# POSTER SESSION BOOKLET


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

University of Catania - University of Cambridge

International Computer Vision Summer School 2023

*From Perception to Action*

Sicily, 9 - 15 July 2023

## 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 2023.

***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 2023***

***Roberto Cipolla***
***Sebastiano Battiato***
***Giovanni Maria Farinella***

## List of Posters [1]

1. MONITORING SHORTCUT LEARNING USING MUTUAL INFORMATION Adnan Mohammed., Ioannou Yani., Tsai Kenyon., Galloway Angus., Tizhoosh Hamid., Taylor Graham.

2. QUALITY ASSESSMENT OF ENHANCED VIDEOS GUIDED BY AESTHETICS AND TECHNICAL QUALITY ATTRIBUTES Agarla Mirko, Celona Luigi, Rota Claudio, Schettini Raimondo

3. ZERO-SHOT COMPOSED IMAGE RETRIEVAL WITH TEXTUAL INVERSION Baldrati A., Agnolucci L., Bertini M., Del Bimbo A.

4. 3D SCENE PERCEPTION FROM SINGLE TO MULTI-MODALITIES Ahmad J.

5. PLAYING REPEATED GAMES WITH LARGE LANGUAGE MODELS Akata E., Schulz L., Coda-Forno J., Oh S.J., Bethge M., Schulz E.

6. A KALMAN VARIATIONAL AUTOENCODER MODEL ASSISTED BY ODOMETRIC CLUSTERING FOR VIDEO FRAME PREDICTION AND ANOMALY DETECTION Slavic Giulia., Alemaw Abrham., Marcenaro Lucio., Martin David., Regazzoni Carlo

7. LOOKING AT WORDS AND POINTS WITH ATTENTION: A BENCHMARK FOR TEXT-TO-SHAPE COHERENCE Amaduzzi A., Lisanti G., Salti S., Di Stefano L.

8. SAOR: SINGLE-VIEW ARTICULATED OBJECT RECONSTRUCTION AYGUN MEHMET, MAC AODHA OISIN

9. BELFUSION: LATENT DIFFUSION FOR BEHAVIOR-DRIVEN HUMAN MOTION PREDICTION Barquero G., Escalera S., Palmero C.

10. DISTANCE MATTERS FOR IMPROVING OUT-OF-DISTRIBUTION PERFORMANCE ESTIMATION Bernhardt M., Glocker B.

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

11. A LIGHT TOUCH APPROACH TO TEACHING TRANSFORMERS MULTI-VIEW GEOMETRY Bhalgat Y., Henriques J., Zisserman A.

12. PHYSICS-INFORMED NEURAL NETWORK FOR SEISMIC IMAGE SEGMENTATION Boillot L., Thouvenot A.

13. HERO: A MULTI-MODAL APPROACH ON MOBILE DEVICES FOR VISUAL-AWARE CONVERSATIONAL ASSISTANCE IN INDUSTRIAL DOMAINS Bonanno C., Ragusa F., Furnari A., Farinella G.M.

14. REHEARSAL CONTINUAL LEARNING IN REAL WORLD SCENARIOS Bonicelli L.

15. LEVERAGING TRIPLET LOSS FOR UNSUPERVISED ACTION SEGMENTATION Bueno-Benito E.,Tura B., Dimiccoli M.

16. VARITEX: VARIATIONAL NEURAL FACE TEXTURES BUEHLER MARCEL, MEKA ABHIMITRA, LI GENGYAN, BEELER THABO, HILLIGES OTMAR

17. SURFACE RECONSTRUCTION BENCHMARK FOR OBJECT POSE ESTIMATION Burde Varun, Benbihi Assia, Sattler Torsten, Burget Pavel

18. EXPLOITING PROXIMITY-AWARE TASKS FOR EMBODIED SOCIAL NAVIGATION Cancelli Enrico, Campari Tommaso, Serafini Luciano, Chang Angel, Ballan Lamberto

19. MITIGATING BIAS WITHOUT DEMOGRAPHICS Capitani G., Bolelli F., Porrello A., Calderara S., Ficarra E.

20. AN EDGE ML SOLUTION FOR AUTOMATIC AND ON-FIELD PARASITE EGG DETECTION Capuozzo S., Marrone S., Sansone C.

21. NEURIPS'22 CROSS-DOMAIN METADL CHALLENGE: RESULTS AND LESSONS LEARNED Carrión-Ojeda D., Escalera S., Guyon I., Mohr F., Nguyen MH., Roth S., Schaub-Meyer S., Ullah I. & Vanschoren J.

22. VIDEO NEURAL ARCHITECTURE SEARCH BY GRAPH GENERATION AND META PERFORMANCE RANKING Casarin S.

23. RESILIENT AND DISTRIBUTED MULTI - ROBOT VISUAL SPATIAL - PERCEPTION SYSTEMS Chang, Yun

24. FACTOR FIELDS: A UNIFIED FRAMEWORK FOR NEURAL FIELDS AND BEYOND Chen Anpei, Xu Zexiang Xu, Wei Xinyue, Tang Siyu, Su Hao, Geiger Andreas

25. TRAINING-FREE LAYOUT CONTROL WITH CROSS-ATTENTION GUIDANCE Chen M., Laina I., Vedaldi A.

26. IMAGE-FREE CLASSIFIER INJECTION FOR ZERO-SHOT CLASSIFICATION Christensen A., Mancini M., Koepke A. S., Winther O., Akata Z.

27. FORECASTING FUTURE INSTANCE SEGMENTATION WITH LEARNED OPTICAL FLOW AND WARPING Ciamarra A., Becattini F., Seidenari L., Del Bimbo A.

28. LEVERAGING AI FOR ANOMALY DETECTION AND ASSET INVENTORY IN INDUSTRIAL INFRASTRUCTURES Corrochano J., Sampedro C., Cordón S.

29. LEARNING DEPTH ESTIMATION FOR TRANSPARENT AND MIRROR SURFACES Costanzino A., Zama Ramirez P., Poggi M., Tosi F., Mattoccia S., Di Stefano L.

30. MULTIMODAL MOTION CONDITIONED DIFFUSION MODEL FOR SKELETON-BASED VIDEO ANOMALY DETECTION Flaborea A., Collorone L., D'Amely G., D'Arrigo S., Prenkaj B., Galasso F.

31. WEAKLY-SUPERVISED DOMAIN ADAPTIVE SEMANTIC SEGMENTATION WITH PROTOTYPICAL CONTRASTIVE LEARNING Das A., Xian Y., Dai D., Schiele B.

32. LEARN THE FORCE WE CAN: MULTI-OBJECT VIDEO GENERATION FROM PIXEL-LEVEL INTERACTIONS Davtyan A., Favaro P.

33. LOGIT-GUIDED FEATURE-BASED KNOWLEDGE DISTILLATION Dell'Eva A., Orsingher M., Lee L., Bertozzi M.

34. POSESCRIPT: 3D HUMAN POSES FROM NATURAL LANGUAGE Delmas G., Weinzaepfel P., Lucas T., Moreno-Noguer F., Rogez G.

35. CONTENT-AWARE TOKEN SHARING FOR EFFICIENT SEMANTIC SEGMENTATION WITH VISION TRANSFORMERS Lu C., de Geus D., Dubbelman G.

36. FORENSIC DEEPFAKE DETECTION USING PROTOTYPES De Leeuw den Bouter M.L., Lloret Pardo J., Geradts Z., Worring M.

37. BEST PRACTICES FOR 2-BODY POSE FORECASTING Rahman M.R.U., Scofano L., De Matteis E., Flaborea A., Sampieri A., Galasso F.

38. SPHERE-GUIDED TRAINING OF NEURAL IMPLICIT SURFACES Dogaru A., Ardelean AT., Ignatyev S., Zakharov E., Burnaev E.

39. AVATARS GROW LEGS: GENERATING SMOOTH HUMAN MOTION FROM SPARSE TRACKING INPUTS WITH DIFFUSION MODEL Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, Artsiom Sanakoyeu

40. A DATASET FOR DEXTEROUS BIMANUAL HAND-OBJECT MANIPULATION Fan Z., Taheri O., Tzionas D., Kocabas M., Kaufmann M., Black M., Hilliges O.

41. SCENEGENIE: SCENE GRAPH GUIDED DIFFUSION MODELS Farshad A., Yeganeh Y., Chi Y., Shen C., Ommer B., Navab N.

42. COMPOSITIONAL AVATAR CREATION FROM IMAGERY Feng Yao

43. UNCERTAINTY-AWARE VISION-BASED METRIC CROSS-VIEW GEOLOCALIZATION Fervers F., Bullinger S., Bodensteiner C., Arens M., Stiefelhagen R.

44. OVTRACK: OPEN-VOCABULARY MULTIPLE OBJECT TRACKING Li S., Fischer T., Ke L., Ding H., Danelljan M., Yu F.

45. DO ALL CNNS LEARN THE SAME FILTERS? A LARGE-SCALE STUDY Gavrikov P., Keuper J.

46. DOC2GRAPH: A TASK AGNOSTIC DOCUMENT UNDERSTANDING FRAMEWORK BASED ON GRAPH NEURAL NETWORKS Gemelli A., Biswas S., Civitelli E., Lladós J. and Marinai S.

47. HOOD: HIERARCHICAL GRAPHS FOR GENERALIZED MODELLING OF CLOTHING DYNAMICS Artur Grigorev, Bernhard Thomaszewski, Michael J. Black, Otmar Hilliges

48. THE FLORENCE INTERPERSONAL RELATION VIDEO DATASET Guerdelli H., Ferrari C., Berretti S., Del Bimbo A.

49. THE BENEFITS & LIMITATIONS OF PSEUDO LABEL ADAPTATION ON EPIC-KITCHENS-100 Guerrier R., Plizzari C., Damen D., Perrett T.

50. MULTI-PERSON HUMAN MOTION PREDICTION GUO W., DU Y., Bie X., Shen X., Lepetit V., Alameda-Pineda X., Moreno-Noguer F.

51. CONTINUAL LEARNING FOR THE ICUB HUMANOID ROBOT – AN EMPIRICAL ANALYSIS Gupta S., Biggio B.

52. HUMAN-OBJECT INTERACTION CAPTURE FROM WEARABLE SENSORS Guzov V., Chibane J., Marin R., He Y., Sattler T., Pons-Moll G.

53. AUTONOMOUS DRIVING AS A SEQUENCE TO SEQUENCE PROBLEM Hamdan S., Barin M. R., Güney F.

54. HIGH-FIDELITY 3D HUMAN DIGITIZATION FROM SINGLE 2K RESOLUTION IMAGES Han S., Park M., Yoon J., Kang J., Park Y., Jeon H.

