Xu Hongli., Junwen Huang., Jung HyunJun., Yu Peter KT., Navab Nassir., Busam Benjamin.

89. MULTIMODALSTUDIO: A DATASET AND FRAMEWORK FOR MULTIMODAL NEURAL RENDERING Lincetto F., Agresti G., Rossi M., Zanuttigh P.

90. INVERSE VIRTUAL TRY-ON: GENERATING MULTI-CATEGORY PRODUCT-STYLE IMAGES FROM CLOTHED INDIVIDUALS Lobba D., Sanguigni F., Ren B., Cornia M., Cucchiara R., Sebe N.

91. ALLIGAT0R: PRE-TRAINING THROUGH COVISIBILITY SEGMENTATION FOR RELATIVE CAMERA POSE REGRESSION Thibaut L., Guillaume B., Vincent L.

92. DARK NOISE DIFFUISION: NOISE SYNTHESIS FOR LOW-LIGHT IMAGE DENOISING Liying Lu, Raphaël Achddou, Sabine Süsstrunk

93. SPEQ: OFFLINE STABILIZATION PHASES FOR EFFICIENT Q-LEARNING IN HIGH UPDATE-TO-DATA RATIO REINFORCEMENT LEARNING Romeo C.*, Macaluso G.*, Sestini A., Bagdanov A. D.

94. HM3: HIERARCHICAL MODELING OF MULTIMEDIA METAVERSES ON 10000 THEMATIC MUSEUMS VIA THEME-AWARE CONTRASTIVE LOSS FUNCTION Macrì G., Bazzana L., Falcon A., Serra G.

95. METRIC-SEMANTIC 3D SCENE UNDERSTANDING Maggio D.

96. FRED: THE FLORENCE RGB-EVENT DRONE DATASET Magrini G., Marini N., Becattini F., Berlincioni L., Biondi N., Pala P., Del Bimbo A.

97. GROUND TRUTH-FREE FINE-TUNING HUMAN MOTION DIFFUSION MODELS WITH REINFORCEMENT LEARNING Mandelli, Macaluso, Bicchierai

98. 4DEFORM: NEURAL SURFACE DEFORMATION FOR ROBUST SHAPE INTERPOLATION Sang L., Canfas Z., Cao D., Marin R., Bernard F., Cremers D.

99. CONTEXT-AWARE EVENT DETECTION, VIDEO UNDESTANDING WITH ROBUSTNESS TO CONTEXTUAL BIAS Mattia Marseglia

100. DOCWAVEDIFF Marulli Matteo, Marco Bertini

101. GAZING INTO MISSTEPS: LEVERAGING EYE-GAZE FOR UNSUPERVISED MISTAKE DETECTION IN EGOCENTRIC VIDEOS OF SKILLED HUMAN ACTIVITIES Mazzamuto M., Furnari A., Sato Y., Farinella G.M.

102. TRAIN TILL YOU DROP: TOWARDS STABLE AND ROBUST SOURCE-FREE UNSUPERVISED 3D (AND IMAGE) DOMAIN ADAPTATION Michele Björn, Boulch Alexandre, Vu Tuan-Hung, Puy Gilles, Marlet Renaud, Courty Nicolas

103. HUMAN MOTION UNLEARNING De Matteis E., Migliarini M., Sampieri A., Spinelli I., Galasso F.

104. PATHOGEN CLASSIFICATION IN HYPERSPECTRAL DATA CUBES Mikulić M., Štajduhar I.

105. REALFRED: AN EMBODIED INSTRUCTION FOLLOWING BENCHMARK IN PHOTO-REALISTIC ENVIRONMENT Kim T.*, Min C.*, Kim B., Kim J., Jeung W., Choi J.

106. MONOCULAR DEPTH ESTIMATION IN ADVERSE CONDITIONS Gasperini S.*, Morbitzer N.*, Jung H., Navab N., Tombari F.

107. EXAMINING GRANULARITY FOR PRIVACY PROTECTION Murrugarra-LLerena Jeffri., Niu Haoran., Barber K. Suzanne., Daumé III Hal., Trista Cao Yang., Cascante-Bonilla Paola.

108. NARRATIVEBRIDGE: ENHANCING VIDEO CAPTIONING WITH CAUSAL-TEMPORAL NARRATIVE Nadeem, Asmar, Sardari Faegheh, Dawes, Robert, Husain Sameed Syed, Hilton, Adrian, Mustafa, Armin

109. OCCLUSION AND CONFUSION: EVENT DETECTION AND PERFORMANCE ANALYSIS IN AMATEUR RUGBY FOOTAGE Ní Dheoráin C.

110. AVHBENCH: A CROSS-MODAL HALLUCINATION BENCHMARK FOR AUDIO-VISUAL LARGE LANGUAGE MODELS Sung-Bin K., Hyun-Bin O., Jung-Mok L., Senocak A., Chung J.S., Oh T.-H.

111. SAMFUSION: SENSOR-ADAPTIVE MULTIMODAL FUSION FOR 3D OBJECT DETECTION IN ADVERSE WEATHER Palladin E., Dietze R., Narayanan P., Bijelic M., Heide F.

112. LENGTH-AWARE MOTION SYNTHESIS VIA LATENT DIFFUSION Sampieri A., Palma A., Spinelli I., Galasso F.

113. TOWARDS EFFICIENT AND GENERIC STRUCTURE-FROM-MOTION Linfei Pan

114. FLYSEARCH: EXPLORING HOW VISION-LANGUAGE MODELS EXPLORE Pardyl, A., Matuszek, D., Przebieracz, M., Cygan, M., Zieliński, B., Wołczyk, M.

115. STYLE-EDITOR: TEXT-DRIVEN OBJECT-CENTRIC STYLE EDITING Park J., Gim J., Lee K., Lee S., Im S.

116. BEYOND THE LIPS: ROBUST SPEAKER DETECTION VIA DISENTANGLED LATENT REPRESENTATIONS Park J., Hong J., Kwon J

117. HIERO: UNDERSTANDING THE HIERARCHY OF HUMAN BEHAVIOR ENHANCES REASONING ON EGOCENTRIC VIDEOS Peirone S. A., Pistilli F., Averta G.

118. DEEP VISUAL ODOMETRY WITH EVENTS AND FRAMES Pellerito R., Cannici M., Gehrig D., Belhadj J., Dubois-Matra O., Casasco M., Scaramuzza D.

119. ADAPTING VISION TRANSFORMERS TO ULTRA-HIGH RESOLUTION SEMANTIC SEGMENTATION WITH RELAY TOKENS Perron Y., Sydorov V., Pottier C., Landrieu L.

120. PRIVILEGED INFORMATION AND MULTIPLE SCLEROSIS LESION SEGMENTATION Pignedoli V., Moro M., Noceti N., Odone F.

121. RENDBEV: SEMANTIC PERSPECTIVE VIEW RENDERING AS SUPERVISION FOR BIRD'S EYE VIEW SEGMENTATION Pineiro Monteagudo, H., Taccari L., Pjetri, A., Sambo, F. and Salti, S.

122. HYPERBOLIC SAFETY-AWARE VISION-LANGUAGE MODELS Poppi T., Kasarla T., Mettes P., Baraldi L., Cucchiara R.

123. PROBPOSE A PROBABILISTIC APPROACH TO 2D HUMAN POSE ESTIMATION Purkrabek M., Matas J.

124. WEATHEREDIT: CONTROLLABLE WEATHER EDITING WITH 4D GAUSSIAN FIELD Qian C.*, Li W.†, Guo Y., Markkula G.

125. TEST-TIME ADAPTATION FOR COMBATING MISSING MODALITIES IN EGOCENTRIC VIDEOS Ramazanova M., Pardo A., Ghanem B., Alfarra M.

126. ART2MUS: BRIDGING VISUAL ARTS AND MUSIC THROUGH CROSSMODAL GENERATION Rinaldi I., Fanelli N., Castellano G., Vessio G.

127. SHOW OR TELL? A BENCHMARK TO EVALUATE VISUAL AND TEXTUAL PROMPTS IN SEMANTIC SEGMENTATION Rosi G., Cermelli F.

128. TAKUNET: AN ENERGY-EFFICIENT CNN FOR REAL-TIME INFERENCE ON EMBEDDED UAV SYSTEMS IN EMERGENCY RESPONSE SCENARIOS Rossi D., Borghi G., Vezzani R.

129. MAMBA-ST: STATE SPACE MODEL FOR EFFICIENT STYLE TRANSFER Botti F., Ergasti A., Rossi L., Fontanini T., Ferrari C., Bertozzi M., Prati A.

130. NEURALATEX: A MACHINE LEARNING LIBRARY WRITTEN IN PURE LATEX Gardner J., Rowan W., Smith W.

131. LAM3D: LEVERAGING ATTENTION FOR MONOCULAR 3D OBJECT DETECTION Diana-Alexandra S., Leandro Di B., Yangxintong L., Florin O., Adrian M.

132. EXPLORATION-DRIVEN GENERATIVE INTERACTIVE ENVIRONMENTS Savov N., Kazemi N., Mahdi M., Paudel D.P., Wang X., Gool L.V.

133. EFFICIENT ATTENTION VISION TRANSFORMERS FOR MONOCULAR DEPTH ESTIMATION ON RESOURCE-LIMITED HARDWARE Schiavella C., Cirillo L., Papa L., Russo P., Amerini I.

134. IT'S A (BLIND) MATCH! TOWARDS VISION–LANGUAGE CORRESPONDENCE WITHOUT PARALLEL DATA Schnaus Dominik, Araslanov Nikita, Cremers Daniel

135. POEM: PRECISE OBJECT-LEVEL EDITING VIA MLLM CONTROL Schouten M., Kaya M.O., Belongie S., Papadopoulos D. P.

136. OBJECTSPLAT: GENERALIZABLE OBJECT-CENTRIC 3D GAUSSIAN SPLATTING Schröppel P., Wewer C., Ilg E., Lenssen J.E.

137. TASK GRAPH MAXIMUM LIKELIHOOD ESTIMATION FOR PROCEDURAL ACTIVITY UNDERSTANDING IN EGOCENTRIC VIDEOS Seminara L., Farinella G. M., Furnari A.

138. BODY MEASUREMENT AND SURFACE RECONSTRUCTION IN MICROWAVE IMAGING Miriam S.

139. CANONICALFUSION: GENERATING DRIVABLE 3D HUMAN AVATARS FROM MULTIPLE IMAGES Shin J., Lee J., Lee S., Park M., Kang J., Yoon J., Jeon H.

140. INTERWEAVING INSIGHTS: HIGH-ORDER FEATURE INTERACTION FOR FINE-GRAINED VISUAL RECOGNITION (I2HOFI) Sikdar A., Liu Y., Kedarisetty S., Zhao Y., Ahmed A., Behera A.,

141. PRADA: PROJECTIVE RADIAL DISTORTION AVERAGING Sinitsyn D., Härenstam-Nielsen L., Cremers D.

142. TCC-DET: TEMPORARILY CONSISTENT CUES FOR WEAKLY-SUPERVISED 3D DETECTION Skvrna Jan, Neumann Lukas

143. CROSS-MODAL SEMANTIC GROUNDING EXPLOITING CONTEXT LEARNING FOR REAL-TIME VISUAL UNDERSTANDING Spingola Camilla

144. SEGMENTATION UNDER LOW-DATA CONSTRAINTS FOR NICHE DOMAINS Sterzinger R., Sablatnig R.

145. WORLD MODEL PREDICTIVE CONTROL FOR INTERPRETABLE AUTONOMOUS DRIVING Sun Jiangxin, Xue Feng, Long Teng, Sebe Nicu

146. BILLBOARDS SPLATTING (BBSPLAT): LEARNABLE TEXTURED PRIMITIVES FOR NOVEL VIEW SYNTHESIS Svitov D., Morerio P., Agapito L., Del Bue A.

147. DIFFFNO: DIFFUSION FOURIER NEURAL OPERATOR Xiaoyi Liu, Hao Tang

148. RGB AND IR FUSION FOR MULTIMODAL AERIAL TARGET DETECTION Tavaris Denis, De Zan Alberto, Ivan Scagnetto, Gian Luca Foresti

149. EGO-R1: CHAIN-OF-TOOL-THOUGHT FOR ULTRA-LONG EGOCENTRIC VIDEO REASONING Tian Shulin* extsuperscript1, extsuperscript2, Wang Ruiqi extsuperscript1, extsuperscript3, Guo Hongming extsuperscript4, Wu Penghao extsuperscript1, Dong Yuhao extsuperscript1, Wang Xiuying extsuperscript1, Yang Jingkang extsuperscript1, Zhang Hao extsuperscript3, Zhu Hongyuan extsuperscript2, Liu Ziwei extsuperscript1

150. OPTIMIZING PROSTHETIC VISION USING VISION TRANSFORMERS Tomas-Barba, J, Perez-Yus, A., Bermudez-Cameo, J.

151. ITERATIVE SELF-SUPERVISION FOR SPARSE VIEW NERF & GAUSSIAN SPLATTING Felix Tristram, Stefano Gasperini, Nassir Navab, Federico Tombari

152. ACTMERGE: TASK-AWARE ACTIVATION INFORMED MODEL MERGING Verasani M.

153. UNIBEV: MULTI-MODAL 3D OBJECT DETECTION WITH UNIFORM BEV ENCODERS FOR ROBUSTNESS AGAINST MISSING SENSOR MODALITIES Wang, Shiming; Caesar, Holger; Nan, Liangliang; Kooij, Julian F. P.

154. 3D RECONSTRUCTION WITH SPATIAL MEMORY Wang H., Agapito L.

155. LEARNING SPATIAL REPRESENTATINOS FOR EMBODIED PERCEPTION IN 3D WORLDS Weijler L.

156. TOWARDS SAFER AUTONOMOUS SYSTEMS: UNCERTAINTY QUANTIFICATION FOR REGRESSION Xiong Z., Johnander J., Forssén P.-E.

157. DEMO: DENSE MOTION CAPTIONING FOR COMPLEX HUMAN MOTIONS Shiyao Xu, Benedetta Liberatori, Gul Varol, Paolo Rota

158. 3D-MOOD: LIFTING 2D TO 3D FOR MONOCULAR OPEN-SET OBJECT DETECTION Yang Y.H., Pollefeys M.

159. SMALLGS GAUSSIAN SPLATTING-BASED CAMERA POSE ESTIMATION FOR SMALL-BASELINE VIDEO Yao Y., Zhang Y, Huang Z., Lasenby J.

160. HOLOMOCAP: LOW-COST AUGMENTED REALITY MOTION CAPTURE WITH HOLOLENS 2 Zaccardi S., Jansen B.

161. CONTROLNET-XS: OPTIMISED IMAGE-BASED CONTROL FOR IMAGE SYNTHESIS Zavadski D., Feiden JF., Rother C.

162. MARSLAM: MORE ACCURATE & ROBUST SLAM Zhu Z.

163. TOWARDS PERSONALIZED EMBODIED AI AGENTS Filippo Ziliotto, Jelin Akkara Raphael, Lamberto Ballan, Luciano Serafini, Tommaso Campari

164. GG-SSMS: GRAPH-GENERATING STATE SPACE MODELS Zubić N., Scaramuzza D.

165. EXPLOITING ADVERSARIAL LEARNING AND TOPOLOGY AUGMENTATION FOR OPEN-SET VISUAL RECOGNITION Zuccarà R., Fagetta G., Ortis A., Battiato S.

LOOK&LEARN: BRIDGING PERCEPTION AND GROUNDING GAP IN VISION-LANGUAGE MODELS

Abdelrahman E., Li Y., Iyer S. S., Zhao N., Shechtman E., Singh K. K., Elhoseiny M

Abstract: Vision-Language Models (VLMs) have demonstrated remarkable perception and reasoning capabilities, yet their ability to precisely ground visual concepts remains limited. Existing grounding approaches typically introduce a segmentation decoder, forcing the model to balance between perception and grounding, often at the cost of degraded reasoning performance. In this work, we introduce a segmentation-free grounding approach that enables VLMs to attend to the correct image regions while generating text without requiring an explicit mask generation module. At the core of our method is Where to Look?, a plug-and-play loss function that operates directly on the attention heatmaps of an arbitrary VLM, guiding the model to focus on relevant visual entities during text generation. To enable this, we develop a scalable pseudo-data collection pipeline that generates reliable grounding signals using a mixture of state-of-the-art grounding models and LLM-based adjudication. Our approach achieves strong grounding performance while preserving perception capabilities, outperforming segmentation-based grounding models despite never being trained on ground-truth segmentation masks. Extensive experiments demonstrate that achieves a +3

Moreover, it reduces hallucination and bridges the gap between grounding and reasoning, outperforming LLaVA-G despite having access to fewer datasets. These results validate our hypothesis that grounding and reasoning can be unified within a single model without explicit segmentation supervision, paving the way for more interpretable and multimodal-aware VLMs.

Contact: eslam.abdelrahman@kaust.edu.sa

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 1

COMMONLY INTERESTING IMAGES

Abdullahu F., Grabner H.