55. GESTURE-SYNC: DETERMINING WHO IS SPEAKING WITHOUT A TALKING HEAD Hegde Sindhu, Zisserman Andrew

56. EUREKA-MOMENTS IN TRANSFORMERS: MULTI-STEP TASKS REVEAL SOFTMAX INDUCED OPTIMIZATION PROBLEMS Hoffmann David T., Schrodi Simon, Behrmann Nadine, Fischer Volker, Brox Thomas

57. WHERE ARE MY NEIGHBORS? EXPLOITING PATCHES RELATIONS IN SELF-SUPERVISED VISION TRANSFORMER Camporese G., Izzo E., Ballan L.

58. EXPLORING SSL STABILITY VIA CENTER VECTORS Jha Abhishek, Asano Yuki M., Tuytelaars Tinne

59. TETRAHEDRAL DIFFUSION MODELS FOR 3D SHAPE GENERATION Kalischek N., Peters T., Wegner D., Schindler K.

60. ANALYSIS OF CONTROLLABILITY AND FAIRNESS IN DIFFUSION MODELS Kara O., Yurtseven S., Yesiltepe H., Stojanov S., Rehg J

61. FAITHFUL TEXT-TO-IMAGE GENERATION BY SELECTION Karthik S., Roth K., Mancini M., Akata Z.

62. THE CALTECH FISH COUNTING DATASET: A BENCHMARK FOR MULTIPLE-OBJECT TRACKING AND COUNTING Kay J., Kulits P., Stathatos S., Deng S., Young E., Beery S., Van Horn G., Perona P.

63. OSSO: OBTAINING SKELETAL SHAPE FROM OUTSIDE Keller M., Zuffi S., Pujades S., Black M.

64. SOUND TO VISUAL SCENE GENERATION BY AUDIO-TO-VISUAL LATENT ALIGNMENT Sung-Bin K., Senocak A., Ha H., Owens A., Oh T-H.

65. CUBOIDS REVISITED: LEARNING ROBUST 3D SHAPE FITTING TO SINGLE RGB IMAGES Kluger F., Ackermann H., Brachmann E., Yang M., Rosenhahn B.

66. NEURAL FIELDS FOR LATENT FORCE FIELD DISCOVERY IN INTERACTING SYSTEMS Kofinas M., Bekkers E. J., Nagaraja N. S., Gavves E.

67. TETRA-NERF: REPRESENTING NEURAL RADIANCE FIELDS USING TETRAHEDRA Kulhanek J., Sattler T.

68. AUTONOMOUS NANO-DRONES Lorenzo Lamberti

69. ADAS: A DIRECT ADAPTATION STRATEGY FOR MULTI-TARGET DOMAIN ADAPTATION Lee S., Choi W., Kim C., Choi M., Im S.

70. MONOCULAR REACTIVE COLLISION AVOIDANCE FOR MAV TELEOPERATION WITH DEEP REINFORCEMENT LEARNING Brilli R., Legittimo M., Crocetti F., Leomanni M., Fravolini M. L., Costante G.

71. EXPLOITING MULTIMODAL SYNTHETIC DATA FOR EGOCENTRIC HUMAN-OBJECT INTERACTION DETECTION IN AN INDUSTRIAL SCENARIO Leonardi R., Ragusa F., Furnari A., Farinella G.M.

72. MATCH, EXPAND AND IMPROVE: UNSUPERVISED FINETUNING FOR ZERO-SHOT ACTION RECOGNITION WITH LANGUAGE KNOWLEDGE Lin Wei, Karlinsky Leonid, Shvetsova Nina, Possegger Horst, Kozinski Mateusz, Panda Rameswar, Feris Rogerio, Kuehne Hilde, Bischof Horst

73. COUNT-LEVEL WEAKLY-SUPERVISED DENSITY ESTIMATION FOR CROWD COUNTING Litrico M., Cheng F., Pound M., Tsaftaris S. A., Battiato S., Giuffrida M. V.

74. SEGMENTATION OF THE INFERIOR ALVEOLAR CANAL Lumetti L., Bolelli F., Grana C.

75. GENERATION OF A PSEUDO-REFERENCE DATA FOR A SUPERVISED DENOISING IN DIFFUSION MRI Machnio J., Pieciak T.

76. SELF SUPERVISED MASKED CONVOLUTIONAL TRANSFORMER BLOCK FOR ANOMALY DETECTION Madan Neelu, Ristea Nicolae-Catalin, Ionescu Tudor Radu, Nasrollahi Kamal, Khan Shahbaz Fahad, Moeslunf B. Thomas, Shah Mubarak

77. TACKLING WITH DATA HETEROGENEITY IN MULTI-DOMAIN FEDERATED LEARNING Madni Hussain Ahmad, Foresti Gian Luca

78. TRIPLETTRACK: 3D OBJECT TRACKING USING TRIPLET EMBEDDINGS AND LSTM Marinello N., Proesmans M., Van Gool L.

79. A MASKED POSE APPROACH FOR MOTION RETARGETING Martinelli G., Garau N., Bisagno N., Conci N.

80. EVENT-BASED CAMERA POSE TRACKER BY $\nabla$T NERF Mana Masuda, Yusuke Sekikawa, Hideo Satio

81. LEARNING TO DETECT ATTENDED OBJECTS IN CULTURAL SITES WITH GAZE SIGNALS AND WEAK OBJECT SUPERVISION Mazzamuto M., Ragusa F., Furnari A., Farinella G. M.

82. DATA-DRIVEN FEATURE TRACKING FOR EVENT CAMERAS Messikommer N.*, Fang C.*, Gehrig M., Scaramuzza D.

83. IMPACT OF DIGITAL FACE BEAUTIFICATION IN BIOMETRICS Mirabet-Herranz N., Galdi C., Dugelay J-L.

84. LAFTER: LABEL-FREE TUNING OF ZERO-SHOT CLASSIFIERS USING LANGUAGE AND UNLABELED IMAGE COLLECTIONS Mirza M. J., Leonid K., Lin W., Kozinski M., Possegger H., Feris R., Bischof H.

85. DUAL CONDITION DIFFUSION MODELS FOR OOD DETECTION Mishra D, Zhao He, Saha P, Noble J.A

86. DEEP FEATURES FOR COLONOSCOPIC SLAM Morlana, Javier

87. ACCURATE 3D BODY SHAPE REGRESSION USING METRIC AND SEMANTIC ATTRIBUTES Choutas Vasileios, Müller Lea, Huang Chun-Hao Paul, Tang Siyu, Tzionas Dimitrios, Black Michael J.

88. ASSEMBLYHANDS: TOWARDS EGOCENTRIC ACTIVITY UNDERSTANDING VIA 3D HAND POSE ESTIMATION Ohkawa T., He K., Sener F., Hodan T., Tran L., and Keskin C.

89. LINE FEATURES: FROM DETECTION TO APPLICATIONS Pautrat R., Suárez I., Lin J., Yu Y., Barath D., Oswald M., Larsson V., Pollefeys M.

90. DOMAIN GENERALIZATION TO CLASSIFY 14TH-CENTURY ITALIAN PUNCHMARKS Peaslee W. E., Cheng Y., Aviles-Rivero A. I., Wrapson L., Schönlieb C. B.

91. LEARNING TO RECONSTRUCT AND UNDERSTAND THE 3D WORLD Songyou Peng

92. SUPER-RESOLUTION 3D HUMAN SHAPE FROM A SINGLE LOW-RESOLUTION IMAGE Pesavento M., Volino M., Hilton A.

93. HANDWRITTEN TEXT GENERATION FROM VISUAL ARCHETYPES Pippi V., Cascianelli S., Cucchiara R.

94. INTERACTION IN MIXED REALITY ENVIRONMENTS Pizzo M., Solari F., Chessa M.

95. SELF-SUPERVISED ACCIDENT ANTICIPATION FROM DASHCAM VIDEOS Pjetri A., Abbondandolo D., Coimbra de Andrade D., Caprasecca S., Sambo F., Bagdanov A. D.

96. MACHINE UNLEARNING FOR EXPLAINABILITY AND SUSTAINABILITY Poppi S., Sarto S., Cornia M., Baraldi L., Cucchiara R.

97. DINER: DEPTH-AWARE IMAGE-BASED NEURAL RADIANCE FIELDS Prinzler Malte, Hilliges Otmar, Thies Justus

98. USE OF DEEP LEARNING FOR 3D FETAL CARDIAC MRI RECONSTRUCTION Prokopenko D., Hammernik K., Roberts T., Lloyd D., Rueckert D., Hajnal J.

99. THERMAL IMAGE SYNTHESIS Qazi Tayeba., Lall Brejesh.

100. PUT YOUR PPE ON: A TOOL FOR SYNTHETIC DATA GENERATION AND RELATED BENCHMARK IN CONSTRUCTION SITE SCENARIOS Quattrocchi C., Di Mauro D., Furnari A., Lopes A., Moltisanti M., Farinella G. M.

101. STILLFAST: AN END-TO-END APPROACH FOR SHORT-TERM OBJECT INTERACTION ANTICIPATION Ragusa F., Furnari A., Farinella G. M.

102. VISUAL DNA: REPRESENTING AND COMPARING IMAGES USING DISTRIBUTIONS OF NEURON ACTIVATIONS Ramtoula B., Gadd M., Newman P., De Martini D.

103. HIERARCHICAL AVERAGE PRECISION TRAINING FOR PERTINENT IMAGE RETRIEVAL Ramzi E., Audebert N., Thome N., Rambour C., Bitot X.

104. USING EXPLANATIONS TO GUIDE MODELS Rao S., Böhle M., Parchami-Araghi A., Schiele B.

105. TOWARDS DATA-DRIVEN REGULARIZATION IN ABDOMINAL IMAGE REGISTRATION Reithmeir A.,Zimmer V.A., Schnabel J.A.

106. VOLRECON: VOLUME RENDERING OF SIGNED RAY DISTANCE FUNCTIONS FOR GENERALIZABLE MULTI-VIEW RECONSTRUCTION Ren Y., Wang F., Zhang T., Pollefeys M., Süsstrunk S.

107. PLANT: EXPLAINABLE PLANNING TRANSFORMERS VIA OBJECT-LEVEL REPRESENTATIONS Renz K., Chitta K., Mercea O.B., Koepke A.S., Akata Z., Geiger A.

108. MOVING: A MODULAR AND FLEXIBLE PLATFORM FOR EMBODIED VISUAL NAVIGATION Rosano M., Ragusa F., Furnari A., Farinella G.M.

109. EGOCENTRIC AUDITORY ATTENTION LOCALIZATION IN CONVERSATIONS Ryan F., Jiang H., Shukla A., Rehg J.M., Ithapu V.

110. TIME DOES TELL, A LOT: SELF-SUPERVISED LEARNING OF DENSE IMAGE REPRESENTATIONS Mohammadreza Salehi, Efstratios Gavves, Cees G. M. Snoek, Yuki M. Asano

111. NLX-GPT: A MODEL FOR NATURAL LANGUAGE EXPLANATIONS IN VISION AND VISION-LANGUAGE TASKS Sammani, Fawaz., Mukherjee Tanmoy., Deligiannis Nikos

112. IMPLICIT SHAPE COMPLETION VIA ADVERSARIAL SHAPE PRIORS Saroha Abhishek, Eisenberger Marvin, Yenamandra Tarun, Cremers Daniel

113. COMPUTATIONAL DESIGN OF FABRICABLE GEOMETRIC PATTERNS Scandurra E., Callieri M., Giorgi D., Laccone F., Malomo L., Cignoni P.

114. DISTRACTING DOWNPOUR: ADVERSARIAL WEATHER ATTACKS FOR MOTION ESTIMATION Schmalfuss J., Mehl L., Bruhn A.

115. LIFELONG LOOP CLOSURE DETECTION Scucchia M.

116. AN AR-BASED TOOL FOR ACQUISITION AND AUTOMATIC LABELING OF HUMAN-OBJECT INTERACTIONS FROM FIRST PERSON VISION Seminara L., Ragusa F., Leonardi R., Farinella G. M., Furnari A.

117. MATERIALISTIC: SELECTING SIMILAR MATERIALS IN IMAGES Sharma P., Philip J., Gharbi M., Freeman B., Durand F., Deschaintre V.

118. GAZENERF: 3D-AWARE GAZE REDIRECTION WITH NEURAL RADIANCE FIELDS Ruzzi A., Shi X., Wang X., Li G., De Mello S., Chang H. J., Zhang X., Hilliges O.

119. VIEWSET DIFFUSION: 0 TO N IMAGE-CONDITIONED 3D GENERATION FROM 2D DATA Szymanowicz S., Rupprecht C., Vedaldi A.

120. ANOMALY DETECTION IN INDUSTRIAL PROCESSES Taylor Alex

121. FAST AND EASY REGISTRATION OF MICROSCOPY TIME-LAPSE DATA WITH NAPARI-ROI REGISTRATION PLUGIN Giorgia Tortora (1), Alessia Candeo (2), Andrea Bassi (3), Stefano Buratti (4), Matteo Grenzi (5), Alex Costa (6)

122. INCORPORATING DIAGNOSTIC PRIOR WITH SEGMENTATION: A DEEP LEARNING PIPELINE FOR THE AUTOMATIC CLASSIFICATION OF AUTOIMMUNE BULLOUS SKIN DISEASES Pastore V P, Touijer L, Capurro N, Cozzani E, Gasparini G, Parodi A, Odone F

123. SPARF: NEURAL RADIANCE FIELDS FROM SPARSE AND NOISY POSES Truong Prune, Rakotosaona Marie-Julie, Manhardt Fabian, Tombari Federico

124. AUTOMATED MONITORING OF IVD DEVICE ANALYSIS PROCESSES USING MACHINE LEARNING AND COMPUTER VISION TECHNIQUES Tufo G.

125. ANOMALY DETECTION FOR VISUAL INSPECTION Ugwu C. I.

126. SCADE: NERFS FROM SPACE CARVING WITH AMBIGUITY-AWARE DEPTH ESTIMATES Uy Mikaela Angelina, Martin-Brualla Ricardo, Guibas Leonidas, Li Ke

127. ALIGNMIXUP: IMPROVING REPRESENTATIONS BY INTERPOLATING ALIGNED FEATURES Venkataramanan Shashanka, Kijak Ewa, Amsaleg Laurent, and Avrithis Yannis

128. COVR: LEARNING COMPOSED VIDEO RETRIEVAL FROM WEB VIDEO CAPTIONS Ventura L., Yang A., Schmid C., Varol G.

129. POP-3D: OPEN-VOCABULARY 3D OCCUPANCY PREDICTION FROM IMAGES Vobecky A., Siméoni O., Hurych D., Gidaris S., Bursuc A., Pérez P., Sivic J.

130. DISCO: DISENTANGLED CONTROL FOR HUMAN DANCE GENERATION Wang T., Li L., Lin K., Lin C., Yang Z., Zhang H., Liu Z., Wang L.

131. POSEDIFFUSION: SOLVING POSE ESTIMATION VIA DIFFUSION-AIDED BUNDLE ADJUSTMENT Wang J., Rupprecht C., Novotny D.

132. SYMNP: LEARNING SYMMETRIES BETWEEN NEURAL POINTS Wewer C., Ilg E., Schiele B., Lenssen J.

133. LEARNING DEPTH FROM FOCUS IN THE WILD Won C., Jeon H.

134. SEGMENTING MOVING OBJECTS VIA AN OBJECT-CENTRIC LAYERED REPRESENTATION Xie J., Xie W., Zisserman A.

135. FINDING AND NAVIGATING TO HUMANS IN COMPLEX ENVIRONMENTS FOR ASSISTIVE TASKS Yaar A., Furnari A., Rosano M., Härmä A., Farinella GM

136. VID2SEQ: LARGE-SCALE PRETRAINING OF A VISUAL LANGUAGE MODEL FOR DENSE VIDEO CAPTIONING Yang Antoine, Nagrani Arsha, Seo Paul, Miech Antoine, Pont-Tuset Jordi, Laptev Ivan, Sivic Josef, Schmid Cordelia

137. TEXTMANIA: ENRICHING VISUAL FEATURES BY TEXT-DRIVEN MANIFOLD AUGMENTATION Ye-Bin M., Kim J., Kim H., Son K., Oh T.-H.

138. BEYOND PIXELS Yousef Yeganeh

139. SYNTHESIZING DIVERSE HUMAN MOTIONS IN 3D INDOOR SCENES Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, Siyu Tang

140. EXPLOITING STABLE GRASPS FOR HAND-OBJECT RECONSTRUCTION IN EGOCENTRIC VIDEOS Zhu Z., Damen D.

141. AI-BASED MONITORING SYSTEM FOR ENHACING INDUSTRIAL PROCESSES: A FOCUS ON VIALS INSPECTION Zribi M., Pagliuca P.

# MONITORING SHORTCUT LEARNING USING MUTUAL INFORMATION

Adnan Mohammed., Ioannou Yani., Tsai Kenyon., Galloway Angus., Tizhoosh Hamid., Taylor Graham.

**Abstract:** The failure of deep neural networks to generalize to out-of-distribution data is a well-known problem and raises concerns about the deployment of trained networks in safety-critical domains such as healthcare, finance and autonomous vehicles. We study a particular kind of distribution shift – shortcuts or spurious correlations in the training data. In this work, we propose to use the mutual information (MI) between the learned representation and the input as a metric to find shortcut learning.

**Contact:** mohammed.adnan1@ucalgary.ca

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 1

# QUALITY ASSESSMENT OF ENHANCED VIDEOS GUIDED BY AESTHETICS AND TECHNICAL QUALITY ATTRIBUTES

Agarla Mirko, Celona Luigi, Rota Claudio, Schettini Raimondo

**Abstract:** In this work we propose a novel method to evaluate the quality of enhanced videos Our approach involves the use of three deep learning models that encode video sequences in terms of overall technical quality quality related attributes and aesthetic quality The resulting feature vectors are combined and used as input to a SVR to estimate the video quality score Quantitative results on the recently released VQA Dataset the NTIRE 2023 demonstrates the effectiveness of the proposed method

**Contact:** m.agarla@campus.unimib.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 2

# ZERO-SHOT COMPOSED IMAGE RETRIEVAL WITH TEXTUAL INVERSION

Baldrati A., Agnolucci L., Bertini M., Del Bimbo A.

**Abstract:** Composed Image Retrieval (CIR) aims to retrieve a target image based on a query composed of a reference image and a relative caption that describes the difference between the two images. In this work, we propose a new task, Zero-Shot CIR (ZS-CIR), that aims to address CIR without requiring a labeled training dataset. Our approach, named SEARLE, maps the visual features of the reference image into a pseudo-word token in CLIP token embedding space and integrates it with the relative caption.

**Contact:** lorenzo.agnolucci@unifi.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 3

# 3D SCENE PERCEPTION FROM SINGLE TO MULTI-MODALITIES

Ahmad J.

**Abstract:** 3D object detection and localization within the desired scene, either from crowdsourced images or onboard multimodal sensors, has a wide range of interesting applications such as creating maps, safer traffic environments, city planning, and self-driving. We present our works: multi-view 3D object localization from street-level scenes, and ongoing research activities on multimodal fusion for 3D object detection.

**Contact:** javed.ahmad@iit.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 4

# PLAYING REPEATED GAMES WITH LARGE LANGUAGE MODELS

Akata E., Schulz L., Coda-Forno J., Oh S.J., Bethge M., Schulz E.

**Abstract:** LLMs are transforming society and permeating into diverse applications. As a result, they will frequently interact with us and other agents. To understand how LLMs behave in interactive social settings, we propose to use behavioral game theory to study LLM's cooperation and coordination behavior. To do so, we let different LLMs (GPT-3, GPT-3.5, and GPT-4) play finitely repeated games with each other and with other, human-like strategies.

**Contact:** elif.akata@uni-tuebingen.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 5

# A KALMAN VARIATIONAL AUTOENCODER MODEL ASSISTED BY ODOMETRIC CLUSTERING FOR VIDEO FRAME PREDICTION AND ANOMALY DETECTION

Slavic Giulia., Alemaw Abrham., Marcenaro Lucio., Martin David., Regazzoni Carlo

**Abstract:** The fusion of sensory channels to be in a dynamic stability is an innate behavior of intelligent beings, which can be, imitated by artificial agents like Autonomous Vehicles. We propose a method for video-frame prediction and anomaly detection leveraging odometric data. A Bayesian framework is adopted and combined with Variational Autoencoder to learn an appropriate latent space representation. For inference and anomaly detection MJPF is built. iCab and Driveset datasets are used for evaluation.

**Contact:** abrham.shiferaw.alemaw@edu.unige.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 6

# LOOKING AT WORDS AND POINTS WITH ATTENTION: A BENCHMARK FOR TEXT-TO-SHAPE COHERENCE

Amaduzzi A., Lisanti G., Salti S., Di Stefano L.