Abstract: Images tell stories, trigger emotions, and let us recall memories – they make us think. Thus, they have the ability to attract and hold one's attention, which is the definition of being "interesting". Yet, the appeal of an image is highly subjective. Looking at the image of my son taking his first steps will always bring me back to this emotional moment, while it is just a blurry, quickly taken snapshot to most others. Preferences vary widely: some adore cats, others are dog enthusiasts, and a third group may not be fond of either. We argue that every image can be interesting to a particular observer under certain circumstances. This work particularly emphasizes subjective preferences. However, our analysis of 2.5k image collections from diverse users of the photo-sharing platform Flickr reveals that specific image characteristics make them commonly more interesting. For instance, images, including professionally taken landscapes, appeal broadly due to their aesthetic qualities. In contrast, subjectively interesting images, such as those depicting personal or niche community events, resonate on a more individual level, often evoking personal memories and emotions.

Contact: fitim.abdullahu@zhaw.ch

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 2

MEASURING RELIABILITY AND GENERALIZATION BEYOND IMAGE CLASSIFICATION

Agnihotri S., Keuper M.

Abstract: We benchmark the reliability and generalization of DL models for semantic segmentation, object detection, disparity, and optical flow estimation. Most prior work has focused on image classification, real-world applications in autonomous driving, and medical imaging demand beyond classification tasks. Using SEMSEGBENCH, DETECBENCH, DISPBENCH, and FLOWBENCH, we show that rising i.i.d. accuracy over the years has not improved robustness, calling for new works focused on real-world applicability.

Contact: shashank.agnihotri@uni-mannheim.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 3

CRACK SEGMENTATION FOR STRUCTURAL HEALTH MONITORING

Ahmad T.

Abstract: This study presents a deep learning-based approach for crack segmentation in real-world images using Fully Convolutional Networks (FCN) and DeepLabV3 with ResNet backbones. Enhanced with super-pixel pooling and bootstrapped CNN, the method outperforms existing models on multiple benchmark datasets.

Contact: tasveerahmad@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 4

PRIVACY-PRESERVING PAIN ESTIMATION VIA FACIAL LANDMARK GRAPHS

Fatemah Alhamdoosh, Pietro Pala, Stefano Berretti

Abstract: Automatic pain estimation for nonverbal patients (e.g., newborns, ICU) is crucial. We introduce a privacy-preserving video-based method using facial landmarks. Our model captures dynamic pain via spatio-temporal graph convolution for short-term features and a GRU for long-term patterns. Validated on BioVid Heat Pain and MIntPain datasets, it achieves robust multi-class and binary pain classification, even with occlusions.

Contact: fatemah.alhamdoosh@unifi.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 5

TOWARDS REALISTIC TEST-TIME ADAPTATION: A TRACKLET-BASED BENCHMARK

Alhuwaider S., Alfarra M., Perez J., Ramazanova M., Ghanem B.

Abstract: What's the Problem? Models struggle when real-world inputs don't match their training. That's where Test-Time Adaptation (TTA) comes in.

But... Most TTA benchmarks ignore temporal correlations. Real life? It's sequential, noisy, and repetitive.

What Did We Do? Built a tracklet-based benchmark to mimic real streams. Introduced AdvMem — a memory-initialized boost for stable adaptation.

Contact: shyma.alhuwaider@kaust.edu.sa

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 6

UNCERTAINTY-AWARE KNOWLEDGE DISTILLATION FOR EFFICIENT 6DOF POSE ESTIMATION

ALI OUSALAH N., KACEM A., GHORBEL E., KOUMANDAKIS M., AOUADA D.

Abstract: We propose an uncertainty-aware Knowledge Distillation (KD) framework for keypoint-based 6DoF object pose estimation. We utilize uncertainty in teacher-predicted keypoints to enhance student accuracy while maintaining lightweight structure. Experiments on LINEMOD and SPEED+ benchmarks demonstrate the effectiveness of our method, achieving superior 6DoF object pose estimation with lightweight models compared to state-of-the-art approaches.

Contact: nassim.aliousalah@uni.lu

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 7

PDISCOFORMER: RELAXING PART DISCOVERY CONSTRAINTS WITH VISION TRANSFORMERS

Ananthu Aniraj, Cassio F. Dantas, Dino Ienco, Diego Marcos

Abstract: Computer vision methods that explicitly detect object parts and reason on them are a step towards inherently interpretable models. Existing approaches that perform part discovery make restrictive assumptions about the geometric properties of the discovered parts. We find that such restrictive priors are not required to detect consistent parts using a pre-trained ViT such as DinoV2. We use the total variation (TV) prior which enforces that the parts form spatially connected components.

Contact: ananthu.aniraj@inria.fr

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 8

B-COSIFICATION: TRANSFORMING NEURAL NETWORKS TO BE INHERENTLY INTERPRETABLE

Arya S., Rao S., Böhle M., Schiele B.

Abstract: We propose B-cosification, a method to convert pre-trained models into inherently interpretable ones by replacing linear layers with B-cos transformations, which eliminates the need to train them from scratch. We evaluate design choices for CNNs and ViTs, showing B-cosified models match or exceed B-cos models trained from scratch in interpretability and accuracy, and perform well even in low-resource settings such as zero-shot transfer.

Contact: shreyash.arya@cispa.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 9

TEXT-DRIVEN 3D HAND MOTION GENERATION

Bensabath L, Petrovich M, Varol G

Abstract: Our goal is to train a generative model of 3D hand motions, conditioned on natural language descriptions specifying motion characteristics such as hand shapes, locations, finger/hand/arm movements. We automatically build pairs of 3D hand motions and their associated textual labels with unprecedented scale. We leverage a large-scale sign language video dataset, along with noisy pseudo-annotated sign categories, which we translate into hand motion descriptions via an LLM that utilises a dictionary of sign attributes, as well as our complementary motion-script cues. This data enables training a text-conditioned hand motion diffusion model (HMDM), that is robust across domains such as unseen sign categories from the same sign language, but also signs from another sign language and non-sign hand movements.

Contact: leore.bensabath@enpc.fr

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 10

HIERASURG: WORLD MODELS FOR SURGICAL DATA SCIENCE

Biagini D., Farshad A., Navab N.

Abstract: Synthetic data generation in data-scarce settings, like internal surgery, greatly benefits from the implicit approach enabled by learning-based methods. We first present HieraSurg, a video generation framework consisting of two specialized diffusion models that can generate realistic surgical videos. While visual quality is great, faithfully replicating physics, interactions and domain peculiarities requires a more profound understanding; which we believe ought to be fulfilled by simulation.

Contact: diego.biagini@tum.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 11

WHAT HAPPENS NEXT? ANTICIPATING FUTURE MOTION BY GENERATING POINT TRAJECTORIES

Boduljak G., Karazija L., Laina I., Rupprecht C., Vedaldi A.

Abstract: Recent video generators struggle with accurate motion forecasting, even in simple physical tasks. We attribute this to the overhead of pixel generation. We propose a flow matching model that directly predicts quasi-dense motion trajectories, offering better accuracy and efficiency. We show that our method outperforms both video generation models and prior motion forecasters, challenging the prevailing paradigm of pre-training large video generators for world modelling.

Contact: boduljak.g@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 12

3D FACE RECONSTRUCTION FROM RADAR IMAGES

Braeutigam V., Wirth V., Ullmann I., Schuessler C., Vossiek M., Berking M., Egger B.

Abstract: We propose a novel model-based method for 3D reconstruction from radar images. In our approach, we generate a dataset of synthetic radar images with a physics-based but non-differentiable radar renderer. We use the data to train a CNN-based encoder to predict face model parameters and extend it in an Analysis-by-Synthesis fashion to a model-based autoencoder. This is enabled by learning the rendering process in the decoder network. We evaluate our model on synthetic and real data.

Contact: valentin.braeutigam@fau.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 13

CROSS-SPECTRAL GATED-RGB STEREO DEPTH ESTIMATION

Brucker S., Walz S., Bijelic M., Heide F.

Abstract: We propose a novel stereo-depth estimation method that combines high-resolution HDR RCCB stereo imaging with gated near-infrared (NIR) sensing. By leveraging multi-view cues from both RGB and NIR images, along with active illumination captured through gated imaging, the system achieves dense and accurate depth estimation at long ranges. The proposed method uses only low-cost CMOS sensors and flood-illumination, making it both scalable and practical. It outperforms existing approaches by 39% in mean absolute error (MAE) over the 100–220 m range on accumulated LiDAR ground truth.

Contact: samuel.brucker@torc.ai

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 14

SUPEREVENT: CROSS-MODAL LEARNING OF EVENT-BASED KEYPOINT DETECTION

Burkhardt Y., Schaefer S., Leutenegger S.

Abstract: We propose SuperEvent, a data-driven method for predicting stable keypoints and expressive descriptors from event data. We leverage existing frame-based keypoint detectors on event-synchronized grayscale frames for self-supervision, generating temporally sparse pseudo-labels considering that events are a product of both scene appearance and camera motion. We integrate SuperEvent into a modern sparse SLAM framework, initially developed for traditional cameras, and achieve state-of-the-art results in event-based SLAM.

Contact: yannick.burkhardt@tum.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 15

CONDITIONAL DDPMS FOR LOW-LIGHT ENHANCEMENT

Cabassa G.

Abstract: This work introduces a new supervised approach to address Low-Light Image Enhancement (LLIE) by exploiting conditional Denoising Diffusion Probabilistic Models (DDPMs), with the aim of reproducing the distribution of normal-light images given the low-light one. The second part of the work focuses on obtaining larger images, exploiting Latent Diffusion Models, and formalizing a new generative process based on darkening images, more suitable to low-light scenario.

Contact: greta.cabassa01@universitadipavia.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 16

RECURRENCE-ENHANCED TRANSFORMERS FOR ROBUST MULTIMODAL DOCUMENT RETRIEVAL

Caffagni D., Sarto S., Cornia M., Baraldi L., Cucchiara R.

Abstract: Multimodal information retrievers typically rely on late feature fusion schemes that may overlook low-level, fine-grained details. Conversely, we propose ReT, a Recurrence-enhanced Transformer that fuses and weights the importance of visual and textual features at each layer of unimodal backbones, as well as the contribution of previous layers. Experiments on M2KR and M-BEIR benchmarks show that ReT achieves state-of-the-art performance across diverse settings.

Contact: davide.caffagni@unimore.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 17

INCREMENTAL AND DECREMENTAL CONTINUAL LEARNING FOR PRIVACY-PRESERVING VIDEO RECOGNITION

Caselli L., Magistri S., Bianconcini T., Benericetti A., de Andrade DC., and Bagdanov AD.

Abstract: Continual Learning (CL) enables models to continuously learn new data while retaining prior knowledge. To address privacy, traditional methods discard ALL prior data, unlike industrial practices with gradual changes. We propose novel Incremental and Decremental CL scenarios where new data sources expand datasets, and removal policies gradually reduce past data. We adopt a category-subcategory label setup, to allow a finer control of data changes in our scenarios.

Contact: lorenzo.caselli@unifi.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 18

SAEMNESIA: ERASING CONCEPTS IN DIFFUSION MODELS WITH SPARSE AUTOENCODERS

Cassano E., Renzulli R., Grangetto M.

Abstract: Diffusion models can generate harmful or copyrighted content, requiring concept unlearning. Traditional methods lack transparency and need curated datasets. Sparse Autoencoders (SAEs) provide interpretable features enabling targeted concept unlearning without supervised training. We propose enhanced fine-tuning strategies improving neuron-level interpretability, allowing specific concept removal without extensive hyperparameters tuning or performance degradation, advancing safe unlearning.

Contact: enrico.cassano@unito.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 19

GROCO: GROUND CONSTRAINT FOR METRIC SELF-SUPERVISED MONOCULAR DEPTH

Aurélien Cecille, Stefan Duffner, Franck Davoine, Thibault Neveu, Rémi Agier

Abstract: Metric monocular depth estimation struggles to generalize across diverse camera poses and datasets. Leveraging the success of supervised methods that utilize ground prior information, we propose a novel constraint to integrate this knowledge into the self-supervised setting, addressing the additional challenge of scale recovery. This mechanism ensures the coherence between depth prediction and ground prior, and improves generalization.

Contact: aurelien.cecille@liris.cnrs.fr

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 20

FORCE-AWARE 3D CONTACT MODELING FOR STABLE GRASP GENERATION

Chen Z., Zhang Z., Cheng Y., Leonardis A., Chang H.

Abstract: We focus on stable grasp generation using explicit contact force predictions. We first define a force-aware contact representation, then formulate the stability problem as minimizing accelerations and obtain physical constraints. The constraints then help identify key contact points for stability which provide effective initialization and guidance for stable pose optimization. Experiments show that our method brings an over 30% improvement in stability and adapts well to novel objects.

Contact: zxc417@student.bham.ac.uk

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 21

SPLATFORMER: POINT TRANSFORMER FOR ROBUST 3D GAUSSIAN SPLATTING (ICLR 2025 SPOTLIGHT)

Chen Yutong, Mihajlovic Marko, Chen Xiyi, Wang Yiming, Prokudin Sergey, Tang Siyu

Abstract: 3D Gaussian Splatting (3DGS) has recently transformed photorealistic reconstruction, achieving high visual fidelity and real-time performance. However, rendering quality significantly deteriorates when test views deviate from the camera angles used during training, posing a major challenge for applications in immersive free-viewpoint rendering and navigation. In this work, we conduct a comprehensive evaluation of 3DGS and related novel view synthesis methods under out-of-distribution (OOD) test camera scenarios. By creating diverse test cases with synthetic and real-world datasets, we demonstrate that most existing methods, including those incorporating various regularization techniques and data-driven priors, struggle to generalize effectively to OOD views. To address this limitation, we introduce SplatFormer, the first point transformer model specifically designed to operate on Gaussian splats. SplatFormer takes as input an initial 3DGS set optimized under limited training views and refines it in a single forward pass, effectively removing potential artifacts in OOD test views. To our knowledge, this is the first successful application of point transformers directly on 3DGS sets, surpassing the limitations of previous multi-scene training methods, which could handle only a restricted number of input views during inference. Our model significantly improves rendering quality under extreme novel views, achieving state-of-the-art performance in these challenging scenarios and outperforming various 3DGS regularization techniques, multi-scene models tailored for sparse view synthesis, and diffusion-based frameworks. The project url is https://sergeyprokudin.github.io/splatformer.

Contact: chenytjudy@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 22

BRIDGING DOMAIN GAP IN 6-DOF POSE ESTIMATION VIA CONTRASTIVE ALIGNMENT AND PSEUDO-LABEL REFINEMENT

Chenni N., Rathinam A., Aouada D.

Abstract: Deep models for 6-DoF pose estimation often fail to generalize from synthetic to real images due to domain shifts. We propose CAPLR, a keypoint-based unsupervised domain adaptation (UDA) framework that bridges this gap using contrastive alignment and pseudo-label refinement. Unlike existing methods, CAPLR improves adaptation under large appearance changes, occlusions and complex geometries. It achieves state-of-the-art results on LINEMOD and strong performance on SPEED+ with fewer parameters.

Contact: nidhal.chenni@uni.lu

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 23

ARTIFICIAL INTELLIGENCE FOR ROBOTIC SURGERY

Chiesa G., Renzulli R., Grangetto M.

Abstract: Robotic-assisted surgery has revolutionized modern medicine by enabling minimally invasive procedures with enhanced precision and control. Despite the widespread adoption of systems like the da Vinci robot, challenges remain—particularly in integrating real-time anatomical models within the surgeon's operative field without obstructing visibility. This PhD project proposes an augmented reality (AR) framework to enhance the visualization capabilities of robotic surgery using AI-powered models. By leveraging convolutional and vision transformer networks, we aim to improve tissue detection, segmentation, and tracking through multimodal preoperative data. Our pipeline involves the creation of a dataset from endoscopic videos, fine-tuning of Medical SAM 2, and the distillation of a lightweight, real-time segmentation model. A key objective is the real-time segmentation of robotic instruments to enable accurate and unobtrusive overlay of 3D anatomical models. We also explore memory-enhanced architectures for dynamic frame analysis. This work represents a step toward AI-integrated AR systems for more intuitive, efficient, and autonomous surgical navigation.

Contact: giorgio.chiesa@unito.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 24

GRAMIAN MULTIMODAL LEARNING AND ALIGNMENT

Cicchetti Giordano

Abstract: Multimodal learning integrates data from multiple modalities to enhance model capabilities. Existing methods often use pairwise cosine similarity to align encoder representations. Despite its efficiency, this approach could not be naturally extended to whatever number of modalities. Gramian Representation Alignment Measure (GRAM) overcomes this limitation by aligning n modalities directly in the higher-dimensional space in which modality embeddings lie simultaneously.

Contact: giordano.cicchetti@uniroma1.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 25

SHEDDING LIGHT ON DEPTH: EXPLAINABILITY ASSESSMENT IN MONOCULAR DEPTH ESTIMATION

Cirillo L., Schiavella C., Papa L., Russo P., Amerini I.

Abstract: Explaining Monocular Depth Estimation (MDE) remains largely unexplored despite its real-world use. We investigate Saliency Maps, Integrated Gradients, and Attention Rollout on different computationally complex models for MDE: METER, a lightweight network, and PixelFormer, a deep network. To evaluate the explanation reliability, we introduce the Attribution Fidelity, a metric that identifies failures of an explainability method, even when conventional metrics might suggest satisfactory results.