**Abstract:** While text-conditional 3D object generation and manipulation have seen rapid progress, there is currently no clear benchmark for evaluating the coherence between generated 3D shapes and input textual descriptions. To address this, we propose the first benchmark for text-shape generation, which includes a dataset of 3D shapes with paired textual descriptions, a quantitative metric for the evaluation of text-shape coherence and a human-validated test set of text-shape pairs.

**Contact:** andrea.amaduzzi4@unibo.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 7

# SAOR: SINGLE-VIEW ARTICULATED OBJECT RECONSTRUCTION

AYGUN MEHMET, MAC AODHA OISIN

**Abstract:** SAOR can efficiently predict object pose, 3D shape reconstruction, and unsupervised part-level assignment using only a single forward pass per image at test time.

**Contact:** mehmet.aygun@ed.ac.uk

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 8

# BELFUSION: LATENT DIFFUSION FOR BEHAVIOR-DRIVEN HUMAN MOTION PREDICTION

Barquero G., Escalera S., Palmero C.

**Abstract:** Most prior stochastic human motion prediction (HMP) works aim at predicting highly coordinate-wise diverse motion. This has led to predictions of fast and divergent movements, often unrealistic and incoherent with past motion. BeLFusion addresses these issues by using a latent diffusion model to conditionally sample from a disentangled behavioral latent space. As a result, predictions display a variety of behaviors that are more realistic and coherent than the state of the art.

**Contact:** germanbarquero@ub.edu

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 9

# DISTANCE MATTERS FOR IMPROVING OUT-OF-DISTRIBUTION PERFORMANCE ESTIMATION

Bernhardt M., Glocker B.

**Abstract:** OOD performance estimation is a crucial component of safe AI deployment. Most current solutions use softmax confidences to derive accuracy estimates. However, these may become ill-calibrated for samples far from the training distribution. In this work, we show that taking into account distances of test samples to the training distribution can significantly improve accuracy estimation. We show the effectiveness of our method on 13 classification tasks, across a wide-range of distribution shifts, with SOTA performance on 10 out of 13 tasks.

**Contact:** m.bernhardt21@imperial.ac.uk

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 10

# A LIGHT TOUCH APPROACH TO TEACHING TRANSFORMERS MULTI-VIEW GEOMETRY

Bhalgat Y., Henriques J., Zisserman A.

**Abstract:** We propose a "light touch" approach to incorporate multi-view geometric priors into Vision Transformers. Epipolar lines are used to guide the Transformer's cross-attention maps, penalizing attention values outside the epipolar lines and encouraging higher attention along these lines since they contain geometrically plausible matches. Unlike previous methods, our method doesn't rely on camera pose information at test-time and it outperforms existing approaches in pose-invariant object retrieval.

**Contact:** yashsb@robots.ox.ac.uk

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 11

# PHYSICS-INFORMED NEURAL NETWORK FOR SEISMIC IMAGE SEGMENTATION

Boillot L., Thouvenot A.

**Abstract:** Seismic interpretation of geological units corresponds to the subsurface layers' detection and classification, colored by age ranges. Deep learning techniques based on local textures recognition struggle to ensure geological coherency, leading to the appearance of many tiny mispredicted areas like bubbles, see [1]. In this work, we proposed to add an unsupervised loss term that penalizes these bubbles using a geological constrained heuristic (aka PINN). In combination with a supervised training, the hybrid strategy drastically improves the overall geological perception.

**Contact:** lionel.boillot@gmail.com

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 12

# HERO: A MULTI-MODAL APPROACH ON MOBILE DEVICES FOR VISUAL-AWARE CONVERSATIONAL ASSISTANCE IN INDUSTRIAL DOMAINS

Bonanno C., Ragusa F., Furnari A., Farinella G.M.

**Abstract:** We present HERO, an artificial assistant designed to communicate with users with both natural language and images to aid them carrying out procedures in industrial contexts. We deployed and evaluated the system in an industrial laboratory. In this setting, our system allows the user to retrieve information on tools, equipment, and procedures. Experiments, as well as a user study, suggest that its use can be beneficial for users over classic methods for retrieving information and guide workers.

**Contact:** claudia.bonanno@phd.unict.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 13

# REHEARSAL CONTINUAL LEARNING IN REAL WORLD SCENARIOS

Bonicelli L.

**Abstract:** Continual Learning (CL) methods learn from a stream of tasks while avoiding catastrophic forgetting. Among these, rehearsal mitigates forgetting by storing and then replaying a subset of data from the past.

In real-world environments, rehearsal faces some challenges: (I) human labeling limits the amount of annotations available; (II) knowledge from samples previously stored should be updated as new relationships are learned; (III) pretrain needs to be carefully considered for future tasks to benefit from it; (IV) the repeated optimization of the buffer causes severe overfit.

**Contact:** lorenzo.bonicelli@unimore.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 14

# LEVERAGING TRIPLET LOSS FOR UNSUPERVISED ACTION SEGMENTATION

Bueno-Benito E.,Tura B., Dimiccoli M.

**Abstract:** This paper [1] presents a novel fully unsupervised framework for action segmentation. Our approach learns action representations from single videos without training data. By leveraging a shallow network, triplet loss, and a novel triplet selection strategy, we capture temporal and semantic priors for action discovery. Our method outperforms existing approaches, yielding higher-quality temporal boundaries. Evaluation of benchmark datasets shows competitive performance using a generic clustering algorithm.

**Contact:** ebueno@iri.upc.edu

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 15

# VARITEX: VARIATIONAL NEURAL FACE TEXTURES

BUEHLER MARCEL, MEKA ABHIMITRA, LI GENGYAN, BEELER THABO, HILLIGES OTMAR

**Abstract:** Deep generative models synthesize photorealistic faces of novel identities. A key step in making such models practically useful is to provide independent control over semantically meaningful parameters: appearance, head pose, face shape, and facial expressions. State-of-the-art generative models for faces fail to synthesize extreme head poses. We solve this problem using neural face textures. Our key insight is that UV mapping and neural textures allow synthesizing novel faces and rendering them in extreme poses and expressions while maintaining identity consistency, despite using datasets with limited pose distribution.

**Contact:** marcel.buehler@inf.ethz.ch

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 16

# SURFACE RECONSTRUCTION BENCHMARK FOR OBJECT POSE ESTIMATION

Burde Varun, Benbihi Assia, Sattler Torsten, Burget Pavel

**Abstract:** Object pose estimation is a crucial task in robotics. Many state-of-the-art object pose algorithms rely on a cad model of the object. This work aims to measure to what degree 3D models reconstructed from images can be used instead. We present a benchmark, based on the YCBV dataset, to evaluate state-of-the-art reconstruction algorithms. Early results show that for some objects, reconstructed 3D models perform close to cad models. For more complex objects, there is a considerable performance gap.

**Contact:** varun.burde@cvut.cz

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 17

# EXPLOITING PROXIMITY-AWARE TASKS FOR EMBODIED SOCIAL NAVIGATION

Cancelli Enrico, Campari Tommaso, Serafini Luciano, Chang Angel, Ballan Lamberto

**Abstract:** Learning how to navigate among humans in indoor environment is required to embodied agent to be integrated into our society. We propose an end-to-end architecture that exploits Proximity-Aware Tasks, called Risk and Proximity Compass, to inject into a reinforcement learning navigation policy the ability to infer social behaviors. Our tasks exploit the notion of immediate and future dangers of collision. We validate our approach on Gibson4+ and HM3D datasets.

**Contact:** enrico.cancelli@phd.unipd.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 18

# MITIGATING BIAS WITHOUT DEMOGRAPHICS

Capitani G., Bolelli F., Porrello A., Calderara S., Ficarra E.

**Abstract:** DNNs failures can result from biases in data or algorithmic choices. Existing debiasing methods rely on prior knowledge or information, which may be impractical. We propose a new debiasing approach that does not require external hints about biases. Our approach modifies the standard ERM by introducing per-example weights, indicating the significance of each example from the majority. These weights consider the model's difficulty in inferring the correct pseudo-label obtained through clustering.

**Contact:** giacomo.capitani@unimore.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 19

# AN EDGE ML SOLUTION FOR AUTOMATIC AND ON-FIELD PARASITE EGG DETECTION

Capuozzo S., Marrone S., Sansone C.

**Abstract:** Parasite egg detection in farm animals fecal samples is a veterinary medicine task that can be automated with CNNs with human-like performances. The Kubic FLOTAC Microscope (KFM) is a compact, low-cost, versatile and portable digital microscope designed to autonomously analyze samples in FLOTAC apparatuses, through a camera with an embedded AI solution, to detect eggs and generate clinical reports in a few minutes, even in field settings without electricity, for different parasites and hosts.

**Contact:** salvatore.capuozzo@unina.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 20

# NEURIPS'22 CROSS-DOMAIN METADL CHALLENGE: RESULTS AND LESSONS LEARNED

Carrión-Ojeda D., Escalera S., Guyon I., Mohr F., Nguyen MH., Roth S., Schaub-Meyer S., Ullah I. & Vanschoren J.

**Abstract:** Most of the evaluation settings for few-shot learning, focus only on the within-domain scenario. However, our NeurIPS'22 Cross-Domain MetaDL Competition challenged participants to generalize across domains. The competition yielded valuable insights, such as the critical role of pre-trained backbones, the necessity of preventing overfitting, and the importance of using data augmentation or domain adaptation techniques alongside extra optimizations to enhance performance.

**Contact:** dustin.carrion@visinf.tu-darmstadt.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 21

# VIDEO NEURAL ARCHITECTURE SEARCH BY GRAPH GENERATION AND META PERFORMANCE RANKING

Casarin S.

**Abstract:** This work tackles the limitations of task generalization and computational costs in Neural Architecture Search (NAS). It efficiently searches neural networks (NNs) for new datasets by leveraging a cross-modal latent space. The selection of NNs occurs without direct target dataset training, through a meta-performance ranking predictor. Impressively, it identifies top-performing NNs in the image domain within seconds. The performance is validated by preliminary results from the CVPR-NAS Challenge.

**Contact:** scasarin@unibz.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 22

# RESILIENT AND DISTRIBUTED MULTI - ROBOT VISUAL SPATIAL - PERCEPTION SYSTEMS

Chang, Yun

**Abstract:** This poster summarizes published and ongoing work on multi-robot visual spatial-perception systems. Kimera-Multi is a multi-robot visual SLAM system that is (i) is capable of rejecting incorrect loop closures from perceptual aliasing, (ii) is fully distributed and only relies on local communication to achieve distributed localization and mapping, and (iii) builds a globally consistent semantic 3D mesh of the environment in real-time. Hydra-Multi is an initial presentation of an ongoing effort on a multi-robot system capable of constructing a 3D scene graph online.