Contact: cirillo@diag.uniroma1.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 26

JOINT OPTIMIZATION OF FILTER ATTACHMENTS AND SUPER-RESOLUTION FOR SPECTRAL IMAGING WITH STEREO RGB CAMERAS

Cogo L., Buzzelli M., Bianco S., Schettini R.

Abstract: Context → Spectral imaging captures and processes images across multiple wavelengths of the electromagnetic spectrum, uncovering information crucial for applications in fields such as smart agriculture, remote sensing, healthcare, and industrial quality control [1]. Open issues → Various designs for spectral imaging devices have been proposed, often relying on scanning along spatial, spectral, or temporal dimensions. However, these approaches typically face trade-offs in resolution, acquisition time, portability, and cost. Our solution → We propose a novel spectral imaging framework based on a stereo camera setup equipped with filter attachments. Spectral information is captured and subsequently reconstructed using a spectral super-resolution module. Crucially, we jointly optimize both the filters' transmittances and the super-resolution model parameters through backpropagation, enabling an efficient and integrated design. Our approach allows to deliver high-quality spectral reconstructions while maintaining portability, low cost, and fast acquisition.

Contact: luca.cogo@unimib.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 27

MULTIMODAL SAM-ADAPTER FOR SEMANTIC SEGMENTATION

Curti Iacopo, Zama Ramirez Pierluigi, Petrelli Alioscia, Di Stefano Luigi

Abstract: Semantic segmentation is vital in fields like autonomous driving and medical imaging but struggles under adverse conditions. Multimodal approaches incorporating auxiliary sensor data have emerged to address this limitation. We propose MM SAM-Adapter, which enhances the Segment Anything Model by injecting multimodal fused features into its RGB backbone. This allows selective use of auxiliary data when beneficial. It achieves state-of-the-art results when evaluated on DeLiVER, FMB, and MUSES.

Contact: iacopo.curti2@unibo.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 28

JUST DANCE WITH π ! A POLY-MODAL INDUCTOR FOR WEAKLY-SUPERVISED VIDEO ANOMALY DETECTION

Mahji, S., D'Amicantonio, G., Dantcheva, A., Kong, Q., Garattoni, L., Francesca, G., Bondarev, E., Bremond, F.,

Abstract: Weakly-supervised methods for video anomaly detection (VAD) are conventionally based merely on RGB spatio-temporal features, which continues to limit their reliability in real-world scenarios. This is due to the fact that RGB-features are not sufficiently distinctive in setting apart anomalies from visually similar events. Therefore, it is essential to augment RGB spatio-temporal features by additional modalities. We introduce the Poly-modal Induced framework for VAD: π-VAD, a novel approach that augments RGB representations by five additional modalities. Specifically, the modalities include sensitivity to fine-grained motion (pose), three dimensional scene and entity representation (depth), surrounding objects (panoptic masks), global motion (optical flow), as well as language cues (VLM). π-VAD includes two plug-in modules, namely Pseudo-modality Generation module and Cross Modal Induction module, which generate modality-specific prototypical representation and, thereby, induce multi-modal information into RGB cues. These modules operate by performing anomaly-aware auxiliary tasks and necessitate five modality backbones – only during training. π-VAD achieves sota on three prominent VAD datasets encompassing real-world scenarios, without requiring the computational overhead of five modality backbones at inference.

Contact: g.d.amicantonio@tue.nl

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 29

NESYLAD: A NEURO-SYMBOLIC APPROACH FOR UNSUPERVISED LOGICAL ANOMALY DETECTION

Dahmardeh M., Setti F.

Abstract: Detecting logical anomalies in industrial settings remains challenging for conventional methods. We propose NeSyLAD, a neuro-symbolic framework combining deep learning component segmentation with symbolic rule extraction and logical reasoning. Our approach extracts interpretable rules from CNN activations using ERIC and applies LTN reasoning to detect anomalies. Evaluated on MVTec LOCO dataset, achieving 0.89 AUC with explainable results.

Contact: MALIHE.DAHMARDEH@UNIVR.IT

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 30

SUPERVISING 3D TALKING HEAD AVATARS WITH ANALYSIS-BY-AUDIO-SYNTHESIS

Daněček R., Schmitt C., Polikovsky S., Black M.

Abstract: We present THUNDER, a 3D talking head framework that achieves accurate lip-sync and expressive facial animation via a novel audio-based self-supervision mechanism. By training a mesh-to-speech model to predict audio from facial motion, we enable a differentiable loop that compares generated audio representations to those of the input audio. This analysis-by-audio-synthesis approach improves lip-sync accuracy in a stochastic diffusion-based avatar model without sacrificing expressiveness.

Contact: radek.danecek@tuebingen.mpg.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 31

VIDEO OBJECT DETECTION IN MARITIME SCENARIOS

Denk F., Moser D., Rankl C., Sablatnig R.

Abstract: To assess if temporal context improves detection over static methods in maritime imagery, this work evaluates the performance of two Video Object Detection models, T-YOLOX and YOLOV++, on three maritime datasets and studies positional encoding to enhance tiny object detection. Two datasets show minor F1-score gains, while the third demonstrates a decline of up to 10%. Positional encoding improves tiny object detection but worsens detection performance for other object size categories.

Contact: fdenk@cvl.tuwien.ac.at

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 32

RESOLUTION WHERE IT COUNTS: HASH-BASED GPU-ACCELERATED 3D RECONSTRUCTION VIA VARIANCE-ADAPTIVE VOXEL GRIDS

De Rebotti L., Giacomini E., Grisetti G., Di Giammarino L.

Abstract: We introduce a novel 3D surface reconstruction method using a variance-adaptive, multi-resolution voxel grid, which adjusts voxel size based on local SDF variance and stores data in a flat spatial hash table for efficient GPU use. This approach enables real-time, memory-efficient reconstruction and GPU-accelerated rendering, achieving up to 13× speedup and 4× less memory than fixed-resolution methods, while maintaining reconstruction accuracy.

Contact: derebotti@diag.uniroma1.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 33

PARAMETRIC SHAPE MODELS FOR VESSELS LEARNED FROM SEGMENTATIONS VIA A DIFFERENTIABLE VOXELIZATION LOSS

DIMA Alina, SHIT Suprosanna, QIU Huaqi, HOLLAND Robbie, MUELLER Tamara, MUSIO Fabio, YANG Kaiyuan, MENZE Bjoern, BRAREN Rickmer, MAKOWSKI Marcus, RUECKERT Daniel

Abstract: We propose an approach for extracting parametric models of 3D vessels from segmentations.

By leveraging differentiable voxelization, we perform shape-to-segmentation fitting without the explicit need for ground-truth shape parameterization.

The model parameters (centerlines and radii) can be manipulated post-fit, which is valuable in downstream applications.

Our method accurately captures the geometry of complex vessels, as demonstrated by experiments on the aortas, aneurysms, and brain vessels.

Contact: alina.dima@tum.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 34

MACHINE LEARNING FOR HUMAN ANALYSIS AND BIOMETRICS

Di Domenico N.

Abstract: Despite the high accuracy of Face Recognition Systems, they still face some open challenges. Firstly, their reliance on large-scale face datasets raises ethical, legal, and demographic bias concerns, highlighting the need for high-quality synthetic datasets. Secondly, they are vulnerable to adversarial attacks such as face morphing, which can deceive both human operators and automated checks, thus raising the need for robust detection techniques.

Contact: nicolo.didomenico@unibo.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 35

SELF-SUPERVISED PRE-TRAINING WITH DIFFUSION MODEL FOR FEW-SHOT LANDMARK DETECTION IN X-RAY IMAGES

Di Via R., Odone F., Pastore V. P.

Abstract: Medical landmark detection faces challenges due to limited annotated datasets. We introduce the first application of denoising diffusion probabilistic models (DDPMs) to this task, using them in a novel self-supervised pre-training setup. This approach enables few-shot learning, achieving superior accuracy with fewer than 50 labeled images and outperforming traditional self-supervised and ImageNet-based methods across three benchmark datasets.

Contact: roberto.divia@edu.unige.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 36

SAIL: SELF-SUPERVISED ALBEDO ESTIMATION FROM REAL IMAGES WITH A LATENT DIFFUSION MODEL

Djeghim.H, Piasco.N, Roldao.L, Bennehar.M, Tsishkou.D, Loscos.C, Sidibé.D

Abstract: Intrinsic decomposition in real-world images is challenged by lighting variations and the lack of ground-truth data. We introduce SAIL, a latent diffusion-based method trained on unlabeled multi-illumination data for real images albedo estimation.

Contact: hala.djeghim@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 37

TOWARDS ROBUST MULTIMODAL OUT-OF-DISTRIBUTION GENERALIZATION AND DETECTION FOR REAL-WORLD SYSTEMS

Dong H., Chatzi E., Fink O.

Abstract: Out-of-distribution (OOD) generalization and detection, integrated with multimodal learning, are essential for building robust and safe AI systems in complex real-world environments, especially in safety-critical domains such as autonomous driving and robotics. This research tackles these challenges across four core areas within multimodal settings, offering comprehensive benchmarks and introducing novel solutions to improve model reliability.

Contact: hao.dong@ibk.baug.ethz.ch

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 38

METPOSE: TESTING POSE ESTIMATION ON UNLABELLED DATA

Duran M.

Abstract: Testing Computer Vision systems under varied conditions is essential given their impact in our society. The high cost of data labelling motivates testing without ground truth, especially for Pose Estimation, whose output keypoints' list differ between datasets. MetPose tackles this problem, using metamorphic testing to test image characteristics relevant for each application. MetPose finds problems that ground truth testing does not, such as Mediapipe Hand's sensitivity to mirrored images.

Contact: mduran@tcd.ie

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 39

INTERACTVLM: 3D INTERACTION REASONING FROM 2D FOUNDATIONAL MODEL

Dwivedi S.K., Antić D., Tripathi S., Taheri O., Schmid C., Black M.J., Tzionas D.

Abstract: We introduce InteractVLM, a novel method to estimate 3D contact points on human bodies and objects from single in-the-wild images, enabling accurate human-object joint reconstruction in 3D. We do so by leveraging the broad visual knowledge of a large Visual Language Model and a novel Render-Localize-Lift module. Our method goes beyond traditional binary contact estimation methods by estimating contact points on a human in relation to a specified object.

Contact: sai.dwivedi@tuebingen.mpg.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 40

HD-EPIC: A HIGHLY-DETAILED EGOCENTRIC VIDEO DATASET

Perrett T, Darkhalil A, Sinha S, Emara O, Pollard S, Parida K, Liu K, Gatti P, Bansal S, Flanagan K, Chalk J, Zhu Z, Guerrier R, Abdelazim F, Zhu B, Moltisanti D, Wray M, Doughty H, Damen D

Abstract: We present a validation dataset of newly-collected kitchen-based egocentric videos, manually annotated with highly detailed and interconnected ground-truth labels covering: recipe steps, fine-grained actions, ingredients with nutritional values, moving objects, and audio annotations. Importantly, all annotations are grounded in 3D through digital twinning of the scene, fixtures, object locations, and primed with gaze. Footage is collected from unscripted recordings in diverse home environments, making HDEPIC the first dataset collected in-the-wild but with detailed annotations matching those in controlled lab environments. We show the potential of our highly-detailed annotations through a challenging VQA benchmark of 26K questions assessing the capability to recognise recipes, ingredients, nutrition, fine-grained actions, 3D perception, object motion, and gaze direction. The powerful long-context Gemini Pro only achieves 38.5% on this benchmark, showcasing its difficulty and highlighting shortcomings in current VLMs. We additionally assess action recognition, sound recognition, and long-term video-object segmentation on HD-EPIC. HD-EPIC is 41 hours of video in 9 kitchens with digital twins of 413 kitchen fixtures, capturing 69 recipes, 59K fine-grained actions, 51K audio events, 20K object movements and 37K object masks lifted to 3D. On average, we have 263 annotations per minute of our unscripted videos.

Contact: omar.emara@bristol.ac.uk

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 41

VIDEO UNLEARNING VIA LOW-RANK REFUSAL VECTOR

Facchiano S., Saravalle S., Migliarini M., De Matteis E., Sampieri A., Pilzer A., Rodolà E., Spinelli I., Franco L., Galasso F.

Abstract: Video generative models can replicate harmful concepts from training data. This paper introduces the first unlearning method for video diffusion models using 5 safe/unsafe prompt pairs. By averaging latent differences and applying a low-rank factorization, it computes a "refusal vector" to suppress unsafe concepts without retraining or access to original data, while preserving generation quality and resisting adversarial bypasses.

Contact: simone.facchiano@uniroma1.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 42

INTRAOPERATIVE REGISTRATION BY CROSS-MODAL INVERSE NEURAL RENDERING

Fehrentz M., Azampour M., Dorent R., Rasheed H., Galvin C., Golby A., Wells W., Frisken S., Navab N., Haouchine N.

Abstract: We present a novel approach for 3D/2D registration during neurosurgery via cross-modal neural rendering. Our approach separates implicit neural representation into two components, handling anatomical structure preoperatively and appearance intraoperatively. This disentanglement is achieved by controlling a Neural Radiance Field's appearance with a multi-style hypernetwork. The implicit neural representation serves as a differentiable rendering engine to estimate the surgical camera pose.

Contact: maximilian.fehrentz@tum.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 43

G-SOLVER: GAUSSIAN BELIEF PROPAGATION AND GAUSSIAN PROCESSES FOR CONTINUOU TIME SLAM

Ceriola D., Ferrari S., Di Giammarino L., Brizi L., Grisetti G.

Abstract: Continuous-time Simultaneous Localization and Mapping (CT-SLAM) combines data from multiple asynchronous sensors to estimate the state over time. Traditional methods use Nonlinear Least Squares (NLLS) solvers, which are unsuitable for handling asynchronous measurements. In contrast, distributed Gaussian Belief Propagation (GBP) offers a scalable, decentralized approach that naturally manages uncertainty. However, existing GBP methods for continuous SLAM rely on spline-based interpolation, which requires manual tuning and does not capture uncertainty well. Gaussian Processes (GPs) provide a robust alternative by modeling dynamics and their uncertainty. In this paper, we introduce a distributed GBP solver that uses GP priors for continuous-time trajectory estimation, resulting in improved accuracy and efficiency without compromising execution times, fundamental for real-world applications. We release an open-source implementation.

Contact: s.ferrari@diag.uniroma1.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 44

PATCH SIZE CURRICULUM IN 3D PATCH-BASED SEGMENTATION

Fischer S., Kiechle J., Peeken J. Schnabel J.

Abstract: Curriculum learning enhances model training convergence. We propose a novel curriculum for 3D patch-based medical image segmentation. This approach progressively increases patch size during training. We operate it in two modes. First, we apply it as a resource-efficient strategy, reducing runtime by 50% while keeping comparable performance to standard training. In the second mode, we maximize segmentation performance by improving convergence over standard training within the same runtime.

Contact: stefan.mi.fischer@tum.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 45

LEARN YOUR SCALES

Forghani F, Yu J, Aumentado-Armstrong T, Derpanis K, Brubaker M

Abstract: Depth-free multiview data captured with uncalibrated monocular cameras has an ambiguous scale. We demonstrate the effect of such scale ambiguity when used to train generative novel view synthesis methods (GNVS). We propose a framework to estimate scene scales jointly with the GNVS model in an end-to-end fashion. We further define two metrics based on optical flow and epipolar geometry to quantify the scale inconsistency present in a GNVS model.

Contact: fereshtforghani@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 46

ONLINE EPISODIC MEMORY VISUAL QUERY LOCALIZATION WITH EGOCENTRIC STREAMING OBJECT MEMORY

Zaira Manigrasso, Matteo Dunnhofer, Antonino Furnari, Moritz Nottebaum, Antonio Finocchiaro, Davide Marana, Rosario Forte, Giovanni Maria Farinella, Christian Micheloni

Abstract: Episodic memory retrieval enables wearable devices to recall objects or events previously observed in video (e.g., "where did I last see my smartphone?"). However, existing formulations assume an "offline" scenario in which the full video history can be accessed when the user makes a query, hindering applicability in real settings, where wearable devices are limited in power and storage capacity. Towards more application-ready episodic memory systems, we propose Online Visual Query 2D (OVQ2D), a new task requiring models to process video streams online, observing video frames only once, storing relevant information, and relying on a compact memory, rather than past video, to retrieve object localizations upon user query. To tackle OVQ2D, we propose ESOM (Egocentric Streaming Object Memory), a novel framework integrating an object discovery module to detect key objects, a visual object tracker to track their position through the video, and a memory module to store spatio-temporal object coordinates and image frames, which can be queried efficiently at any moment. Experiments on Ego4D show ESOM's effectiveness, surpassing other online methods. However, OVQ2D remains challenging, with the best method achieving only 4% success. ESOM's performance improves significantly with perfect object tracking (31.91%), detection (40.55%), or both (81.92%), highlighting the need for advancements in these components. Our study provides a competitive benchmark to advance online episodic memory and assess the real-world applicability of object detection and tracking.

Contact: rosario.forte@phd.unict.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 47

DENSE 3D MAPPING FOR SPATIAL AI

Fry N., Kelly P., Davison A.

Abstract: We investigate efficient 3D scene representations to advance robot intelligence, particularly in the context of emerging vision-chip architectures. We evaluate methods like SuperPrimitive, which lack persistent mapping, and 3D Gaussian Splatting, which is expressive but produces poor geometry. Our goal is to leverage differentiable rendering to generate abstract, higher-level "primitives" that balance efficiency, reusability, and semantic understanding.