**Contact:** yunchang@mit.edu

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 23

# FACTOR FIELDS: A UNIFIED FRAMEWORK FOR NEURAL FIELDS AND BEYOND

Chen Anpei, Xu Zexiang Xu, Wei Xinyue, Tang Siyu, Su Hao, Geiger Andreas

**Abstract:** We present Factor Fields, a novel framework for modeling and representing signals. Factor Fields decomposes a signal into a product of factors, each represented by a classical or neural field representation which operates on transformed input coordinates. This decomposition results in a unified framework that accommodates several recent signal representations including NeRF, Plenoxels, EG3D, Instant-NGP, and TensoRF. Additionally, our framework allows for the creation of powerful new signal representations, such as the "Coefficient-Basis Factorization" (CoBaFa) which is a second contribution of this paper. Our experiments show that CoBaFa leads to improvements in approximation quality, compactness, and training time when compared to previous fast reconstruction methods. Experimentally, our representation achieves better image approximation quality on 2D image regression tasks, higher geometric quality when reconstructing 3D signed distance fields, and higher compactness for radiance field reconstruction tasks.

**Contact:** anpei.chen@inf.ethz.ch

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 24

# TRAINING-FREE LAYOUT CONTROL WITH CROSS-ATTENTION GUIDANCE

Chen M., Laina I., Vedaldi A.

**Abstract:** In this paper, we present a method for controlling the layout of images generated by large pre-trained text-to-image models by guiding the cross-attention patterns produced by the model in a spatially-directed manner. Our method requires no further training or finetuning.

**Contact:** minghao@robots.ox.ac.uk

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 25

# IMAGE-FREE CLASSIFIER INJECTION FOR ZERO-SHOT CLASSIFICATION

Christensen A., Mancini M., Koepke A. S., Winther O., Akata Z.

**Abstract:** We pose the practical question: given a specific image classification task and a generic pre-trained model, can we extend it with missing categories without using images? We achieve this with our proposed ICIS framework. Our model injects classifiers for new, unseen classes into fixed classification models in a post-hoc fashion. This is done without image data. Instead, ICIS learns from existing classifier weights and simple class-wise descriptors, such as class names or attributes.

**Contact:** anders.christensen321@gmail.com

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 26

# FORECASTING FUTURE INSTANCE SEGMENTATION WITH LEARNED OPTICAL FLOW AND WARPING

Ciamarra A., Becattini F., Seidenari L., Del Bimbo A.

**Abstract:** For an autonomous vehicle it is essential to observe the ongoing dynamics of a scene and consequently predict imminent future scenarios to ensure safety to itself and others. In this paper we investigate the usage of optical flow for predicting future semantic segmentations. To do so, we design a model that forecasts flow fields autoregressively. Such predictions are then used to guide the inference of a learned warping function that moves instance segmentations on to future frames.

**Contact:** andrea.ciamarra@unifi.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 27

# LEVERAGING AI FOR ANOMALY DETECTION AND ASSET INVENTORY IN INDUSTRIAL INFRASTRUCTURES

Corrochano J., Sampedro C., Cordón S.

**Abstract:** Companies related to electrical and railway infrastructure allocate a large budget for maintenance. Expensive inspection processes capture vast amounts of data as images, videos or point clouds. These data are very heterogeneous as they can be captured with different devices, under any weather condition, assuming a challenge for AI-based solutions. To tackle this, unusuals proposes a system combining computer vision and deep learning for fully automated asset inventory and anomaly detection.

**Contact:** jcorrochano@unusuals.ai

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 28

# LEARNING DEPTH ESTIMATION FOR TRANSPARENT AND MIRROR SURFACES

Costanzino A., Zama Ramirez P., Poggi M., Tosi F., Mattoccia S., Di Stefano L.

**Abstract:** Inferring depth of transparent or mirror (ToM) surfaces from images represents a hard challenge. We propose a pipeline for learning to estimate depth for such surfaces with neural networks, without requiring ground-truth annotation. We unveil how to obtain reliable pseudo labels by in-painting ToM objects and processing them with a monocular depth estimation model. Such labels are used to fine-tune existing models, to let them learn how to deal with ToM surfaces, achieving dramatic improvements.

**Contact:** alex.costanzino@unibo.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 29

# MULTIMODAL MOTION CONDITIONED DIFFUSION MODEL FOR SKELETON-BASED VIDEO ANOMALY DETECTION

Flaborea A., Collorone L., D'Amely G., D'Arrigo S., Prenkaj B., Galasso F.

**Abstract:** We propose MoCoDAD, a novel generative model for video anomaly detection. We consider skeletal representations of human motion and take advantage of the improved mode coverage capabilities of a past motion conditioned diffusion model to generate different but plausible (multimodal) future motions. On the basis of the statistical aggregation of future modes, the abnormality is detected when the generated set of motions is not pertinent to the actual future.

**Contact:** stefano.darrigo@uniroma1.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 30

# WEAKLY-SUPERVISED DOMAIN ADAPTIVE SEMANTIC SEGMENTATION WITH PROTOTYPICAL CONTRASTIVE LEARNING

Das A., Xian Y., Dai D., Schiele B.

**Abstract:** There is a huge gap in performance of Unsupervised Domain Adaptive Semantic Segmentation and Supervised Learning. In this work, we propose a common framework to use different weak labels, e.g., image, point and coarse labels from the target domain to reduce this performance gap. Our experiments on various benchmarks show that our framework achieves significant improvement compared to existing works and can reduce the performance gap with supervised learning.

**Contact:** andas@mpi-inf.mpg.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 31

# LEARN THE FORCE WE CAN: MULTI-OBJECT VIDEO GENERATION FROM PIXEL-LEVEL INTERACTIONS

Davtyan A., Favaro P.

**Abstract:** We propose a novel unsupervised method to autoregressively generate videos from a single frame and a sparse motion input. Our trained model can generate realistic object-to-object interactions and separate the dynamics and the extents of multiple objects despite only observing them under correlated motion activities. Key components in our method are the randomized conditioning scheme, the encoding of the input motion control, and the randomized and sparse sampling to break correlations. Our model, which we call YODA, has the ability to move objects without physically touching them. We show both qualitatively and quantitatively that YODA accurately follows the user control, while yielding a video quality that is on par with or better than state-of-the-art video generation prior work on several datasets.

**Contact:** aram.davtyan@unibe.ch

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 32

# LOGIT-GUIDED FEATURE-BASED KNOWLEDGE DISTILLATION

Dell'Eva A., Orsingher M., Lee L., Bertozzi M.

**Abstract:** Knowledge distillation has emerged as a powerful technique for transferring knowledge from a complex teacher network to a more compact student network, improving its performance. We propose a novel approach that leverages the complementary strengths of feature-based and logit-based KD methods to enhance the knowledge transfer process. The features extracted from the student network are firstly aligned with a simple convolutional layer to match the teacher resolution and then fed to the corresponding subsequent layers of the frozen teacher network to compute a response-based loss. The logit-based loss for each branch acts as a regularizer, providing additional guidance to the feature-based loss.

**Contact:** anthony.delleva2@unibo.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 33

# POSESCRIPT: 3D HUMAN POSES FROM NATURAL LANGUAGE

Delmas G., Weinzaepfel P., Lucas T., Moreno-Noguer F., Rogez G.

**Abstract:** PoseScript is a dataset pairing 3D human poses with both automatically generated and human-written descriptions. We use it to train a text-to-pose retrieval model and a text-conditioned pose generative model. We show better performance on human data when pretraining on automatic descriptions generated by our captioning pipeline

**Contact:** gdelmas@iri.upc.edu

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 34

# CONTENT-AWARE TOKEN SHARING FOR EFFICIENT SEMANTIC SEGMENTATION WITH VISION TRANSFORMERS

Lu C., de Geus D., Dubbelman G.

**Abstract:** We introduce Content-aware Token Sharing (CTS), a token reduction approach that improves the computational efficiency of semantic segmentation networks that use Vision Transformers (ViTs). CTS first identifies neighboring image patches that contain the same semantic class, and then lets these redundant neighboring patches share a token. By doing so, CTS can reduce the number of tokens by up to 44% and improve the throughput by up to 110%, without decreasing the segmentation quality.

**Contact:** d.c.d.geus@tue.nl

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 35

# FORENSIC DEEPFAKE DETECTION USING PROTOTYPES

De Leeuw den Bouter M.L., Lloret Pardo J., Geradts Z., Worring M.

**Abstract:** The increasing fidelity of deepfake videos poses a serious risk for our society. Our work focuses on video forensics, where the aim is to make judgments on the authenticity of a video, which can be defended in court. Forensic investigators are expected to get confronted with large volumes of disputed material in the near future, due to the rapidly improving quality of deepfakes, combined with their proliferation and the accessibility of the tools to generate them. Therefore, we see potential for XAI-based methods that can aid the investigators in the detection of deepfake videos. We use DPNet, a prototype-based neural network, to classify videos as pristine or manipulated. Activation maps (created using interpolation or Prototypical Relevance Propagation) demonstrate which parts of the input video correspond most to each of the prototypes. While the results are promising, the prototypes could be improved in terms of explainability and correspondence to features most often used by human experts to determine whether a video is pristine or manipulated (e.g., face wobble, blurred features, overly smooth skin).

**Contact:** m.l.deleeuwdenbouter@uva.nl

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 36

# BEST PRACTICES FOR 2-BODY POSE FORECASTING

Rahman M.R.U., Scofano L., De Matteis E., Flaborea A., Sampieri A., Galasso F.

**Abstract:** Task: Predict future human poses of two interacting agents based on previous frames. Problem: Previous works primarily focused on predicting individual poses, neglecting collaborative scenarios with two agents. Motivation: Predicting interactions between two agents can lead to improved performance due to body-body motion correlations. Results: Up to 21.9% with respect to the state of the art. Code: https://github.com/edodema/BestPractices2Body Project Page: https://www.pinlab.org/bestpractices2body

**Contact:** edoardo.dematteis@uniroma1.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 37

# SPHERE-GUIDED TRAINING OF NEURAL IMPLICIT SURFACES

Dogaru A., Ardelean AT., Ignatyev S., Zakharov E., Burnaev E.

**Abstract:** Neural distance functions trained via volumetric ray marching have been widely adopted for multi-view 3D reconstruction. However, these methods apply the ray marching procedure for the entire scene volume, leading to reduced sampling efficiency. We address this problem by using a coarse sphere-based representation to exclude the empty volume of the scene from ray marching without additional forward passes, resulting in an increased fidelity of the reconstructions compared to the base systems.