Contact: nf20@imperial.ac.uk

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 48

SKELETON-BASED ACTION RECOGNITION FOR BIOMECHANICAL RISK ASSESSMENT

Gennarelli I., Ranavolo A., Micheloni C., Martinel N.

Abstract: This work evaluates biomechanical risk in lifting by combining multimodal pose estimation (IMUs, IR, RGB) with traditional methods and several skeleton-based action recognition networks, including the newly proposed SkelMamba [1]. Results demonstrate the feasibility of vision-based approaches for scalable, automated ergonomic risk assessment.

Contact: gennarelli.irene@spes.uniud.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 49

PIXEL3DMM: VERSATILE SCREEN-SPACE PRIORS FOR SINGLE-IMAGE 3D FACE RECONSTRUCTION

Giebenhain Simon, Kirschstein Tobias, Rünz Martin, Agapito Lourdes, Nießner Matthias

Abstract: We propose Pixel3DMM, a set of highly-generalized ViTs which predict per-pixel geometric cues to guide the optimization of a 3DMM. We train specialized surface normal and uv-coordinate prediction heads on top of DINOv2. We register 3 high-quality 3D face datasets to FLAME topology, resulting in over 1,00a0 identities and 976K images. We introduce a new benchmark for single-image face reconstruction. Crucially, our benchmark is the first to evaluate posed and neutral facial geometry.

Contact: simon.giebenhain@tum.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 50

TOWARDS A PERCEPTUAL EVALUATION FRAMEWORK FOR LIGHTING ESTIMATION

Giroux J., Dastjerdi M., Hold-Geoffroy Y., Vazquez-Corral J., Lalonde J.-F.

Abstract: Lighting estimation is often evaluated using IQA metrics on benchmark datasets. Yet, we show these metrics don't align with human preference when relighting virtual scenes into real photographs. We conduct a controlled psychophysical study where observers compare scenes rendered with different lighting methods. Results show no single metric reflects human perception of lighting. However, combining them better matches human preference, offering a perceptual framework for evaluation of lighting.

Contact: justine.giroux.2@ulaval.ca

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 51

UNDERLOC: IMAGE BASED RELOCALIZATION AND ALIGNMENT FOR DYNAMIC UNDERWATER ENVIRONMENTS

Gorry B., Fischer T., Milford M., Fontan A.

Abstract: Underwater ecosystem monitoring is crucial but challenging to automate due to the complexities of underwater imagery which hinder traditional visual methods. Our integrated Visual Place Recognition (VPR), feature matching, and image segmentation pipeline enables robust identification of revisited areas and change analysis from video-derived images. Furthermore, we introduce the SQUIDLE+ VPR Benchmark—a large-scale underwater VPR benchmark with diverse data from multiple robotic platforms.

Contact: beverley.gorry@hdr.qut.edu.au

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 52

EVENT-BASED PHOTOMETRIC BUNDLE ADJUSTMENT

Guo S., Gallego G.

Abstract: We tackle the problem of bundle adjustment for a purely rotating event camera. Starting from first principles, we formulate this BA problem as a classical non-linear least squares optimization. The photometric error is defined using event generation model directly in camera rotations and semi-dense intensity map that triggers events.

Contact: shuang.guo@campus.tu-berlin.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 53

VID2AVATAR-PRO: AUTHENTIC AVATAR FROM VIDEOS IN THE WILD VIA UNIVERSAL PRIOR

Guo C., Li J., Kant Y., Sheikh Y., Saito S., Cao C.

Abstract: Building high-quality animatable avatars from a monocular video is challenging because the pose diversity and view points are limited, leading to poor generalization when animated. We address these limitations by leveraging a universal prior model learned from a large corpus of multi-view clothed human performance data. Once trained, we fine-tune the model with an in the-wild video to obtain a personalized avatar that can be faithfully animated to novel motions and rendered from novel views.

Contact: chen.guo@inf.ethz.ch

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 54

ETAP: EVENT-BASED TRACKING OF ANY POINT

Hamann F., Gehrig D., Febryanto F., Daniilidis K., Gallego G.

Abstract: We introduce the first event-only method for tracking any point (TAP). RGB-based TAP works well in good conditions but fails in challenging scenarios like fast movements & low light. Event cameras handle these with high temporal resolution, low motion blur, and high dynamic range. Our model uses a new synthetic event dataset and event-specific feature alignment loss. We outperform SOTA event-based feature tracking methods by ¿20%.

Contact: f.hamann@tu-berlin.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 55

THE INVISIBLE EGOHAND: 3D HAND FORECASTING THROUGH EGOBODY POSE ESTIMATION

Hatano M., Zhu Z., Saito H., Damen D.

Abstract: Forecasting hand motion and pose from an egocentric perspective is essential for understanding human intention. However, existing methods focus solely on predicting positions without considering articulation, and only when the hands are visible in the field of view. In this paper, we propose a method to forecast the 3D trajectories and poses of both hands from an egocentric video, both in and out of the field of view.

Contact: hatano1210@keio.jp

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 56

LIMO: LIFELIKE HUMAN MOTION GENERATION WITH CONTINUOUS-SPACE GENERATIVE MODELS

He Y., Tiwari G., Zhang X., Bora P., Birdal T., Lenssen J., Pons-Moll G.

Abstract: LiMo is a generative text-to-motion model that generates realistic lifelike motions. Current generative models either use diffusion on full sequences or auto-regressive models in quantized spaces. Both have difficulties generating detailed motions. With LiMo, we combine both paradigms and present an auto-regressive diffusion model in a continuous latent space, which can successfully generate long motions while keeping high-frequency motions. To analyze generated motions with respect to their frequency components, we introduce a novel metric on the power spectral densities (PSDs) of motions. We demonstrate that our PSD metric is more aligned to human perception compared to previous metrics, and show that LiMo outperforms previous works on common generative metrics such as FID, our novel PSD-based metric, and human evaluation.

Contact: yannan.he@uni-tuebingen.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 57

ADVANCING PERSONAL HUMAN-CENTRIC AI WITH GENERATIVE MODELING

Ho H., Kaufman M., Zhang L., Hilliges O.

Abstract: Personalization is a key aspect of future human-centric AI systems. It becomes increasingly important to explore how human-centric models can enhance personalized understanding and interaction. In my PhD research, I investigate how generative models can be leveraged to improve two human-centric tasks: single-view 3D reconstruction and 3D pose estimation. In the SiTH project, I propose an image-conditioned diffusion model for 3D reconstruction. This method requires only a small number of 3D scans for training and can reconstruct high-quality, textured 3D meshes in under two minutes. SiTH enables easy generation of personalized 3D avatars, supporting various downstream applications. In the PHD project, I develop a new pipeline for personalized 3D pose estimation. The approach uses a point-based diffusion model that incorporates both image features and personal body shape information to generate plausible 3D points. These serve as a robust prior for accurate pose estimation in diverse, in-the-wild settings. PHD advances the goal of building perceptive AI systems capable of understanding individuals' body states and positions more accurately. Together, these works contribute to enabling more adaptive, personalized, and perceptive AI systems for the future.

Contact: hohs@ethz.ch

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 58

NEURAL RENDERING FOR SENSOR ADAPTATION IN 3D OBJECT DETECTION

Embacher F., Holtz D., Uhrig J., Cordts M., Enzweiler M.

Abstract: Autonomous vehicles face a cross-sensor domain gap due to varying camera setups, affecting 3D object detector accuracy. We introduce CamShift, a dataset simulating this gap between subcompact vehicles and SUVs. CamShift reveals significant performance degradation and highlights robustness dependencies on model architecture. Additionally, we propose a neural rendering-based sensor adaptation pipeline, improving performance across detectors and reducing the need for new data collection.

Contact: holtz_david@web.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 59

MONOCULAR CAMERA-BASED SIDEWALK: WIDTH ESTIMATION FOR WHEELCHAIR ACCESSIBILITY

Houshyar Yazdian S.H., Jacquet W., Stiens J

Abstract: There is a critical lack of publicly available data on sidewalk infrastructure, limiting navigation apps from offering safe, accessible routes—especially for wheelchair users. We propose a low-cost, automated method to estimate sidewalk widths using a single monocular camera, enabling real-time and scalable data collection in urban environments. This approach facilitates inclusive navigation, supports accessibility audits, and contributes to smart city planning. A simple camera becomes a powerful tool to map sidewalk accessibility and enhance mobility for vulnerable populations.

Contact: seyedeh.hiva.houshyar.yazdian@vub.be

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 60

LITEREALITY: GRAPHIC-READY 3D SCENE RECONSTRUCTION FROM RGB-D SCANS

Huang Z., Wu X., Zhong F., Zhao H., Niessner M., Lasenby J.

Abstract: LiteReality converts RGB-D scans into compact, realistic, and interactive 3D scenes with object individuality, articulation, and high-quality PBR materials. The output is fully compatible with graphics pipelines, supporting applications in AR/VR, gaming, robotics, and digital twins.

Contact: zh340@cam.ac.uk

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 61

HOW TO PREDICT SOCIO-ECONOMIC DEVELOPMENT FROM SPACE?

Janisiów Ł., Wójcik P.

Abstract: Access to up-to-date statistical data is crucial for understanding and responding to socio-economic trends. However, in many parts of the world, official statistics are often delayed, and in some regions, reliable data may be scarce or entirely unavailable. This makes it difficult for researchers, policymakers, and organizations to make informed and timely decisions. Traditionally, nighttime lights (NTL) satellite imagery has been used as a proxy for economic activity, but this approach lacks precision. Recent advances in high-resolution satellite imagery, combined with increased computational power, offer new opportunities for near real-time socio-economic monitoring. Compared to traditional statistical data, high-resolution imagery provides two key advantages: it is available almost in real time, and predictions can be made for any area, even if it does not align with official administrative boundaries. In this study, we propose two approaches for extracting features from high-resolution satellite imagery to predict socio-economic indicators. The first is a classification model that assigns land cover classes to satellite images. The second is a transfer learning model, initially trained to predict NTL intensity, whose learned image representations are repurposed for the socio-economic prediction task. This research uses satellite imagery from the European Space Agency's Sentinel-2 program and focuses on predicting socio-economic indicators across all administrative units in Poland.

Contact: l.janisiow@student.uw.edu.pl

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 62

ADVANCING FOREST TYPOLOGY FROM SPACE

Jiang Y., Neumann M.

Abstract: Existing forest maps focus on where forests are; we ask what kinds of forests exist. We develop deep learning methods for dense geospatial prediction, addressing temporal, multi-modal, and label-scarce challenges. Our benchmark (FORTY) and transformer model (MTSViT) enable accurate forest type mapping at scale, showcasing how vision models can advance environmental monitoring and global sustainability.

Contact: yuchang.jiang@uzh.ch

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 63

GEO4D: LEVERAGING VIDEO GENERATORS FOR GEOMETRIC 4D SCENE RECONSTRUCTION

Jiang Zeren, Zheng Chuanxia, Laina Iro, Larlus Diane, Vedaldi Andrea

Abstract: We introduce Geo4D, a method to repurpose video diffusion models for dynamic scene reconstruction. By leveraging the strong dynamic prior captured by such video models, Geo4D can be trained using only synthetic data while generalizing well to real data in a zero-shot manner. Geo4D predicts several complementary geometric modalities. We introduce a new multi-modal alignment algorithm to align and fuse these modalities, thus obtaining robust and accurate 4D reconstruction of long videos.

Contact: zeren.jiang99@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 64

SELF-SUPERVISED COLLABORATIVE DISTILLATION FOR LIGHTNING-ROBUST AND 3D-AWARE 2D REPRESENTATIONS

Jo W., Ha H., Kim J.-Y., Jeong H., Oh T.-H.

Abstract: As deep learning advances, self-supervised learning enables 2D image encoders to extract features for vision tasks. However, they struggle with nighttime conditions and limited 3D awareness. We propose Collaborative Distillation, a novel self-supervised method that improves light-invariance and 3D understanding by integrating 2D and 3D LiDAR data. Our method outperforms others across lighting conditions and generalizes well.

Contact: jo1jun@postech.ac.kr

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 65

ADVANCING GENERALIZABILITY AND FAIRNESS IN BREAST CANCER: MAMA-MIA CHALLENGE

Garrucho Lidia, Joshi Smriti, Kushibar Kaisar, Bobowicz Maciej, Bargalló Xavier, Jaruševičius Paulius, Lekadir Karim

Abstract: We present a T1-weighted dynamic contrast-enhanced MRI (DCE-MRI) dataset with pre-treatment scans from 1,506 biopsy-confirmed breast cancer patients. To advance research in automated diagnosis and treatment response prediction, we are hosting the MICCAI 2025 MAMA-MIA challenge focused on reproducible benchmarks for tumor segmentation and response prediction to neoadjuvant chemotherapy (NAC).

Contact: smriti.joshi@ub.edu

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 66

UNFOLDING CLOTH: NEURAL DEFORMATION FIELDS FOR SIMULATION AND MONOCULAR TRACKING

Kairanda N., Habermann M., Naik S., Theobalt C., Golyanik V.

Abstract: We represent clothes as continuous neural fields instead of discrete meshes to address core challenges in cloth simulation and 3D surface tracking.

Contact: nkairand@mpi-inf.mpg.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 67

PRIMEDEPTH: EFFICIENT MONOCULAR DEPTH ESTIMATION WITH A STABLE DIFFUSION PREIMAGE

Zavadski D., Kalšan D., Rother C.

Abstract: Text-to-Image generative models provide a rich and generic image representation, which we dub preimage and use for monocular depth estimation. We extract the preimage at a single denoising step and integrate it into a network with an architectural inductive bias. Our method is two orders of magnitude faster at test time, more robust, and marginally superior to the competing diffusion-based method, while retaining detailed depth predictions and requiring little labelled training data.

Contact: damjan.kalsan@iwr.uni-heidelberg.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 68

SELF-SUPERVISED PRETRAINING FOR FINE-GRAINED PLANKTON RECOGNITION

Kareinen J., Eerola T., Kraft K., Lensu L., Suikkanen S., Kälviäinen H.

Abstract: Plankton recognition is crucial for ocean monitoring due to plankton's role in ocean food webs and carbon capture. Automated imaging systems collect large-scale plankton data requiring computer vision. The task is challenging due to fine-grained species and dataset shifts across instruments and locations. We study self-supervised pretraining on diverse data to learn a general encoder that improves accuracy over ImageNet baselines, especially when unlabeled target data is used during pretraining.

Contact: joona.kareinen@lut.fi

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 69

SEED4D: A SYNTHETIC EGO-EXO DYNAMIC 4D DATA GENERATOR, DRIVING DATASET AND BENCHMARK

Kästingschäfer M., Gieruc T., Bernhard S., Campbell D., Insafutdinov E., Najafli E., Brox T.

Abstract: We propose a Synthetic Ego–Exo Dynamic 4D (SEED4D) data generator and dataset. SEED4D is a customizable, easy-to use data generator for spatio-temporal multi-view data creation. The open-source data generator allows the creation of synthetic data for camera setups commonly used in the NuScenes, KITTI360, and Waymo datasets. Additionally, SEED4D encompasses two large-scale multi-view synthetic urban scene datasets. Our static (3D) dataset encompasses 212k inward- and outward-facing vehicle images from 2k scenes, while our dynamic (4D) dataset contains 16.8M images from 10k trajectories.

Contact: marius.kaestingschaefer@online.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 70

DUALPM: DUAL POSED-CANONICAL POINT MAPS FOR 3D SHAPE AND POSE RECONSTRUCTION

Kaye, B., Jakab, T., Wu, S., Rupprecht, C., Vedaldi, A.

Abstract: We introduce Dual Point Maps (DualPM), paired multi-layer point maps representation well suited for deformable object 3D reconstruction. We train a fully supervised models on our representation with single template synthetic data of each category for which we use depth peeling to extract an exact ground truth.

Contact: benkaye001@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 71

INCORPORATING PROPERTIES OF HUMAN VISION INTO IMAGE GENERATION

Kergaßner S., Tariq T., Didyk P.

Abstract: The vision for real-time virtual and augmented reality is to reproduce our visual reality in its entirety on immersive displays. However, we are highly constrained by the capabilities of display technologies and rendering algorithms. To overcome these limitations, modern graphics pipelines leverage the limitations of human vision to allocate computational resources to the fovea while reducing quality in the periphery.

Poster: https://drive.google.com/file/d/1yXFmFJkCqUEQJwhyhMWQwtq3EP–44mN/view?usp=drive_link

Contact: sophie.kergassner@usi.ch

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 72

YOUR VIT IS SECRETLY AN IMAGE SEGMENTATION MODEL

Kerssies T., Cavagnero N., Hermans A., Norouzi N., Averta G., Leibe B., Dubbelman G., de Geus D.

Abstract: Vision Transformers (ViTs) excel at vision tasks, but segmentation methods add an adapter, pixel decoder, and Transformer decoder. We show these components can be removed, as large ViTs learn their inductive biases with sufficient pre-training. We introduce the Encoder-only Mask Transformer (EoMT), which repurposes the plain ViT for segmentation. With large-scale models and pre-training, EoMT achieves accuracy similar to models with task-specific components while running significantly faster.

Contact: t.kerssies@tue.nl

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 73

KAIROSAD: A SAM-BASED MODEL FOR INDUSTRIAL ANOMALY DETECTION ON EMBEDDED DEVICES

Khan U., Fummi F., Capogrosso L

Abstract: In intelligent manufacturing, anomaly detection is crucial for quality control on production lines. Existing models are too large and computationally heavy for embedded devices. We present KairosAD, a supervised approach that uses the Mobile Segment Anything Model (MobileSAM) for image-based anomaly detection. KairosAD is 78% smaller, 4× faster, and maintains comparable AUROC performance to state-of-the-art models. It was successfully tested on real production lines at the University of Verona.

Contact: uzair.khan@univr.it

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 74

DISENTANGLING MODALITY RELATIONS: A TWO-STAGE GRAPH FOR EMOTION RECOGNITION

Khan Mohammad Mohammed Rahman Sherif., Kumar Swagat., Behera Ardhendu

Abstract: Concurrently learning entangled inter- and intra-modal signals fundamentally limits current recognition models. Our two-stage graph network explicitly disentangles these signals, first mastering representations within each modality before modeling dependencies between them. This hierarchical strategy sets a new state-of-the-art on challenging emotion datasets, confirming the superiority of a disentangled approach.

Contact: Mohamm@edgehill.ac.uk

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 75

MEDICAL MULTI-VIEW GNN: ADVANCING TUMOR MALIGNANCY PREDICTION USING SPATIAL-AWARE DINOV2 REPRESENTATIONS

Kiechle J., Fischer S.M., Peeken J.C., Schnabel J.A.

Abstract: To date, no open-source foundation model has yet been trained specifically on oncologic MRI to capture features pertinent to precise tumor characterization. While DINOv2 shows promise in 2D medical image analysis, its appropriate application to 3D data remains unclear. We explore how DINOv2 can be adapted to 3D MRI while preserving its spatial structure. A GNN on a spherical graph is evaluated against non-linear MLPs for its ability to learn features predictive of tumor malignancy.

Contact: johannes.kiechle@tum.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 76

ON-SENSOR OPTICAL FLOW FOR ALWAYS-ON ROBOT VISION

Kim S., Kelly P., Davison A.

Abstract: Conventional vision pipelines inefficiently consume energy by transmitting raw pixels redundant and noisy for most downstream tasks. We advocate on-sensor processing: pixels generate high-level cues ready for downstream tasks. We instantiate this with a distributed optical flow algorithm where each pixel exchanges local beliefs to converge via Gaussian Belief Propagation. Tested on synthetic data under varied priors and messaging patterns, it offers design insights for next-gen pixel processors.

Contact: s.kim@imperial.ac.uk

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 77

UNI-DVPS: UNIFIED MODEL FOR DEPTH-AWARE VIDEO PANOPTIC SEGMENTATION

Ji-Yeon K., Hyun-Bin O., Byung-Ki K., Kim D., Kwon Y., Oh T.-H.

Abstract: We propose Uni-DVPS, a unified model for Depth-aware Video Panoptic Segmentation (DVPS) that jointly tackles fundamental vision tasks: 1) video panoptic segmentation, 2) monocular depth estimation, and 3) object tracking. The key idea is to design a single Transformer decoder network for multi-task learning, maximizing shared computation and promoting synergy between tasks while maintaining high efficiency. Our unified queries learn instance-level representations, enabling query-based tracking without requiring an extra tracking module.

Contact: jiyeon.kim@postech.ac.kr

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 78

NERSEMBLE: MULTI-VIEW RADIANCE FIELD RECONSTRUCTION OF HUMAN HEADS

KIRSCHSTEIN T., QIAN S., GIEBENHAIN S., WALTER T., NIESSNER M.

Abstract: We introduce a high-quality multi-view capture system with 16 cameras at 7.1 MP and 73 fps, used to collect a large-scale dataset of dynamic human head recordings from 424 subjects, totaling over 65 million frames across diverse demographics and expressions. The dataset is publicly released. We also present a benchmark for 3D head avatar research with two tasks: Dynamic Novel View Synthesis and Monocular FLAME Avatar Reconstruction, aiming to support progress in the field.

Contact: tobias.kirschstein@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 79

TACKLING DATA CHALLENGES IN DEEP LEARNING FOR ELECTRON MICROSCOPY

Kniesel H.

Abstract: Electron microscopy (EM) produces highly detailed images that are critical for advancing biological research, enabling visualization of cellular structures at nanometer resolution. However, the application of deep learning to EM data faces three major challenges: noisy data, scarce annotations, and expensive data acquisition. This work investigates effective strategies to address these challenges, aiming to support more reliable and efficient analysis of electron microscopy data in biological contexts through deep learning.

Contact: hannah.kniesel@uni-ulm.de

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 80

TOWARDS AUTONOMOUS MULTI-ROBOT EXPLORATION IN UNSTRUCTURED ENVIRONMENTS

Lasheras Hernández B., Giubilato R., Schuster M., Triebel R., Civera J.

Abstract: Planetary exploration increasingly relies on robotic systems as intelligent extensions of human astronauts, capable of operating with high-level guidance while maintaining significant autonomy. These systems must therefore be able to navigate and interpret complex, unstructured environments with low levels of human supervision. At the same time, multi-robot systems enhance redundancy, coverage, and resilience. However, autonomy is constrained by challenges in perception, decision-making, collaboration, and interaction. This research aims to advance autonomous exploration through open-set semantic scene understanding, knowledge representation, intelligent task allocation, and minimal human-in-the-loop interaction.

Contact: zuria.98@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 81

ENHANCING THE OLDEST ICE CLIMATE SIGNALS THROUGH SUPER-RESOLUTION IMAGING TECHNIQUES

Latif Hasan, Larkman Piers, Bohleber Pascal, Vascon Sebastiano

Abstract: Ice cores from polar regions are important archives of past climates. Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS) allows the acquisition of chemical images of the ice core. The usage of a bigger spot-size laser in LA-ICP-MS decreases the acquisition time but at the cost of the low-resolution chemical image. We explored several super-resolution techniques on chemical images, ranging from bicubic interpolation, CNNs, and diffusion models, showing promising results.

Contact: hasanlatif.pk@gmail.com

Presentation Type: Poster

Date: Monday 7 July 2025

Time: 21:30

Poster Session: 1

Poster Number: 82

ENIGMA-360: A MULTI-VIEW DATASET FOR HUMAN BEHAVIOR UNDERSTANDING IN INDUSTRIAL SCENARIOS

Ragusa F., Leonardi R., Mazzamuto M., Di Mauro D., Quattrocchi C., Passanisi A., D'Ambra I., Furnari A., Farinella G.M.

Abstract: In this paper we propose ENIGMA-360, a new multi-view dataset acquired in a real industrial scenario. The dataset is composed of 180 egocentric and 180 exocentric procedural videos temporally synchronized offering complementary information of the same scene. The 360 videos have been labeled with temporal and spatial annotations, enabling the study of different aspects of human behavior in industrial domain. We provide baseline experiments for 3 tasks: 1) Temporal Action Segmentation, 2) Keystep Recognition and 3) Egocentric Human-Object Interaction Detection, showing the limits of state-of-the-art approaches on this challenging scenario. We publicy release the dataset and its annotations at: https://iplab.dmi.unict.it/ENIGMA-360

Contact: rosario.leonardi@unict.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 83

PROBABILISTIC CONTRASTIVE LEARNING VIA REGULARIZED VON MISES-FISHER DISTRIBUTIONS

Li H. B., Ouyang C., Amiranashvili T., Rosen M., Menze B., Iglesias J. E.

Abstract: Traditional contrastive learning methods are deterministic and lack mechanisms to model uncertainty, limiting their reliability in high-stakes applications. We propose a probabilistic contrastive learning framework based on the von Mises-Fisher (vMF) distribution, which models representations as directional distributions on the hypersphere. To stabilize the learning dynamics of the vMF concentration parameter κ in high dimensions, we introduce a regularized vMF formulation with theoretical guarantees on equilibrium concentration dynamics. A novel probabilistic alignment loss integrates both the mean direction and concentration parameters of paired vMF embeddings, enabling adaptive alignment based on sample-level uncertainty. Empirically, our method captures both aleatoric and epistemic uncertainty, enhancing failure prediction and out-of-distribution detection across various corruption types and datasets. The proposed method preserves the geometric benefits of hyperspherical embeddings while providing a principled and computationally efficient route to uncertainty-aware representation learning.

Contact: holi2@mgh.harvard.edu

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 84

UNIMOTION: UNIFYING 3D HUMAN MOTION SYNTHESIS AND UNDERSTANDING

Li C., Chibane J., He Y., Pearl N., Geiger A., Pons-Moll G.

Abstract: Unimotion is the first model to jointly support global text and fine-grained frame-level text control for 3D human motion generation. It uniquely outputs frame-level text paired with the generated motion, enabling users to understand what happens and when. This allows applications such as hierarchical motion control, motion captioning from MoCap or video, and text-based motion editing. Unimotion achieves state-of-the-art results on the HumanML3D benchmark.

Contact: coralli1153410101@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 85

MEDBRIDGE: BRIDGING FOUNDATION VISION-LANGUAGE MODELS TO MEDICAL IMAGE DIAGNOSIS

Li Yitong, Ghahremani Morteza, Wachinger Christian

Abstract: Recent vision-language foundation models deliver state-of-the-art results on natural image classification but falter on medical images due to pronounced domain shifts. At the same time, training a medical foundation model requires substantial resources, including extensive annotated data and high computational capacity. To bridge this gap with minimal overhead, we introduce MedBridge, a lightweight multimodal adaptation framework re-purposing pretrained VLMs for accurate medical image diagnosis.

Contact: yi_tong.li@tum.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 86

FROG: FIELD ROBOTICS GROUP FOR REAL-WORLD 3D PERCEPTION SYSTEM

Li W., Fusaro D., Mosco S., Pretto A.

Abstract: Field RObotics Group develops 3D perception systems that are data-efficient, edge-compatible, and synthetic-driven. Avoiding dense annotations and heavy computation, we propose lightweight methods for LiDAR segmentation, traversability analysis, and sonar-based underwater localization. Our solutions achieve robust performance under unsupervised or minimally supervised settings, and run efficiently on low-power hardware, advancing practical 3D perception in real-world environments.

Contact: wanmeng.li@studenti.unipd.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 87

GCE-POSE: GLOBAL CONTEXT ENHANCEMENT FOR CATEGORY-LEVEL OBJECT POSE ESTIMATION

Li Weihang., Xu Hongli., Junwen Huang., Jung HyunJun., Yu Peter KT., Navab Nassir., Busam Benjamin.

Abstract: We present GCE-Pose, a method for category-level object pose estimation that uses global context priors. Our Semantic Shape Reconstruction module deforms category-specific 3D semantic prototypes to recover complete shape and semantics from partial RGBD input. A Global Context Enhanced fusion module then integrates observed and reconstructed features. GCE-Pose improves generalization to unseen instances and outperforms prior methods on HouseCat6D and NOCS-REAL275.

Contact: weihang.li@tum.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 88

MULTIMODALSTUDIO: A DATASET AND FRAMEWORK FOR MULTIMODAL NEURAL RENDERING

Lincetto F., Agresti G., Rossi M., Zanuttigh P.

Abstract: NeRF excels in rendering 3D scenes, but its capability to learn from different imaging modalities has seldom been explored. We present MultimodalStudio (MMS): it comprises MMS-DATA and MMS-FW. MMS-DATA is a multimodal multi-view dataset captured with 5 different modalities. MMS-FW is a novel NeRF framework that handles multimodal raw data. MMS-FW trained on MMS-DATA can transfer information between different modalities and produce higher quality renderings than using single modalities alone.

Contact: federico.lincetto@phd.unipd.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 89

INVERSE VIRTUAL TRY-ON: GENERATING MULTI-CATEGORY PRODUCT-STYLE IMAGES FROM CLOTHED INDIVIDUALS

Lobba D., Sanguigni F., Ren B., Cornia M., Cucchiara R., Sebe N.

Abstract: This paper presents Text-Enhanced Multi-category Virtual Try-Off, a novel architecture featuring a dual DiT backbone with a multimodal attention mechanism for robust garment feature extraction. Our architecture can receive garment information from multiple modalities to work in a multi-category setting. Finally, we propose an additional alignment module to refine the generated visual details. Experiments on VITON-HD and Dress Code datasets show that TEMU-VTOFF sets a new SOTA on the VTOFF task.

Contact: davide.lobba@unitn.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 90

ALLIGAT0R: PRE-TRAINING THROUGH CO-VISIBILITY SEGMENTATION FOR RELATIVE CAMERA POSE REGRESSION

Thibaut L., Guillaume B., Vincent L.

Abstract: We present Alligat0R, a novel pre-training method for binocular vision that segments pixels as covisible, occluded, or outside field-of-view. Unlike CroCo's cross-view completion which struggles with non-covisible regions, our approach learns effectively from all image areas. Combined with our Cub3 dataset of 5M annotated pairs, Alligat0R achieves superior performance on relative pose regression tasks.

Contact: thibaut.loiseau@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 91

DARK NOISE DIFFUISION: NOISE SYNTHESIS FOR LOW-LIGHT IMAGE DENOISING

Liying Lu, Raphaël Achddou, Sabine Süsstrunk

Abstract: Deep low-light image denoising networks require extensive training data. Manually collecting paired clean-noisy samples is labor-intensive. We propose using diffusion models to synthesize realistic camera noise and generate paired samples. Accurately model the mean and the variance of the camera noise is essential. We found that a two-branch network architecture and the noise schedule of the diffusion process play important roles.

Contact: liying.lu@epfl.ch

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 92

SPEQ: OFFLINE STABILIZATION PHASES FOR EFFICIENT Q-LEARNING IN HIGH UPDATE-TO-DATA RATIO REINFORCEMENT LEARNING

Romeo C.*, Macaluso G.*, Sestini A., Bagdanov A. D.

Abstract: High update-to-data (UTD) ratio algorithms improve sample efficiency but are computationally expensive. We introduce SPEQ, a RL method that combines low-UTD online training with periodic offline stabilization phases, where Q-functions are fine-tuned using a fixed replay buffer. This reduces redundant updates on poor data and balance between sample and compute efficiency. SPEQ achieves 40–99% fewer gradient updates and 27–78% less training time than SOTA methods while matching or exceeding their performance.

Contact: girolamo.macaluso@unifi.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 93

HM3: HIERARCHICAL MODELING OF MULTIMEDIA METAVERSES ON 10000 THEMATIC MUSEUMS VIA THEME-AWARE CONTRASTIVE LOSS FUNCTION

Macrì G., Bazzana L., Falcon A., Serra G.

Abstract: Task: Metaverse Retrieval, text-based retrieval of complex 3D scenes with multimedia elements. We focused on virtual art exhibitions.

Previous methods' shortcomings: - Small-scale, randomly aggregated datasets - Focus only on either images or videos, not both - Lack of thematic awareness in both datasets and methods

We address these issues with: - Large-scale thematic dataset with images & videos - Hierarchical method processing rooms' images and videos jointly - Theme-aware contrastive loss to guide cross-modal learning

Contact: gianluca18ma@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 94

METRIC-SEMANTIC 3D SCENE UNDERSTANDING

Maggio D.

Abstract: We present three works taking steps towards a unified scene representation using multi-view information to create an open-set metric-semantic map, enabling a spatial memory at the correct granularity to support an agent's tasks while being memory efficient and computationally tractable for real time deployment. Specifically, we present: (1) VGGT-SLAM - a dense transformer-based SLAM system, (2) Clio - a task-driven open-set real time mapping system, and (3) Bayesian Fields - an approach for aggregating multi-view open-set semantics into a task-driven photorealistic map.

Contact: drmaggio@mit.edu

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 95

FRED: THE FLORENCE RGB-EVENT DRONE DATASET

Magrini G., Marini N., Becattini F., Berlincioni L., Biondi N., Pala P., Del Bimbo A.

Abstract: We introduce the Florence RGB-Event Drone dataset (FRED), a novel multimodal dataset specifically designed for drone detection, tracking, and trajectory forecasting, combining RGB video and event streams. FRED features more than 7 hours of densely annotated drone trajectories, using 5 different drone models and including challenging scenarios such as rain and adverse lighting conditions.

Contact: gabriele.magrini@unifi.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 96

GROUND TRUTH-FREE FINE-TUNING HUMAN MOTION DIFFUSION MODELS WITH REINFORCEMENT LEARNING

Mandelli, Macaluso, Bicchierai

Abstract: We address key limitations in text-to-motion diffusion models—quality and controllability—by fine-tuning with reinforcement learning. Our method leverages external neural evaluators without ground-truth data and adapts RL to the temporal structure of human motion. We also introduce trajectory control, enabling characters to follow target paths in 3D space using diffusion-based generation.

Contact: lorenzo.mandelli@unifi.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 97

4DEFORM: NEURAL SURFACE DEFORMATION FOR ROBUST SHAPE INTERPOLATION

Sang L., Canfas Z., Cao D., Marin R., Bernard F., Cremers D.

Abstract: Generating a continuous 4D non-rigid deformation from just an initial and a final shape is challenging, especially for unstructured data (e.g., point clouds). In 4Deform, we rely on a modified level-set equation, linking a neural implicit representation with a velocity field. Such a connection enables geometrical and physical constraints on the deformation, letting us tackle noisy, partial, topology-changing, non-isometric shapes, enabling applications like 4D Kinect sequence up-sampling.