**Contact:** andreea.dogaru@fau.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 38

# AVATARS GROW LEGS: GENERATING SMOOTH HUMAN MOTION FROM SPARSE TRACKING INPUTS WITH DIFFUSION MODEL

Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, Artsiom Sanakoyeu

**Abstract:** With the recent surge in popularity of AR/VR applications, a particular challenge is that only a sparse tracking signal is available from standalone HMDs (Head Mounted Devices). In this paper, we present AGRoL, a novel conditional diffusion model specifically designed to track full bodies given sparse upper-body tracking signals. Our model can predict accurate and smooth full-body motion, particularly the challenging lower body movement.

**Contact:** yuming.du@enpc.fr

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 39

# A DATASET FOR DEXTEROUS BIMANUAL HAND-OBJECT MANIPULATION

Fan Z., Taheri O., Tzionas D., Kocabas M., Kaufmann M., Black M., Hilliges O.

**Abstract:** We intuitively understand that inanimate objects do not move by themselves, but that state changes are typically caused by human. This is not yet the case for machines. This is because there is no dataset with groundtruth 3D data to study the consistent motion of hands and articulated objects. We present ARCTIC, a dataset of two hands manipulating articulated objects, and propose two tasks: consistent motion reconstruction and interaction field estimation, with corresponding baselines.

**Contact:** zicong.fan@inf.ethz.ch

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 40

# SCENEGENIE: SCENE GRAPH GUIDED DIFFUSION MODELS

Farshad A., Yeganeh Y., Chi Y., Shen C., Ommer B., Navab N.

**Abstract:** We propose a novel guidance approach for the sampling process in the diffusion model that leverages bounding box and segmentation map information at inference time without additional training data. Through a novel loss in the sampling process, our approach guides the model with semantic features and enforces geometric constraints, leading to high-resolution images that accurately represent the scene. To obtain bounding box and segmentation map information, we structure the text prompt as a scene graph and enrich the nodes with CLIP embeddings.

**Contact:** azade.farshad@tum.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 41

# COMPOSITIONAL AVATAR CREATION FROM IMAGERY

Feng Yao

**Abstract:** Inspired by Richard Feynman's quote, 'What I can not create, I do not understand,' my PhD research delves into the recreation of digital humans to gain a profound comprehension of ourselves. Two key observations shape my insights: 1) Our primary mode of observing others is through visual perception, and 2) Human beings are inherently compositional, with distinct elements such as face, hair and external features like clothing. Based on these observations, my research focuses on creating compositional avatars from real-world imagery. I begin by exploring 3D human face learning, developing a framework for generating realistic and animatable 3D faces from single images. Then, I construct expressive 3D human bodies, including detailed faces, bodies, and hands, with minimal clothing. Additionally, I investigate capturing human clothing and hair from videos. My research democratizes avatar creation, making it accessible for everyone to generate avatars from unconstrained images or videos. Furthermore, the compositional nature of these avatars enables realistic movement, clothing changes, and hairstyle modifications, replicating real-world experiences.

**Contact:** yao.feng@tuebingen.mpg.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 42

# UNCERTAINTY-AWARE VISION-BASED METRIC CROSS-VIEW GEOLOCALIZATION

Fervers F., Bullinger S., Bodensteiner C., Arens M., Stiefelhagen R.

**Abstract:** We develop a method for self-localization of vehicles by matching their camera images against static aerial images (e.g. Google Maps). Our novel end-to-end trainable model achieves state-of-the-art performance on this task even in a zero-shot setting. For evaluation, we gather a large amount of diverse data from multiple public datasets and improve their inaccurate ground-truth geo-poses via a pseudo-label approach.

**Contact:** florian.fervers@iosb.fraunhofer.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 43

# OVTRACK: OPEN-VOCABULARY MULTIPLE OBJECT TRACKING

Li S., Fischer T., Ke L., Ding H., Danelljan M., Yu F.

**Abstract:** Contemporary MOT benchmarks and systems are limited to small sets of categories. We extend MOT beyond fixed vocabularies by tackling a novel task, open-vocabulary MOT, and developing OVTrack, an open-vocabulary tracker. Its design is based on two key ingredients: First, leveraging vision-language models for both classification and association via knowledge distillation; second, a data hallucination strategy for robust appearance feature learning from denoising diffusion probabilistic models.

**Contact:** tobias.fischer6@web.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 44

# DO ALL CNNS LEARN THE SAME FILTERS? A LARGE-SCALE STUDY

Gavrikov P., Keuper J.

**Abstract:** We investigate whether CNNs learn unique representations for each problem by exploring the convolution filter patterns on a large, heterogeneous model zoo of pre-trained CNNs. Our results show that the distributions are surprisingly largely independent of many dimensions like the task or training data but overparameterization causes networks to form "degenerated" filters. Finally, we investigate differences in robust CNNs and find a distinct first layer and higher diversity in patterns.

**Contact:** paul.gavrikov@hs-offenburg.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 45

# DOC2GRAPH: A TASK AGNOSTIC DOCUMENT UNDERSTANDING FRAMEWORK BASED ON GRAPH NEURAL NETWORKS

Gemelli A., Biswas S., Civitelli E., Lladós J. and Marinai S.

**Abstract:** Recently, the application of Graph Neural Networks (GNNs) has become crucial in various document-related tasks, thanks to their ability of unravel important structural patterns. Previous works in the literature propose task-driven models and heuristic-based graphs representations. We propose Doc2Graph, a task-agnostic document understanding framework based on a custom GNN model suited for documents.

**Contact:** andrea.gemelli@unifi.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 46

# HOOD: HIERARCHICAL GRAPHS FOR GENERALIZED MODELLING OF CLOTHING DYNAMICS

Artur Grigorev, Bernhard Thomaszewski, Michael J. Black, Otmar Hilliges

**Abstract:** We propose a method that leverages graph neural networks, multi-level message passing, and unsupervised training to enable real-time prediction of realistic clothing dynamics.

Whereas existing methods based on linear blend skinning must be trained for specific garments, our method is agnostic to body shape and applies to tight-fitting garments as well as loose, free-flowing clothing. Our method furthermore handles changes in topology (e.g., garments with buttons or zippers) and material properties at inference time.

As one key contribution, we propose a hierarchical message-passing scheme that efficiently propagates stiff stretching modes while preserving local detail. We empirically show that our method outperforms strong baselines quantitatively and that its results are perceived as more realistic than state-of-the-art methods

**Contact:** agrigorev@ethz.ch

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 47

# THE FLORENCE INTERPERSONAL RELATION VIDEO DATASET

Guerdelli H., Ferrari C., Berretti S., Del Bimbo A.

**Abstract:** In facial expression analysis, most if not all existing datasets focus on macro- and micro-facial expression[1]. There are also some datasets for studying interpersonal relations [2], but no one of these datasets provide multiple annotations. In this work, we propose a dataset that is, to our knowledge, the first annotated dataset with age, gender, social relationship, valence and arousal and 16 emotion terms that describe emotion variation.

**Contact:** hajer.guerdelli@unifi.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 48

# THE BENEFITS & LIMITATIONS OF PSEUDO LABEL ADAPTATION ON EPIC-KITCHENS-100

Guerrier R., Plizzari C., Damen D., Perrett T.

**Abstract:** The EPIC-KITCHENS-100 UDA challenge for action recognition poses an interesting but difficult problem. Not only do models need to adapt to a different distribution without ground truth. they also need to handle multiple domain shifts. Due to recent successes, we examined the benefits and limitations of pseudo labelling methods on adaptation. We find that although adaptation is possible, it is hindered when the pseudo label class distribution does not match the test set distribution very well.

**Contact:** rhodriguerrier98@gmail.com

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 49

# MULTI-PERSON HUMAN MOTION PREDICTION

GUO W., DU Y., Bie X., Shen X., Lepetit V., Alameda-Pineda X., Moreno-Noguer F.

**Abstract:** This work tackles the problem of human motion prediction, consisting in forecasting future body poses from historically observed sequences. We present two contributions: (1) We collect and make publicly available ExPI, a large dataset of highly interacted multi-person extreme dancing poses, annotated with 3D joint locations and body shapes. (2) SOTA approaches of human motion prediction provide good results, however, they rely on deep learning architectures of arbitrary complexity such as RNN, Transformers or GCN, typically requiring multiple training stages and more than 2M parameters. We show that, after combining with a series of standard practices, a lightweight network based on MLPs with only 0.14M parameters can surpass the SOTA performance. Exhaustive evaluations on multiple datasets show that our method(siMLPe) consistently outperforms other approaches.

**Contact:** wen.guo@inria.fr

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 50

# CONTINUAL LEARNING FOR THE ICUB HUMANOID ROBOT – AN EMPIRICAL ANALYSIS

Gupta S., Biggio B.

**Abstract:** Continual Learning (CL) or Lifelong (Incremental) Learning is a machine learning paradigm devised to tackle the problem of learning in a non-stationary environment from a temporal stream of data without forgetting old information even in the absence of old data. This paradigm is particularly interesting in the field of robotics given the nature of the interaction of an autonomous agent with the real world and the expectation to learn continually and adapt in the dynamic setting. In robotic vision, Single Incremental task (SIT) scenarios are highly desirable, as robots can incrementally improve their recognition capabilities when exposed to both known and completely new tasks. In this work, we do an empirical analysis using the iCub humanoid robot which implements object recognition using deep learning, and in particular, pre-trained ImageNet models. We consider the SIT scenario using state-of-the-art continual learning strategies from the ContinualAI Avalanche library [2], and report their performances on the iCubWorld28 dataset, where each of the 28 classes (i.e., objects) is learnt incrementally.

**Contact:** srishti.gupta@unica.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 51

# HUMAN-OBJECT INTERACTION CAPTURE FROM WEARABLE SENSORS

Guzov V., Chibane J., Marin R., He Y., Sattler T., Pons-Moll G.

**Abstract:** Understanding human behavior cannot be possible without tools to capture motions and interactions with the surrounding environment. Traditional multi-camera systems are impractical for large volumes, leading to an interest in wearable sensors and egocentric systems. But this presents a new, so far unexplored problem: How to track humans and objects using only wearable sensors? We subdivide the problem into two: 1) where the human is in the scene and 2) when and how the human interacts with an object. We present HPS and iReplica – our answers to each of those questions.

**Contact:** vladimir.guzov@mnf.uni-tuebingen.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 52

# AUTONOMOUS DRIVING AS A SEQUENCE TO SEQUENCE PROBLEM

Hamdan S., Barin M. R., Güney F.