Contact: riccardo.marin@tum.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 98

CONTEXT-AWARE EVENT DETECTION, VIDEO UNDESTANDING WITH ROBUSTNESS TO CONTEXTUAL BIAS

Mattia Marseglia

Abstract: Understanding complex human and environmental activities in video streams is a key challenge for AI in safety, surveillance, and sustainability. This project advances video anomaly detection by moving beyond rigid label-based recognition, leveraging open-vocabulary and vision-language models to reason about actions and anomalies in a flexible, context-driven way. We aim to bridge visual content and language to enable efficient, real-world deployment.

Contact: mmarseglia@unisa.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 99

DOCWAVEDIFF

Marulli Matteo, Marco Bertini

Abstract: Document images often suffer from degradations such as blurring, unwanted text, and complex background textures, which reduce readability and impair automated processing like Optical Character Recognition (OCR). We present DocWaveDiff, a novel restoration and enhancement document image method based on a predict-and-refine strategy[1], addressing tasks such as inpainting, deblurring, and binarization.

Contact: matteo.marulli@unifi.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 100

GAZING INTO MISSTEPS: LEVERAGING EYE-GAZE FOR UNSUPERVISED MISTAKE DETECTION IN EGOCENTRIC VIDEOS OF SKILLED HUMAN ACTIVITIES

Mazzamuto M., Furnari A., Sato Y., Farinella G.M.

Abstract: We address the challenge of unsupervised mistake detection in egocentric video of skilled human activities through the analysis of gaze signals. While traditional methods rely on manually labeled mistakes, our approach does not require mistake annotations, hence overcoming the need of domainspecific labeled data. Based on the observation that eye movements closely follow object manipulation activities, we assess to what extent eye-gaze signals can support mistake detection, proposing to identify deviations in attention patterns measured through a gaze tracker with respect to those estimated by a gaze prediction model. Since predicting gaze in video is characterized by high uncertainty, we propose a novel gaze completion task, where eye fixations are predicted from visual observations and partial gaze trajectories, and contribute a novel gaze completion approach which explicitly models correlations between gaze information and local visual tokens. Inconsistencies between predicted and observed gaze trajectories act as an indicator to identify mistakes. Experiments highlight the effectiveness of the proposed approach in different settings, with relative gains up to +14%, +11%, and +5% in EPIC-Tent, HoloAssist and IndustReal respectively, remarkably matching results of supervised approaches without seeing any labels. We further show that gaze-based analysis is particularly useful in the presence of skilled actions, low action execution confidence, and actions requiring hand-eye coordination and object manipulation skills. Our method is ranked first on the HoloAssist Mistake Detection challenge.

Contact: michele.mazzamuto@phd.unict.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 101

TRAIN TILL YOU DROP: TOWARDS STABLE AND ROBUST SOURCE-FREE UNSUPERVISED 3D (AND IMAGE) DOMAIN ADAPTATION

Michele Björn, Boulch Alexandre, Vu Tuan-Hung, Puy Gilles, Marlet Renaud, Courty Nicolas

Abstract: We address source-free unsupervised domain adaptation (SFUDA) for semantic segmentation: only a source-trained model and unlabeled target data are available. Existing SFUDA methods often deteriorate with continued training. We mitigate this by regularizing the training and introducing an agreement-based criterion with a reference model that both stops training at the right moment and validates hyperparameters without target labels.

Contact: bjoern.michele@univ-ubs.fr

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 102

HUMAN MOTION UNLEARNING

De Matteis E., Migliarini M., Sampieri A., Spinelli I., Galasso F.

Abstract: Task: Erase violent motions from a motion generation model. Problem: Avoiding retraining while steering generation is hard, especially without hurting performance. Motivation: Motion synthesis models can produce toxic content. We want to remove it permanently from the model. Results: We remove harmful concepts while keeping performance on safe, unrelated actions intact.

Contact: matteo.migliarini@uniroma1.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 103

PATHOGEN CLASSIFICATION IN HYPERSPECTRAL DATA CUBES

Mikulić M., Štajduhar I.

Abstract: Skin and soft tissue infections, caused by microbial pathogens, present a clinical challenge due to diagnostic delays and increasing antimicrobial resistance. Current diagnostic standards can take weeks to identify the causative organism, which hinders timely treatment. [1] This study explores hyperspectral imaging as a method for rapid bacterial identification. We used a hyperspectral camera (400-1000 nm) to capture images of cultured Escherichia coli, Pseudomonas aeruginosa, Staphylococcus aureus, and Staphylococcus epidermidis. The data cubes were preprocessed with denoising, scatter correction and dimensionality reduction techniques. A three-dimensional convolutional neural network was trained on the preprocessed data cube patches, achieving a multiclass classification macro F1 score of 78% on test patches. These findings suggest this technique could be a step toward laboratory automation, enabling faster identification of pathogens in skin and soft tissue infections to guide therapeutic interventions.

Contact: mateo.mikulic@uniri.hr

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 104

REALFRED: AN EMBODIED INSTRUCTION FOLLOWING BENCHMARK IN PHOTO-REALISTIC ENVIRONMENT

Kim T.*, Min C.*, Kim B., Kim J., Jeung W., Choi J.

Abstract: Virtual environments for training robots have limitations: limited object interaction, unrealistic visuals, and small spaces, preventing real-world deployment. We propose ReALFRED, extending the ALFRED benchmark with real-world scenes, objects, and layouts in large, multi-room 3D environments. This reduces the visual domain gap between training and deployment. Analysis shows existing ALFRED methods perform worse in realistic settings, highlighting the need for better approaches.

Contact: cheolhong.min@snu.ac.kr

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 105

MONOCULAR DEPTH ESTIMATION IN ADVERSE CONDITIONS

Gasperini S.*, Morbitzer N.*, Jung H., Navab N., Tombari F.

Abstract: Monocular depth estimation methods struggle in adverse illumination and weather conditions. Therefore, we propose md4all, a simple yet effective solution, enhancing robustness by training models on complex augmented inputs while computing standard losses using corresponding original, ideal-condition images. Extensive experiments on nuScenes and Oxford RobotCar datasets show significant improvements over prior methods.

Contact: nils.morbitzer@tum.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 106

EXAMINING GRANULARITY FOR PRIVACY PROTECTION

Murrugarra-LLerena Jeffri., Niu Haoran., Barber K. Suzanne., Daumé III Hal., Trista Cao Yang., Cascante-Bonilla Paola.

Abstract: As visual assistant systems powered by visual language models (VLMs) become more prevalent, concerns over user privacy have grown, particularly for blind and low vision users who may unknowingly capture personal private information in their images. Existing privacy protection methods rely on coarse-grained segmentation, which uniformly masks entire private objects, often at the cost of usability. In this work, we propose FiG-Priv, a fine-grained privacy protection framework that selectively masks only high-risk private information while preserving low-risk information. Our approach integrates fine-grained segmentation with a data-driven risk scoring. We evaluate our framework using the BIV-Priv-Seg dataset and show that FiG-Priv preserves +26% of image content, enhancing the ability of VLMs to provide useful responses by +11% and identify the image content by +45%, while ensuring privacy protection.

Contact: ju.jeffri.v@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 107

NARRATIVEBRIDGE: ENHANCING VIDEO CAPTIONING WITH CAUSAL-TEMPORAL NARRATIVE

Nadeem, Asmar, Sardari Faegheh, Dawes, Robert, Husain Sameed Syed, Hilton, Adrian, Mustafa, Armin

Abstract: Existing video captioning benchmarks and models lack causal-temporal narrative, which is sequences of events linked through cause and effect, unfolding over time and driven by characters or agents. This lack of narrative restricts models' ability to generate text descriptions that capture the causal and temporal dynamics inherent in video content. To address this gap, we propose NarrativeBridge, an approach comprising of: (1) a novel Causal-Temporal Narrative (CTN) captions benchmark generated using a large language model and few-shot prompting, explicitly encoding cause-effect temporal relationships in video descriptions; and (2) a Cause-Effect Network (CEN) with separate encoders for capturing cause and effect dynamics, enabling effective learning and generation of captions with causal-temporal narrative. Extensive experiments demonstrate that CEN significantly outperforms state-of-the-art models in articulating the causal and temporal aspects of video content: 17.88 and 17.44 CIDEr on the MSVD-CTN and MSRVTT-CTN datasets, respectively. Cross-dataset evaluations further showcase CEN's strong generalization capabilities. The proposed framework understands and generates nuanced text descriptions with intricate causal-temporal narrative structures present in videos, addressing a critical limitation in video captioning. For project details, visit https://narrativebridge.github.io/.

Contact: asmar.nadeem@surrey.ac.uk

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 108

OCCLUSION AND CONFUSION: EVENT DETECTION AND PERFORMANCE ANALYSIS IN AMATEUR RUGBY FOOTAGE

Ní Dheoráin C.

Abstract: My research consists of analysing amateur rugby footage using Computer Vision to improve player performance and safety. This is challenging because of the complexity and ambiguity of objects of analysis in the sport (e.g, what exactly is a tackle?) and low-quality footage. This research requires large amounts of detailed labelled data which is time-consuming and costly to produce. I am currently focusing on approaches to automate the labelling which would be beneficial to all rugby research.

Contact: nidheorc@tcd.ie

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 109

AVHBENCH: A CROSS-MODAL HALLUCINATION BENCHMARK FOR AUDIO-VISUAL LARGE LANGUAGE MODELS

Sung-Bin K., Hyun-Bin O., Jung-Mok L., Senocak A., Chung J.S., Oh T.-H.

Abstract: Audio-Visual LLMs struggle to discern subtle relationships between audio and visual signals, leading to hallucinations and highlighting the need for reliable benchmarks. To address this, we introduce AVHBench, the first comprehensive benchmark that includes tests for assessing hallucinations, as well as the cross-modal matching and reasoning abilities of these models. We also demonstrate that simple training with our AVHBench improves robustness of audio-visual LLMs against hallucinations.

Contact: hyunbinoh@postech.ac.kr

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 110

SAMFUSION: SENSOR-ADAPTIVE MULTIMODAL FUSION FOR 3D OBJECT DETECTION IN ADVERSE WEATHER

Palladin E., Dietze R., Narayanan P., Bijelic M., Heide F.

Abstract: Multimodal sensor fusion is an essential capability for autonomous robots, enabling object detection and decision-making in the presence of failing or uncertain inputs. While recent fusion methods excel in normal environmental conditions, these approaches fail in adverse weather, e.g., heavy fog, snow, or obstructions due to soiling. We introduce a novel multi-sensor fusion approach tailored to adverse weather conditions. In addition to fusing RGB and LiDAR sensors, which are employed in recent autonomous driving literature, our sensor fusion stack is also capable of learning from NIR gated camera and radar modalities to tackle low light and inclement weather. We fuse multimodal sensor data through attentive, depth-based blending schemes, with learned refinement on the Bird's Eye View (BEV) plane to combine image and range features effectively. Our detections are predicted by a transformer decoder that weighs modalities based on distance and visibility. We demonstrate that our method improves the reliability of multimodal sensor fusion in autonomous vehicles under challenging weather conditions, bridging the gap between ideal conditions and real-world edge cases. Our approach improves average precision by 17.2 AP compared to the next best method for vulnerable pedestrians in long distances and challenging foggy scenes.

Contact: edoardo.palladin@torc.ai

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 111

LENGTH-AWARE MOTION SYNTHESIS VIA LATENT DIFFUSION

Sampieri A., Palma A., Spinelli I., Galasso F.

Abstract: We introduce the problem of generating length-aware 3D human motion sequences from text, and propose a novel model to synthesize motions of variable target lengths, which we dub "Length-Aware Latent Diffusion" (LADiff). It consists of two new modules: 1) a length-aware variational auto-encoder to learn motion representations with length-dependent latents; 2) a length-conforming latent diffusion model to generate motions with a richness of details that increases with the required target length.

Contact: a.palma@diag.uniroma1.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 112

TOWARDS EFFICIENT AND GENERIC STRUCTURE-FROM-MOTION

Linfei Pan

Abstract: We present advances towards a more efficient and generic Structure-from-Motion (SfM) pipeline. Our work includes: (1) a gravity-aligned rotation averaging method using circular regression for robust orientation estimation; (2) GLOMAP, a revisited global SfM pipeline that improves robustness and accuracy through efficient global positioning; and (3) GenSfM, a non-parametric SfM framework that handles diverse camera models via adaptive calibration. These components collectively enhance SfM's scalability, generality, and reconstruction quality across challenging scenarios.

Contact: linfei.pan@inf.ethz.ch

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 113

FLYSEARCH: EXPLORING HOW VISION-LANGUAGE MODELS EXPLORE

Pardyl, A., Matuszek, D., Przebieracz, M., Cygan, M., Zieliński, B., Wołczyk, M.

Abstract: The real world is messy and unstructured, requiring active, goal-driven exploration to uncover key information. We investigate whether Vision-Language Models can handle such conditions by introducing FlySearch - a 3D, outdoor, photorealistic environment for object search and navigation. Across three difficulty levels, we find that state-of-the-art VLMs struggle even with simple tasks. We identify key failure modes, including hallucinations, context errors, and planning issues.

Contact: adam.pardyl@doctoral.uj.edu.pl

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 114

STYLE-EDITOR: TEXT-DRIVEN OBJECT-CENTRIC STYLE EDITING

Park J., Gim J., Lee K., Lee S., Im S.

Abstract: We present a Text-driven object-centric style editing model named Style-Editor, a novel method that guides style editing at an objectcentric level using textual inputs. The core of Style-Editor is our Patchwise Co-Directional (PCD) loss, meticulously designed for precise object-centric editing that is closely aligned with the input text. This loss combines a patch directional loss for text-guided style direction and a patch distribution consistency loss for even CLIP embedding distribution across object regions. It ensures a seamless and harmonious style editing across object regions. Key to our method are the Text-Matched Patch Selection (TMPS) and Pre-fixed Region Selection (PRS) modules for identifying object locations via text, eliminating the need for segmentation masks. Lastly, we introduce an Adaptive Background Preservation (ABP) loss to maintain the original style and structural essence of the image's background. This loss is applied to dynamically identified background areas. Extensive experiments underline the effectiveness of our approach in creating visually coherent and textually aligned style editing.

Contact: pjh2857@dgist.ac.kr

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 115

BEYOND THE LIPS: ROBUST SPEAKER DETECTION VIA DISENTANGLED LATENT REPRESENTATIONS

Park J., Hong J., Kwon J

Abstract: Active Speaker Detection (ASD) determines if a person is speaking by analyzing audio and visual cues. Existing models rely on lip motion, making them vulnerable under occlusions. They add extra visual features or spatial context, increasing complexity. We propose an information-theoretic framework that disentangles speaker-relevant from irrelevant visual features using only face-region input. Our model, trained solely on one dataset, achieves strong cross-domain performance on another test set, outperforming prior methods under both clean and lip-masked conditions.

Contact: cosmopark624@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 116

HIERO: UNDERSTANDING THE HIERARCHY OF HUMAN BEHAVIOR ENHANCES REASONING ON EGOCENTRIC VIDEOS

Peirone S. A., Pistilli F., Averta G.

Abstract: Human activities have an underlying hierarchical structure that is mostly overlooked by deep learning models but can be used to better reason about them. Such structure can emerge naturally from unscripted videos of human activities, and can be leveraged to better reason about their content. We present HiERO, a weakly-supervised method to enrich video segment features with the corresponding hierarchical activity threads, achieving SOTA in video-language alignment and procedure learning tasks.

Contact: simone.peirone@polito.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 117

DEEP VISUAL ODOMETRY WITH EVENTS AND FRAMES

Pellerito R., Cannici M., Gehrig D., Belhadj J., Dubois-Matra O., Casasco M., Scaramuzza D.

Abstract: Visual Odometry (VO) is essential for navigation where GPS is unavailable, such as planetary surfaces. In this work, we present the first end-to-end learning-based VO system that fuses events and frames asynchronously. Our method, RAMP-VO, uses a novel Recurrent, Asynchronous, and Massively Parallel (RAMP) encoder to combine the strengths of both modalities. Despite being trained only in simulation, RAMP-VO achieves state-of-the-art performance on real-world and space-inspired datasets, offering a robust solution for low-light and high-speed motion scenarios.

Contact: rpellerito@ifi.uzh.ch

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 118

ADAPTING VISION TRANSFORMERS TO ULTRA-HIGH RESOLUTION SEMANTIC SEGMENTATION WITH RELAY TOKENS

Perron Y., Sydorov V., Pottier C., Landrieu L.

Abstract: Common segmentation methods use sliding-windows (losing global context) or downsampling (losing details). We propose a multi-scale method for vision transformers combining high-res details with global awareness. We process images at local (high-res) and global (low-res) scales simultaneously, exchanging info via relay tokens: learnable vectors aggregating features between scales. This adds ¡2% parameters and integrates with standard backbones (ViT, SWIN).

Contact: yohann.perron@enpc.fr

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 119

PRIVILEGED INFORMATION AND MULTIPLE SCLEROSIS LESION SEGMENTATION

Pignedoli V., Moro M., Noceti N., Odone F.

Abstract: This research investigates the Privileged Information paradigm to enhance machine learning algorithms with a focus on medical image analysis. In this framework training samples are of the form (x_i, x_i∧*, y_i), where x_i∧* is privileged information - available only during training. At inference time, x_i∧* is not accessible. Our goal is to study strategies for integrating PI into machine learning pipelines, with a focus on clinical imaging, where data scarcity make PI especially valuable.

Contact: veronica.pignedoli@edu.unige.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 120

RENDBEV: SEMANTIC PERSPECTIVE VIEW RENDERING AS SUPERVISION FOR BIRD'S EYE VIEW SEGMENTATION

Pineiro Monteagudo, H., Taccari L., Pjetri, A., Sambo, F. and Salti, S.

Abstract: The Bird's Eye View (BEV) is a widely used, compact representation for autonomous and assisted driving and robotics. GT labels for fully supervised training are expensive and difficult to get. RendBEV allows BEV segmentation models to be trained without labels or to work better with few labels.

Contact: henrique.pineiro2@unibo.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 121

HYPERBOLIC SAFETY-AWARE VISION-LANGUAGE MODELS

Poppi T., Kasarla T., Mettes P., Baraldi L., Cucchiara R.

Abstract: We propose HySAC, a vision-language model that enhances safety through hyperbolic embeddings. Instead of unlearning unsafe concepts, HySAC organizes them hierarchically, allowing redirection and classification. This enables moderation of unsafe content while retaining semantic structure.

Contact: tobia.poppi@unimore.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 122

PROBPOSE A PROBABILISTIC APPROACH TO 2D HUMAN POSE ESTIMATION

Purkrabek M., Matas J.

Abstract: Do not forget out-of-image keypoints when modeling people and evaluating pose.

Contact: purkrmir@fel.cvut.cz

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 123

WEATHEREDIT: CONTROLLABLE WEATHER EDITING WITH 4D GAUSSIAN FIELD

Qian C.*, Li W.†, Guo Y., Markkula G.

Abstract: In this work, we present: 1)WeatherEdit Framework: A framework for realistic and controllable weather generation in robotic simulation and vision tasks. 2) Multi-Weather Diffusion Adapter & Temporal-View Attention: An all-in-one adapter that enables a diffusion model to synthesize multiple weather types (snow, rain, fog), enhanced by a Temporal-View Attention mechanism for multi-frame and multi-view consistency. 3) 4D Gaussian Weather Field: A 4D Gaussian field-based weather model that simulates raindrops, snowflakes, and fog with high realism and controllable severity, enabling fine-grained weather manipulation.

Contact: tscq@leeds.ac.uk

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 124

TEST-TIME ADAPTATION FOR COMBATING MISSING MODALITIES IN EGOCENTRIC VIDEOS

Ramazanova M., Pardo A., Ghanem B., Alfarra M.

Abstract: Understanding videos that contain multiple modalities is crucial, especially in egocentric videos, where combining various sensory inputs significantly improves tasks like action recognition and moment localization. However, real-world applications often face challenges with incomplete modalities due to privacy concerns, efficiency needs, or hardware issues. Current methods, while effective, often necessitate retraining the model entirely to handle missing modalities, making them computationally intensive, particularly with large training datasets. In this study, we propose a novel approach to address this issue at test time without requiring retraining. We frame the problem as a test-time adaptation task, where the model adjusts to the available unlabeled data at test time. Our method, MiDl (Mutual information with self-Distillation), encourages the model to be insensitive to the specific modality source present during testing by minimizing the mutual information between the prediction and the available modality. Additionally, we incorporate self-distillation to maintain the model's original performance when both modalities are available. MiDl represents the first self-supervised, online solution for handling missing modalities exclusively at test time. Through experiments with various pretrained models and datasets, MiDl demonstrates substantial performance improvement without the need for retraining.

Contact: merey.ramazanova@kaust.edu.sa

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 125

ART2MUS: BRIDGING VISUAL ARTS AND MUSIC THROUGH CROSS-MODAL GENERATION

Rinaldi I., Fanelli N., Castellano G., Vessio G.

Abstract: Artificial Intelligence has transformed music creation via generative models.While text-to-music works well, complex visual inputs like digital artworks remain challenging. We propose Art2Mus, a model that generates music from artworks or text prompts by extending AudioLDM 2 [1]. We create new datasets using ImageBind [2], pairing digital artworks with music. Results show that Art2Mus produces music that aligns well with input stimuli, enabling new forms of multimedia art and creative tools.

Contact: i.rinaldi4@phd.uniba.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 126

SHOW OR TELL? A BENCHMARK TO EVALUATE VISUAL AND TEXTUAL PROMPTS IN SEMANTIC SEGMENTATION

Rosi G., Cermelli F.

Abstract: Prompt engineering excels in language models but is less explored in vision. For semantic segmentation, textual prompts enable open-vocabulary segmentation, while visual prompts offer intuitive references. Existing benchmarks test them separately. We introduce Show or Tell (SoT), evaluating both across 14 datasets in 7 domains. Our study finds textual prompts perform well on common concepts but struggle with complex ones, whereas visual prompts show variable but solid results, offering insights for future models.

Contact: gabriele.rosi@polito.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 127

TAKUNET: AN ENERGY-EFFICIENT CNN FOR REAL-TIME INFERENCE ON EMBEDDED UAV SYSTEMS IN EMERGENCY RESPONSE SCENARIOS

Rossi D., Borghi G., Vezzani R.

Abstract: Efficient embedded neural networks are essential for real-time drone-based emergency response imaging. TakuNet is a lightweight architecture using depthwise convolutions, early downsampling, dense connections and FP16 optimization to reduce complexity while retaining accuracy. On two public datasets, it achieves near–state-of-the-art accuracy, and when deployed on Jetson Orin Nano and RaspberryPi it yields high throughput, ¿650 FPS on Orin, enabling real-time inference on constrained platforms.

Contact: daniel.rossi@unimore.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 128

MAMBA-ST: STATE SPACE MODEL FOR EFFICIENT STYLE TRANSFER

Botti F., Ergasti A., Rossi L., Fontanini T., Ferrari C., Bertozzi M., Prati A.

Abstract: Style transfer task involves generating a new image that preserves the original content while applying a new style taken from another image. Most SOTA are Transformer-based models, which use cross-attention layers to transfer the style, resulting in a large memory footprint. To overcome the above limits, our work exploits Mamba, an emerging State-Space Model (SSM). By adapting Mamba's equations, we simulate cross-attention behavior while drastically reducing memory usage and time complexity.

Contact: leonardo.rossi@unipr.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 129

NEURALATEX: A MACHINE LEARNING LIBRARY WRITTEN IN PURE LATEX

Gardner J., Rowan W., Smith W.

Abstract: NeuRaLaTeX is a scalar values-based auto-grad library written entirely in LaTeX! You can specify the architecture of a neural network and loss functions, how to generate or load training data, and specify training hyperparameters and experiments. When the document is compiled, the LaTeX compiler will generate or load training data, train the network, run experiments and generate figures. Our evaluation shows NeuRaLaTeX to be the state-of-the-art in LaTeX-based machine learning libraries.

Contact: wrowan46@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 130

LAM3D: LEVERAGING ATTENTION FOR MONOCULAR 3D OBJECT DETECTION

Diana-Alexandra S., Leandro Di B., Yangxintong L., Florin O., Adrian M.

Abstract: Since the introduction of the self-attention mechanism and the adoption of the Transformer architecture for Computer Vision tasks, the Vision Transformer-based architectures gained a lot of popularity in the field, being used for tasks such as image classification, object detection and image segmentation. However, efficiently leveraging the attention mechanism in vision transformers for the Monocular 3D Object Detection task remains an open question. LAM3D is a framework that Leverages self-Attention mechanism for Monocular 3D object Detection. To do so, the proposed method is built upon a Pyramid Vision Transformer v2 (PVTv2) as feature extraction backbone and 2D/3D detection machinery. We evaluate the proposed method on the KITTI 3D Object Detection Benchmark, proving the applicability of the proposed solution in the autonomous driving domain and outperforming reference methods. Moreover, due to the usage of self-attention, LAM3D is able to systematically outperform the equivalent architecture that does not employ self-attention.

Contact: sas.cr.diana@student.utcluj.ro

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 131

EXPLORATION-DRIVEN GENERATIVE INTERACTIVE ENVIRONMENTS

Savov N., Kazemi N., Mahdi M., Paudel D.P., Wang X., Gool L.V.

Abstract: Modern multi-environment world models rely on costly unlabeled human interaction data. Instead, we label 974 virtual environments (RetroAct) and auto-collect a large amount of labeled interactions. To diversify data we introduce AutoExplore Agent, driven by world-model uncertainty. We build an open world model based on Genie - GenieRedux. We enhance it and adapt for virtual environments - our GenieRedux-G variant. Trained on the diverse data, we achieve high video fidelity and controllability.

Contact: nedko.savov@insait.ai

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 132

EFFICIENT ATTENTION VISION TRANSFORMERS FOR MONOCULAR DEPTH ESTIMATION ON RESOURCE-LIMITED HARDWARE

Schiavella C., Cirillo L., Papa L., Russo P., Amerini I.

Abstract: The attention module has a quadratic cost with respect to the processed tokens. In dense tasks like Monocular Depth Estimation, this poses challenges, especially in onboard applications. We use efficient attention to balance model quality and speed. Optimizations target each network module. The Pareto Frontier is used to find the best trade-off between modified models and baselines. Optimized networks often outperform baselines and improve inference speed.

Contact: schiavella@diag.uniroma1.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 133

IT'S A (BLIND) MATCH! TOWARDS VISION–LANGUA CORRESPONDENCE WITHOUT PARALLEL DATA

Schnaus Dominik, Araslanov Nikita, Cremers Daniel

Abstract: Can we match vision and language embeddings without any supervision? According to the platonic representation hypothesis, as model and dataset scales increase, distances between corresponding representations are becoming similar in both embedding spaces. Our study demonstrates that pairwise distances are often sufficient to enable unsupervised matching, allowing vision-language correspondences to be discovered without any parallel data.

Contact: dominik.schnaus@tum.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 134

POEM: PRECISE OBJECT-LEVEL EDITING VIA MLLM CONTROL

Schouten M., Kaya M.O., Belongie S., Papadopoulos D. P.

Abstract: Diffusion models have advanced text-to-image generation, yet object-specific editing remains challenging due to the need for precise modifications without disrupting global context. We propose POEM, a framework that leverages multimodal large language models to enable fine-grained, instruction-driven editing. POEM predicts pre- and post-edit object masks to guide a diffusion-based process. We introduce VOCEdits, a benchmark based on PASCAL VOC 2012, and demonstrate improved precision over prior methods

Contact: marscho@dtu.dk

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 135

OBJECTSPLAT: GENERALIZABLE OBJECT-CENTRIC 3D GAUSSIAN SPLATTING

Schröppel P., Wewer C., Ilg E., Lenssen J.E.

Abstract: Recent generalizable 3D Gaussian Splatting models reconstruct entire 3D scenes in a feed-forward fashion. However, for many downstream applications, 3D reconstructions with semantics and object decomposition are helpful. We introduce objectSplat, a feed-forward model that decomposes input views into objects and reconstructs semantic 3D Gaussian Splats for each object individually. The decomposition into objects is learned without any supervision. Instead, objectSplat builds on foundation models with rich semantic features spaces and strong 3D reconstruction capabilities and uses inductive biases that encourage clustering based on feature similarity and spatial proximity. objectSplat can work with single or multiple input views. We evaluate on CLEVR-3D and CLEVR-567. objectSplat delivers object decompositions and rendering quality on par with the state of the art, while achieving faster rendering speeds.

Contact: schroepp@cs.uni-freiburg.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 136

TASK GRAPH MAXIMUM LIKELIHOOD ESTIMATION FOR PROCEDURAL ACTIVITY UNDERSTANDING IN EGOCENTRIC VIDEOS

Seminara L., Farinella G. M., Furnari A.

Abstract: We present a novel framework for learning task graphs from action sequences using maximum likelihood estimation to enable a fully differentiable training objective. Our approach introduces two complementary methods: one based on direct optimization of the graph's adjacency matrix, and another leveraging key-step representations from text or video embeddings. We demonstrate the effectiveness of our task graph learning paradigm on the Ego-Exo4D Procedure Understanding Benchmark, where our method achieves state-of-the-art performance and ranks first in the official challenge.

Contact: luigi.seminara@phd.unict.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 137

BODY MEASUREMENT AND SURFACE RECONSTRUCTION IN MICROWAVE IMAGING

Miriam S.

Abstract: Accurate body dimension capture is crucial for both apparel and medical fields. Traditional methods, like measuring tapes, often lead to errors, while 3D laser scanners require tight-fitting or no clothing. Microwave Imaging (MI) presents a promising alternative, effectively capturing body surface data even when clothed. This paper explores overcoming MI's limitations to enable precise body measurement and surface reconstruction using an MI scanner, similar to the Rohde&Schwarz security scanner.

Contact: miriam.senne@tum.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 138

CANONICALFUSION: GENERATING DRIVABLE 3D HUMAN AVATARS FROM MULTIPLE IMAGES

Shin J., Lee J., Lee S., Park M., Kang J., Yoon J., Jeon H.

Abstract: CanonicalFusion is a framework for reconstructing animatable human avatars from multiple images. It predicts depth and compressed LBS weight maps via a shared-encoder-dual-decoder network to directly canonicalize 3D meshes. A forward skinning-based differentiable renderer merges results and refines the mesh by minimizing photometric and geometric errors, optimizing both vertex properties and joint angles to reduce pose-related artifacts.

Contact: jsshin98@gm.gist.ac.kr

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 139

INTERWEAVING INSIGHTS: HIGH-ORDER FEATURE INTERACTION FOR FINE-GRAINED VISUAL RECOGNITION (I2HOFI)

Sikdar A., Liu Y., Kedarisetty S., Zhao Y., Ahmed A., Behera A.,

Abstract: I2-HOFI is a graph-based framework for fine-grained visual recognition that captures both global (inter-region) and local (intra-region) feature interactions. It builds graphs connecting feature segments within and across object regions, then processes these graphs with a shared attention-enhanced GNN. Without requiring any part annotations, the model learns discriminative representations and achieves state-of-the-art accuracy on multiple fine-grained benchmarks.

Contact: sikdara@edgehill.ac.uk

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 140

PRADA: PROJECTIVE RADIAL DISTORTION AVERAGING

Sinitsyn D., Härenstam-Nielsen L., Cremers D.

Abstract: We tackle the problem of radially distorted camera self-calibration without solving full Structure-from-Motion or relying on less accurate deep networks. Working in projective space, a single homography absorbs all camera settings except the distortion term, letting us separate distortion estimation from the full 3-D reconstruction. We solve it by averaging pairwise distortion estimates — like pose averaging — so no 3-D points or bundle adjustments are needed while matching SfM accuracy.

Contact: daniil.sinitsyn@tum.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 141

TCC-DET: TEMPORARILY CONSISTENT CUES FOR WEAKLY-SUPERVISED 3D DETECTION

Skvrna Jan, Neumann Lukas

Abstract: Instead of paying costly annotators for 3D labels of vehicles, use TCC-Det to autolabel the dataset while losing less than 10% AP of a 3D Object Detector. If 10% of the human ground truth is added, the same AP is achieved compared to training with 100% ground truth. Our method uses an off-the-shelf 2D detector with temporal aggregation to extract dense point clouds representing vehicles, which are further used in direct optimisation of generic templates to acquire accurate 3D bounding boxes.

Contact: skvrnjan@fel.cvut.cz

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 142

CROSS-MODAL SEMANTIC GROUNDING EXPLOITING CONTEXT LEARNING FOR REAL-TIME VISUAL UNDERSTANDING

Spingola Camilla

Abstract: How can a vision system quickly adapt to unfamiliar environments? This research project explores cross-modal semantic grounding techniques, integrating context and task-adaptive information to guide attention, enhancing Vision Language Model (VLMs) for real-time, task-aware visual understanding.

Contact: cspingola@unisa.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 143

SEGMENTATION UNDER LOW-DATA CONSTRAINTS FOR NICHE DOMAINS

Sterzinger R., Sablatnig R.

Abstract: My PhD addresses segmentation in data scarcity for niche domains such as historical documents. Initial research (Etruscan Mirrors, Handwritten Manuscripts) reaffirmed data as the primary bottleneck. Subsequent work (Historical Maps) suggested leveraging foundation models, despite drastic appearance differences between historical artifacts and natural images, as semantic embeddings remain rich. Outlook: meta-learning with parameter-efficient fine-tuning to improve few-shot learning.

Contact: rsterzinger@cvl.tuwien.ac.at

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 144

WORLD MODEL PREDICTIVE CONTROL FOR INTERPRETABLE AUTONOMOUS DRIVING

Sun Jiangxin, Xue Feng, Long Teng, Sebe Nicu

Abstract: Despite the remarkable progress achieved in autonomous driving, most state-of-the-art methods follow the imitation learning pattern to clone driving behaviors from experts, introducing three critical challenges: heavy dependency on pre-trained experts leveraging privileged information, lack of an assessment mechanism for selected actions, and inability to perform fine-tuning through online exploration. To address these issues, we propose a novel predictive control framework called World Model Predictive Control (WMPC), where the control policy lies in a world model designed to anticipate near-future scene information conditioned on past observations and candidate actions. By explicitly evaluating the consequences of driving actions through these predictions, our method identifies the most suitable one, enhancing interpretability and reliability during both training and inference. In contrast to existing imitation learning methods that heavily depend on experts using privileged information, WMPC supports flexible learning through online exploration, offline expert demonstrations, and their combination, which mitigates the reliance on experts, improves sampling efficiency, and enables potential domain adaptation. Experimental results on closed-loop autonomous driving demonstrate that our approach achieves state-of-the-art performance compared with imitation learning methods while offering enhanced interpretability and online learning capability.