**Abstract:** Causal transformers, such as GPT-3, have shown remarkable performance in language modeling. Furthermore, recent work [1][2] proposes formulations for different reinforcement learning tasks as sequence modeling tasks. In this work, we explore formulating autonomous driving as a sequence-to-sequence problem. We introduce CarFormer, a causal transformer model trained on driving data, and extensively evaluate and compare different input and output representations for the state and action spaces.

**Contact:** shamdan17@ku.edu.tr

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 53

# HIGH-FIDELITY 3D HUMAN DIGITIZATION FROM SINGLE 2K RESOLUTION IMAGES

Han S., Park M., Yoon J., Kang J., Park Y., Jeon H.

**Abstract:** High-quality 3D human reconstruction requires large-scale training data and appropriate network design. To tackle these problems, we propose a simple yet effective 3D human digitization method called 2K2K, which constructs a 2K human dataset and infers 3D human models from 2K resolution images. The proposed method separately recovers the global shape of a human and its details with part-wise normal prediction and coarse-to-fine method. In addition, we also provide 2,050 3D human models.

**Contact:** sanghunhan@gm.gist.ac.kr

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 54

# GESTURE-SYNC: DETERMINING WHO IS SPEAKING WITHOUT A TALKING HEAD

Hegde Sindhu, Zisserman Andrew

**Abstract:** In this work, we introduce a new synchronisation task, Gesture-Sync: determining if a person's gestures are correlated with their speech. We design a dual-encoder model using self-supervised learning, and compare a number of input representations: (i) RGB frames, (ii) keypoint images, and (iii) keypoint vectors. We demonstrate applications of Gesture-Sync for audio-visual synchronisation, and in determining who is the speaker in a crowd, without seeing their faces.

**Contact:** sindhu.hegde@eng.ox.ac.uk

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 55

# EUREKA-MOMENTS IN TRANSFORMERS: MULTI-STEP TASKS REVEAL SOFTMAX INDUCED OPTIMIZATION PROBLEMS

Hoffmann David T., Schrodi Simon, Behrmann Nadine, Fischer Volker, Brox Thomas

**Abstract:** We study rapid improvements of the transformer loss when being confronted with multi-step tasks. In contrast to CNNs, transformers struggle to learn latent tasks and learn a prior instead. When transformers learn the latent task, it's rapid and unexpected after training and validation loss saturated for many epochs. We call such leaps Eureka-Moments. We trace the problem back to the softmax function in the self-attention block of transformers and show ways to alleviate the problem. These fixes improve training speed and result in a higher likelihood to converge to a good solution.

**Contact:** david.hoffmann2@de.bosch.com

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 56

# WHERE ARE MY NEIGHBORS? EXPLOITING PATCHES RELATIONS IN SELF-SUPERVISED VISION TRANSFORMER

Camporese G., Izzo E., Ballan L.

**Abstract:** Vision Transformers (ViTs) enabled the use of the transformer architecture on vision tasks showing impressive performances when trained on big datasets. However, on relatively small datasets, ViTs are less accurate given their lack of inductive bias. To this end, we propose a simple but still effective Self-Supervised Learning (SSL) strategy to train ViTs, that without any external annotation or external data, can significantly improve the results. Specifically, we define a set of SSL tasks based on relations of image patches that the model has to solve before or jointly the supervised task. Differently from ViT, our RelViT model optimizes all the output tokens of the transformer encoder that are related to the image patches, thus exploiting more training signals at each training step. We investigated our methods on several image benchmarks finding that RelViT improves the SSL state-of-the-art methods by a large margin, especially on small datasets.

**Contact:** elena.izzo@phd.unipd.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 57

# EXPLORING SSL STABILITY VIA CENTER VECTORS

Jha Abhishek, Asano Yuki M., Tuytelaars Tinne

**Abstract:** This work examines the stability mechanisms of various self-supervised learning (SSL) techniques: (i) It explores both contrastive techniques (e.g., SimCLR) and non-contrastive techniques (e.g., BYOL, SWAV, SimSiam, Barlow Twins, DINO). (ii) The paper proposes a framework that argues that despite formulation differences, these methods implicitly optimize a shared objective: minimizing the magnitude of the expected representation over all data samples while maximizing expected representation for individual samples over different augmentations. (iii) Mathematical and empirical evidence, supported by experiments on the Imagenet100 dataset, validates the framework. The analysis enhances understanding of SSL principles and reveals commonalities among techniques.

**Contact:** abhishek.jha@esat.kuleuven.be

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 58

# TETRAHEDRAL DIFFUSION MODELS FOR 3D SHAPE GENERATION

Kalischek N., Peters T., Wegner D., Schindler K.

**Abstract:** Denoising diffusion models (DDMs) can be readily applied to 3D point clouds. However, point clouds have a number of drawbacks (no notion of topology). Here, we introduce tetrahedral DDMs. The much more structured representation has several advantages, e.g., it is possible to guide the diffusion. We develop tetrahedral convolutions, down- and up-sampling kernels. With those operators, 3D shape generation amounts to learning displacement vectors and signed distance values on the tetrahedral grid.

**Contact:** nikolai.kalischek@geod.baug.ethz.ch

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 59

# ANALYSIS OF CONTROLLABILITY AND FAIRNESS IN DIFFUSION MODELS

Kara O., Yurtseven S., Yesiltepe H., Stojanov S., Rehg J

**Abstract:** The prevalence of diffusion models in generative-based computer vision applications has raised concerns about biases resulting from imbalanced datasets. These biases can lead to under-representation of certain attributes such as gender, age especially in face datasets. This study focuses on identifying the bias in diffusion models from two perspectives and addressing this issue by proposing a methodology to mitigate bias in diffusion models. The approach involves identifying the global directions in h-space that control the target attribute and subsequently adjusting it through a fine-tuning process.

**Contact:** okara7@gatech.edu

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 60

# FAITHFUL TEXT-TO-IMAGE GENERATION BY SELECTION

Karthik S., Roth K., Mancini M., Akata Z.

**Abstract:** Text-to-image (T2I) models can lack faithfulness, where they do not reflect the textual prompt accurately. In this work, we show that faithfulness can be tackled as a candidate selection problem and introduce a simple pipeline that generates candidate images for a text prompt and picks the best one according to an automatic scoring system that can leverage existing T2I evaluation metrics. Results on diverse benchmarks show improved faithfulness over other post-hoc methods.

**Contact:** shyamgopalkart@gmail.com

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 61

# THE CALTECH FISH COUNTING DATASET: A BENCHMARK FOR MULTIPLE-OBJECT TRACKING AND COUNTING

Kay J., Kulits P., Stathatos S., Deng S., Young E., Beery S., Van Horn G., Perona P.

**Abstract:** We present the Caltech Fish Counting Dataset, a large-scale dataset for detecting, tracking, and counting fish in sonar videos. Fish counting in sonar is an impactful real-world application with open computer vision challenges such as domain generalization and low signal-to-noise data. In our ongoing work we are partnering with fisheries organizations in the United States and Canada to deploy computer vision algorithms for real-world use in salmon conservation and fisheries management.

**Contact:** justinkay92@gmail.com

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 62

# OSSO: OBTAINING SKELETAL SHAPE FROM OUTSIDE

Keller M., Zuffi S., Pujades S., Black M.

**Abstract:** We address the problem of inferring the anatomic skeleton of a person, in an arbitrary pose, from the 3D surface of the body. Existing state-of-the-art methods for inferring the skeleton are not learned from or validated against ground truth data. In contrast, we leverage a dataset of medical scans to build a dataset of body shapes and their corresponding skeletons. From this dataset, we learn to predict a realistic skeleton inside the body.

**Contact:** marilyn.keller@tuebingen.mpg.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 63

# SOUND TO VISUAL SCENE GENERATION BY AUDIO-TO-VISUAL LATENT ALIGNMENT

Sung-Bin K., Senocak A., Ha H., Owens A., Oh T-H.

**Abstract:** We propose Sound2Scene, a method that for generating an image of a scene from sound. We design a model that works by scheduling the learning procedure of each model component to associate audio-visual modalities despite their information gap. The key idea is to enrich the audio features with visual information by learning to align audio to visual latent space. To further improve the quality of our generated images, we use sound source localization to select the audio-visual pairs that have strong cross-modal correlations.

**Contact:** sungbin@postech.ac.kr

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 64

# CUBOIDS REVISITED: LEARNING ROBUST 3D SHAPE FITTING TO SINGLE RGB IMAGES

Kluger F., Ackermann H., Brachmann E., Yang M., Rosenhahn B.

**Abstract:** We propose a robust estimator for primitive fitting which can abstract real-world environments using cuboids. A RANSAC estimator guided by a neural network fits primitives to 3D features such as a depth map. We condition the network on previously detected parts of the scene, parsing it one-by-one. Results on the NYU dataset demonstrate that our algorithm successfully abstracts cluttered real-world 3D scene layouts.

**Contact:** kluger@tnt.uni-hannover.de

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 65

# NEURAL FIELDS FOR LATENT FORCE FIELD DISCOVERY IN INTERACTING SYSTEMS

Kofinas M., Bekkers E. J., Nagaraja N. S., Gavves E.

**Abstract:** Systems of interacting objects often evolve under the influence of field effects, yet previous works assume that systems evolve in a vacuum. In this work, we focus on discovering these fields, and infer them from the observed dynamics alone, without directly observing them. We disentangle equivariant local object interactions from external global field effects. We model interactions with equivariant graph networks, and combine them with neural fields in a novel graph network.

**Contact:** m.kofinas@uva.nl

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 66

# TETRA-NERF: REPRESENTING NEURAL RADIANCE FIELDS USING TETRAHEDRA

Kulhanek J., Sattler T.

**Abstract:** Neural Radiance Fields are a recent and popular approach for novel view synthesis and 3D reconstruction. We propose an adaptive radiance field representation based on tetrahedra obtained by the Delaunay triangulation instead of the uniform subdivision or point-based representations. The approach enables efficient training and leads to state-of-the-art results. Our method elegantly combines concepts from 3D geometry processing, triangle-based rendering, and modern neural radiance fields.

**Contact:** jonas.kulhanek@live.com

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 67

# AUTONOMOUS NANO-DRONES

Lorenzo Lamberti

**Abstract:** Nano-sized drones are gaining relevance in the Internet-of-Things ecosystem as unobtrusive robotic helpers. However, achieving a high degree of autonomy for complex multi-objective missions (e.g., safe flight, exploration, object detection) is very challenging: the limited computational and memory resources available aboard nano-UAVs introduce the challenge of minimizing and optimizing vision-based CNNs. This poster presents my research activities in optimizing and embedding CNNs for visual perception aboard nano-drones.