Contact: jiangxin.sun@unitn.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 145

BILLBOARDS SPLATTING (BBSPLAT): LEARNABLE TEXTURED PRIMITIVES FOR NOVEL VIEW SYNTHESIS

Svitov D., Morerio P., Agapito L., Del Bue A.

Abstract: BBSplat is based on textured planar primitives with learnable RGB textures and alpha-maps to control shape. The proposed primitives close the rendering quality gap between 2D and 3D Gaussian Splatting, enabling the accurate extraction of 3D mesh as in the 2DGS framework. The explicit planar primitives enables ray-tracing effects in rasterization. Novel regularization term encourages textures to have a sparser structure enabling an efficient compression leading to a storage space reduction.

Contact: david.svitov@iit.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 146

DIFFFNO: DIFFUSION FOURIER NEURAL OPERATOR

Xiaoyi Liu, Hao Tang

Abstract: We propose DiffFNO, a diffusion framework for arbitrary-scale image super-resolution powered by a Weighted Fourier Neural Operator (WFNO) with Mode Rebalancing for high-frequency detail recovery. Our Gated Fusion Mechanism combines WFNO's spectral features with spatial cues from an Attention-based Neural Operator. The Adaptive Time-Step ODE solver speeds up inference. DiffFNO achieves state-of-the-art PSNR gains (2–4 dB) and efficiency across scales, including those out of distribution.

Contact: bjdxtanghao@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 147

RGB AND IR FUSION FOR MULTIMODAL AERIAL TARGET DETECTION

Tavaris Denis, De Zan Alberto, Ivan Scagnetto, Gian Luca Foresti

Abstract: The identification of drones using advanced multimodal fusion techniques is increasingly relevant for security and surveillance. In this work, we developed a custom dataset containing RGB and thermal images of UAVs, aimed at improving detection models. We explored two fusion approaches: early fusion, where a single network is trained directly on combined RGT (RGB + Thermal) images, and late fusion, where separate models process RGB and thermal frames independently before merging the resulting bounding boxes. Finally, we tested these solutions on two distinct architectures: one relying on ground-based processing and the other performing onboard computation directly on the drone, analyzing their advantages and limitations.

Contact: 142438@spes.uniud.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 148

EGO-R1: CHAIN-OF-TOOL-THOUGHT FOR ULTRA-LONG EGOCENTRIC VIDEO REASONING

Tian Shulin* extsuperscript1, extsuperscript2, Wang Ruiqi extsuperscript1, extsuperscript3, Guo Hongming extsuperscript4, Wu Penghao extsuperscript1, Dong Yuhao extsuperscript1, Wang Xiuying extsuperscript1, Yang Jingkang extsuperscript1, Zhang Hao extsuperscript3, Zhu Hongyuan extsuperscript2, Liu Ziwei extsuperscript1

Abstract: Ego-R1 tackles day-to-week-long egocentric video reasoning with a Chain-of-Tool-Thought (CoTT) agent trained by reinforcement learning. The agent breaks each question into modular sub-steps and, at every step, invokes the optimal tool for temporal retrieval or multimodal grounding. Training combines supervised fine-tuning on Ego-CoTT-25K traces with RL on Ego-QA-4.4K question–answer pairs. We also release Ego-R1 Bench, a human-verified, week-scale video-QA benchmark. Experiments show that Ego-R1 extends reasoning horizons from a few hours to an entire week, setting a new state of the art for ultra-long egocentric video understanding.

Contact: shulin002@e.ntu.edu.sg

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 149

OPTIMIZING PROSTHETIC VISION USING VISION TRANSFORMERS

Tomas-Barba, J, Perez-Yus, A., Bermudez-Cameo, J.

Abstract: Prosthetic vision aims to restore sight in the visually impaired. However, suboptimal perceptions due to both physical constraints and anatomical characteristics, limit its effectiveness. We propose a novel neural network architecture that integrates a vision transformer with patient-specific information to optimize stimulation parameters and reduce perceptual distortions. The model shows improved performance over existing approaches and demonstrates its potential across various visual tasks.

Contact: j.tomas@unizar.es

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 150

ITERATIVE SELF-SUPERVISION FOR SPARSE VIEW NERF & GAUSSIAN SPLATTING

Felix Tristram, Stefano Gasperini, Nassir Navab, Federico Tombari

Abstract: NeRF and 3D-GS achieve great performance with dense inputs but degrade under sparsity due to geometry errors and artifacts. We propose Iterate Self-Supervision (ISS), which trains a first model to synthesize auxiliary views, applies uncertainty-aware masking to exclude unreliable regions, and augments training of a second model. ISS reinforces multi-view constraints, improves convergence without extra supervision, and yields significant gains on mip-NeRF 360, Tanks and Temples, and LLFF.

Contact: felix.tristram@tum.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 151

ACTMERGE: TASK-AWARE ACTIVATION INFORMED MODEL MERGING

Verasani M.

Abstract: ActMerge is a novel model merging framework that leverages activation signals from both the base and fine-tuned language models. Unlike prior approaches that rely only on base model activations, ActMerge captures task-specific saliency from expert activations to better preserve specialized capabilities. This task-aware strategy should enhance robustness and performance across diverse benchmarks, advancing the state of model merging.

Contact: 80934@studenti.unimore.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 152

UNIBEV: MULTI-MODAL 3D OBJECT DETECTION WITH UNIFORM BEV ENCODERS FOR ROBUSTNESS AGAINST MISSING SENSOR MODALITIES

Wang, Shiming; Caesar, Holger; Nan, Liangliang; Kooij, Julian F. P.

Abstract: Multi-sensor object detection is an active research topic in automated driving, but the robustness of such detection models against missing sensor input (modality missing), e.g., due to a sudden sensor failure, is a critical problem which remains under-studied. In this work, we propose UniBEV, an end-to-end multi-modal 3D object detection framework designed for robustness against missing modalities: UniBEV can operate on LiDAR plus camera input, but also on LiDAR-only or camera-only input without retraining. To facilitate its detector head to handle different input combinations, UniBEV aims to create well-aligned Bird's Eye View (BEV) feature maps from each available modality. Unlike prior BEV-based multi-modal detection methods, all sensor modalities follow a uniform approach to resample features from the native sensor coordinate systems to the BEV features. We furthermore investigate the robustness of various fusion strategies w.r.t. missing modalities: the commonly used feature concatenation, but also channel-wise averaging, and a generalization to weighted averaging termed Channel Normalized Weights. To validate its effectiveness, we compare UniBEV to state-of-the-art BEVFusion and MetaBEV on nuScenes over all sensor input combinations. In this setting, UniBEV achieves 52.5% mAP on average over all input combinations, significantly improving over the baselines (43.5% mAP on average for BEVFusion, 48.7% mAP on average for MetaBEV). An ablation study shows the robustness benefits of fusing by weighted averaging over regular concatenation, and of sharing queries between the BEV encoders of each modality. Our code is available at this https URL.

Contact: s.wang-15@tudelft.nl

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 153

3D RECONSTRUCTION WITH SPATIAL MEMORY

Wang H., Agapito L.

Abstract: We present Spann3R, a transformer-based model that performs feed-forward dense 3D reconstruction from uncalibrated images. The key idea is to manage a spatial memory that stores previous relevant 3D information. Spann3R then queries this spatial memory to reconstruct next frame in a global coordinate system

Contact: wanghengyi@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 154

LEARNING SPATIAL REPRESENTATINOS FOR EMBODIED PERCEPTION IN 3D WORLDS

Weijler L.

Abstract: Effective 3D perception requires: adapting to unfamiliar environments, maintaining spatial consistency under motion and understanding scene composition. Addressing these challenges, the three featured approaches explore test-time training for adaptable 3D segmentation, SE(3)-equivariant convolutions for motion-consistent spatial processing, and hyperbolic embeddings for hierarchical scene understanding—advancing perception for embodied intelligence.

Contact: lisa.weijler@tuwien.ac.at

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 155

TOWARDS SAFER AUTONOMOUS SYSTEMS: UNCERTAINTY QUANTIFICATION FOR REGRESSION

Xiong Z., Johnander J., Forssén P.-E.

Abstract: We present two papers in this poster. First, CATPlan is a lightweight module for collision risk prediction in end-to-end autonomous driving, improving detection by 54.8% on NeuroNCAP and nuScenes. Complementing this, we evaluate uncertainty quantification metrics for deep regression, recommending Calibration Error, AUSE, and NLL for reliable safety-critical AI. Together, these works advance trust and safety in autonomous systems.

Contact: ziliang.xiong@liu.se

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 156

DEMO: DENSE MOTION CAPTIONING FOR COMPLEX HUMAN MOTIONS

Shiyao Xu, Benedetta Liberatori, Gul Varol, Paolo Rota

Abstract: The focus of this paper is 3D dense motion captioning. Although there have been many works on dense video captioning, there is still a gap in human motion captioning, especially dense motion captioning. The challenges here include the lack of a large amount of complex motion data with corresponding dense captions and a clear definition of this task. We therefore design a generation protocol for complex 3D human motion data with dense captions, eliminate the tedious and unrealistic manual captioning, and create our complex motion dataset. CompMo is a synthetic dataset that contains 60k human motion sequences, each of which contains at least 2 10 different actions and lasts from 20s to 80s with 20fps. The descriptions in CompMo come from the human annotations of HumanML3D with our processing, consisting of 11,122 distinguished texts corresponding to the motion segment in a motion. With this dataset, we propose our Dense Motion Captioning with large language models to understand the input 3D motions, generate the corresponding dense captions, and establish a new benchmark on the CompMo evaluation set. Our work demonstrates the effectiveness of 3D dense motion captioning and explores a new way for human motion understanding.

Contact: shiyao.xu@unitn.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 157

3D-MOOD: LIFTING 2D TO 3D FOR MONOCULAR OPEN-SET OBJECT DETECTION

Yang Y.H., Pollefeys M.

Abstract: We address monocular 3D object detection in an open-set setting and introduce the first end-to-end 3D Monocular Open-set Object Detector (3D-MOOD). We propose to lift the open-set 2D detection into 3D space through our designed 3D bounding box head, enabling end-to-end joint training for both 2D and 3D tasks to yield better overall performance. We condition the object queries with geometry prior and overcome the generalization for 3D estimation across diverse scenes. We design the canonical image space for more efficient cross-dataset training. We achieve new state-of-the-art results on both closed-set settings (Omni3D) and open-set settings (Argoverse 2, ScanNet).

Contact: yunghsu.yang@inf.ethz.ch

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 158

SMALLGS GAUSSIAN SPLATTING-BASED CAMERA POSE ESTIMATION FOR SMALL-BASELINE VIDEO

Yao Y., Zhang Y, Huang Z., Lasenby J.

Abstract: Dynamic videos with small baselines are ubiquitous in daily life. However, it presents a challenge to existing camera pose estimation frameworks with ambiguous features and limited constraints. SmallGS focuses on camera pose estimation for small-baseline videos, which exploit the temporal consistency of Gaussian splatting within limited viewpoint change. We further incorporated DINOv2 into Gaussian splatting. We evaluated SmallGS in TUM-Dynamics sequences and achieved SOTA results.

Contact: yy561@cam.ac.uk

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 159

HOLOMOCAP: LOW-COST AUGMENTED REALITY MOTION CAPTURE WITH HOLOLENS 2

Zaccardi S., Jansen B.

Abstract: Low-cost, portable motion capture (MoCap) systems often lack the accuracy and real-time capabilities needed for clinical applications. To address these challenges, we present HoloMoCap [1], the first standalone MoCap application for HoloLens 2, enabling physiotherapists to receive real-time Augmented Reality (AR) feedback on patients' movements. Leveraging on-device Deep Learning (DL) and marker-based MoCap, HoloMoCap accurately estimates hip and knee angles, offering a practical solution for clinical motion analysis.

Contact: silvia.zaccardi@vub.be

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 160

CONTROLNET-XS: OPTIMISED IMAGE-BASED CONTROL FOR IMAGE SYNTHESIS

Zavadski D., Feiden JF., Rother C.

Abstract: The field of image synthesis has made tremendous strides forward in the last years. Besides defining the desired output image with text-prompts, an intuitive approach is to additionally use spatial guidance in form of an image, such as a depth map. In state-of-the-art approaches, spatial guidance of text-to-image diffusion models is often realised by a separate controlling model that controls a pre-trained image generation network. Understanding this process from a control system perspective shows that it forms a feedback-control system, where the control module receives a feedback signal from the generation process and sends a corrective signal back. When analysing existing systems, we observe that the feedback signals are timely sparse and have a small number of bits. As a consequence, there can be long delays between newly generated features and the respective corrective signals for these features. In this work, we take an existing controlling network, ControlNet [Zha23], and change the communication between the controlling network and the generation process to be of high-frequency and with large-bandwidth. By doing so, we are able to considerably improve the quality of the generated images, as well as the fidelity of the control. Also, the controlling network needs noticeably fewer parameters and hence is about twice as fast during inference and training time. We call our proposed network ControlNet-XS. When comparing with the state-of-the-art approaches, we outperform them for pixel-level guidance, such as depth, canny-edges, and semantic segmentation, and are on a par for loose keypoint-guidance of human poses.

Contact: denis.zavadski@iwr.uni-heidelberg.de

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 161

MARSLAM: MORE ACCURATE & ROBUST SLAM

Zhu Z.

Abstract: Simultaneous Localization and Mapping (SLAM) is a core challenge in robotics and computer vision, enabling autonomous agents to perceive and navigate their environment. We present three works aimed at achieving More Accurate & Robust SLAM. NICE-SLAM introduces dense 3D reconstruction and camera tracking with RGB-D input by integrating neural implicit representations. NICER-SLAM removes the dependency on depth sensors by incorporating monocular cues. WildGS-SLAM enables monocular Gaussian Splatting SLAM in dynamic environments.

Contact: zhuzihan2000@gmail.com

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 162

TOWARDS PERSONALIZED EMBODIED AI AGENTS

Filippo Ziliotto, Jelin Akkara Raphael, Lamberto Ballan, Luciano Serafini, Tommaso Campari

Abstract: We introduce the task of personalized embodied object grounding in pre-mapped indoor environments, where an embodied agent must resolve a user's natural-language query (e.g., "Find James's backpack") to precise real-world coordinates. To support this, we extend the GOAT-Bench dataset by incorporating over 50 HM3D home environments, each annotated with rich, narrative scene descriptions and hundreds of user-specific query episodes. Our method leverages a frozen BLIP-2 text encoder to process both the global scene description and the user query, which are then fused with a precomputed 2D spatial feature map via a multi-head attention mechanism to construct a personalized context map. Object locations are predicted by computing cosine similarity between the query embedding and the feature vector at each 1×1 m grid cell. By releasing our extended dataset and proposed architecture, we aim to foster further research in user-aware embodied AI.

Contact: filippo.ziliotto@phd.unipd.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 163

GG-SSMS: GRAPH-GENERATING STATE SPACE MODELS

Zubić N., Scaramuzza D.

Abstract: State Space Models (SSMs) excel at 1D sequential data but struggle with high-dimensional inputs like images. Traditional SSMs use fixed, one-dimensional scanning, limiting their ability to capture non-local interactions. Even recent methods (Mamba, Vim, VMamba) use predetermined scan paths, failing to adapt to complex, data-inherent structures. We solve this problem by introducing the Graph-Generating State Space Models (GG-SSMs).

Contact: zubic@ifi.uzh.ch

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 164

EXPLOITING ADVERSARIAL LEARNING AND TOPOLOGY AUGMENTATION FOR OPEN-SET VISUAL RECOGNITION

Zuccarà R., Fagetta G., Ortis A., Battiato S.

Abstract: This work aims to modify and enrich the feature space in which classes are represented, thereby improving the separability of known classes and enhance the model's robustness to out-of-distribution (OOD) inputs. The characteristic descriptor of a sample is defined as the probability distribution vector produced by the model. The topological enrichment of the feature space is achieved by introducing a new class, referred to as Neutral, whose ideal descriptor is represented by a uniform distribution, such that it cannot be confidently associated with any known class, thus inducing maximum uncertainty in the classifier. The Neutral class is synthetically generated using a custom-designed system that integrates the NEAT technique (NeuroEvolution of Augmenting Topologies), an evolutionary algorithm for the automatic generation of artificial neural networks. A fitness function, named FNEAT, has been implemented to guide the evolution of the networks in producing patterns suitable for the intended objective. The t-SNE technique is used to visualize, in a three-dimensional space, the probability distribution vectors obtained in supervised classification scenarios, both in the context of closed-set and open-set recognition.

Contact: rosa.zuccara@phd.unict.it

Presentation Type: Poster

Date: Tuesday 8 July 2025

Time: 21:30

Poster Session: 2

Poster Number: 165