**Contact:** lorenzo.lamberti@unibo.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 68

# ADAS: A DIRECT ADAPTATION STRATEGY FOR MULTI-TARGET DOMAIN ADAPTATION

Lee S., Choi W., Kim C., Choi M., Im S.

**Abstract:** We aim to directly adapt a single model from a labeled source domain to multiple unlabeled target domains in a semantic segmentation task without pretrained domain-specific models. We first globally adapt the model via image translation, then refine the knowledge of each target domain via curriculum learning with bi-directional adaptive region selection.

**Contact:** lsh5688@dgist.ac.kr

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 69

# MONOCULAR REACTIVE COLLISION AVOIDANCE FOR MAV TELEOPERATION WITH DEEP REINFORCEMENT LEARNING

Brilli R., Legittimo M., Crocetti F., Leomanni M., Fravolini M. L., Costante G.

**Abstract:** Enabling semi-autonomous capabilities in MAVs for teleoperation is crucial. This work introduces a novel setting where operators provide high-level signals. MAVs employ an end-to-end DRL model for trajectory tracking and collision avoidance. It allows free movement in 3D space using RGB images and robot position, without assumptions about obstacle characteristics. Effectiveness and generalization are demonstrated against a SotA baseline in simulated environments.

**Contact:** marco.legittimo@studenti.unipg.it

**Presentation Type:** Poster

**Date:** Monday 10 July 2023

**Time:** 21:30

**Poster Session:** 1

**Poster Number:** 70

# EXPLOITING MULTIMODAL SYNTHETIC DATA FOR EGOCENTRIC HUMAN-OBJECT INTERACTION DETECTION IN AN INDUSTRIAL SCENARIO

Leonardi R., Ragusa F., Furnari A., Farinella G.M.

**Abstract:** We tackle the problem of EHOI detection in an industrial setting. To overcome the lack of datasets in this context, we propose a pipeline for generating synthetic images of EHOIs paired with several annotations. To show the efficacy of our pipeline, we present EgoISM-HOI, a multimodal dataset of EHOI images, and we designed a method that uses these multimodal signals to detect EHOIs. Our study shows that using synthetic data to pre-train our method improves performance when tested on real-world data.

**Contact:** rosario.leonardi@phd.unict.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 71

# MATCH, EXPAND AND IMPROVE: UNSUPERVISED FINETUNING FOR ZERO-SHOT ACTION RECOGNITION WITH LANGUAGE KNOWLEDGE

Lin Wei, Karlinsky Leonid, Shvetsova Nina, Possegger Horst, Kozinski Mateusz, Panda Rameswar, Feris Rogerio, Kuehne Hilde, Bischof Horst

**Abstract:** Large scale Vision-Language (VL) models have shown tremendous success in aligning representations between visual and text modalities. This enables remarkable progress in zero-shot recognition, image generation & editing, and many other exciting tasks. However, VL models tend to over-represent objects while paying much less attention to verbs, and require additional tuning on video data for best zero-shot action recognition performance. While previous work relied on large-scale, fully-annotated data, in this work we propose an unsupervised approach. We adapt a VL model for zero-shot and few-shot action recognition using a collection of unlabeled videos and an unpaired action dictionary. Based on that, we leverage Large Language Models and VL models to build a text bag for each unlabeled video via matching, text expansion and captioning. We use those bags in a Multiple Instance Learning setup to adapt an image-text backbone to video data.

**Contact:** wei.lin@icg.tugraz.at

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 72

# COUNT-LEVEL WEAKLY-SUPERVISED DENSITY ESTIMATION FOR CROWD COUNTING

Litrico M., Cheng F., Pound M., Tsaftaris S. A., Battiato S., Giuffrida M. V.

**Abstract:** To reduce the cost of collecting location-level annotations to predict crowd densities, this paper presents WADE, a training framework that predicts density maps leveraging only count-level annotations. To recover the spatial information, WADE is trained over pseudo-density maps, generated from count-level annotations. Our approach outperforms baseline and competing methods and yields comparable performance w.r.t. state-of-the-art semi-supervised methods that require location-level annotations.

**Contact:** mattia.litrico@phd.unict.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 73

# SEGMENTATION OF THE INFERIOR ALVEOLAR CANAL

Lumetti L., Bolelli F., Grana C.

**Abstract:** The Inferior Alveolar Canal (IAC) is a main interest in the maxillofacial field, as its accurate localization reduces the risks of injury during surgical procedures. Recent literature has focused on developing deep-learning techniques to produce accurate segmentations of the IAC, while are still strong limitations due to the limited amount of publicly available maxillofacial datasets. This work tackles both problems by proposing an annotation tool and a novel DNN for the segmentation of the IAC.

**Contact:** luca.lumetti@unimore.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 74

# GENERATION OF A PSEUDO-REFERENCE DATA FOR A SUPERVISED DENOISING IN DIFFUSION MRI

Machnio J., Pieciak T.

**Abstract:** The study aims to propose a method to generate pseudo-reference data and remove the noise in diffusion MRI data. The proposed method should remove as much noise as possible, with the smallest possible amount of medically valuable information. The described method is similar to the DeepDTI technique, but the spherical harmonic series are considered instead of modelling the pseudo-reference data by diffusion tensor. The proposed technique is model-free and, therefore, it is more likely not to lose the information. The Convolutional Neural Network named DeepSH was proposed as an actual denoising technique. The network studied the differences between the pseudo-ground-truth data reconstructed by SHs for all possible gradient directions and the input - highly noised data reconstructed by spherical harmonics for smaller directions. The model performance is evaluated visually by comparing the input and output data and the RGB Fractional Anisotropy maps as well as parameters from the NODDI algorithm. The proposal proves significant denoising properties with a negligible removed noise part which does not affect the microstructural data. And because of the method simplicity, it can be introduced for clinical use without any additional requirements.

**Contact:** julia.machnio1999@gmail.com

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 75

# SELF SUPERVISED MASKED CONVOLUTIONAL TRANSFORMER BLOCK FOR ANOMALY DETECTION

Madan Neelu, Ristea Nicolae-Catalin, Ionescu Tudor Radu, Nasrollahi Kamal, Khan Shahbaz Fahad, Moeslunf B. Thomas, Shah Mubarak

**Abstract:** We present a novel self-supervised masked convolutional transformer block (SSMCTB) that comprises the reconstruction-based functionality at a core architectural level. The proposed self-supervised block is extremely flexible, enabling information masking at any layer of a neural network and being compatible with a wide range of neural architectures. The block is integrated with multiple baselines for anomaly detection and almost always brings improvement.

**Contact:** nema@create.aau.dk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 76

# TACKLING WITH DATA HETEROGENEITY IN MULTI-DOMAIN FEDERATED LEARNING

Madni Hussain Ahmad, Foresti Gian Luca

**Abstract:** Federated Learning (FL) is a decentralized collaborative machine learning that allows model training by multiple participants without sharing original data. Data are heterogeneous due to domain shift and sometimes have different class distributions for multiple participants in real-world problems that is a challenging scenario of FL. To solve the problem of such heterogeneity in data, we propose a method based on non-convolutional Transformer architectures for model training. Moreover, we use two loss functions to minimize the divergence and optimize the global model.

**Contact:** madni.hussainahmad@spes.uniud.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 77

# TRIPLETTRACK: 3D OBJECT TRACKING USING TRIPLET EMBEDDINGS AND LSTM

Marinello N., Proesmans M., Van Gool L.

**Abstract:** In this paper we introduce a 3D Multi-Object Tracking method that solely relies on cameras for autonomous driving purposes. Our approach combines triplet embeddings, which extract visual and 3D physical attributes, with motion descriptors. These motion descriptors offer a robust representation of object trajectories. Our approach is reliable and accurate in case of occlusions, missed detection and can detect re-appearance across different cameras.

**Contact:** nicola.marinello@esat.kuleuven.be

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 78

# A MASKED POSE APPROACH FOR MOTION RETARGETING

Martinelli G., Garau N., Bisagno N., Conci N.

**Abstract:** Human motion modelling aims at understanding and replicating human kinematics. By masking certain body parts or specific motions, it is possible to isolate and focus on the critical aspects of human behaviours. We present a novel approach for motion retargeting, which aims to transfer the movements from a source skeleton to a target one in a different format. Moreover, we show how our method can effectively be used to transfer real human motion to synthetic characters with humanoid skeletons.

**Contact:** giulia.martinelli-2@unitn.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 79

# EVENT-BASED CAMERA POSE TRACKER BY ∇T NERF

Mana Masuda, Yusuke Sekikawa, Hideo Satio

**Abstract:** How can we utilize sparse event camera observation for efficient recovery of the camera pose? We found that when the input pose to the NeRF coincides with the actual pose, the output of the temporal gradient of NeRF equals the observed intensity changes on the event's points. Using this principle, we propose an event-based camera pose tracking framework called TeGRA which realizes the pose update by using the sparse event's observation.

**Contact:** mana.smile@keio.jp

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 80

# LEARNING TO DETECT ATTENDED OBJECTS IN CULTURAL SITES WITH GAZE SIGNALS AND WEAK OBJECT SUPERVISION

Mazzamuto M., Ragusa F., Furnari A., Farinella G. M.

**Abstract:** Cultural sites offer insights into diverse cultures and arts. Wearable devices can enhance visitor experiences by detecting attention from egocentric images. We collected a dataset of egocentric images from museum visitors and compared unsupervised, weakly supervised, and fully supervised approaches for attended object detection. Evaluating on our and external datasets (COCO, EGO-CH), we found that weakly supervised approaches using 2D gaze labels are effective for attended object detection.

**Contact:** michele.mazzamuto@phd.unict.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 81

# DATA-DRIVEN FEATURE TRACKING FOR EVENT CAMERAS

Messikommer N.*, Fang C.*, Gehrig M., Scaramuzza D.

**Abstract:** Because of their high temporal resolution, increased resilience to motion blur, and very sparse output, event cameras have been shown to be ideal for low-latency and low-bandwidth feature tracking, even in challenging scenarios. In this work, we introduce the first data-driven feature tracker for event cameras, which leverages low-latency events to track features detected in a grayscale frame. We achieve robust performance via a novel frame attention module, which shares information across feature tracks.

**Contact:** nico.messikommer@gmail.com

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 82

# IMPACT OF DIGITAL FACE BEAUTIFICATION IN BIOMETRICS

Mirabet-Herranz N., Galdi C., Dugelay J-L.

**Abstract:** Social media platforms offer different tools to beautify face images by performing operations such as skin smoothing and the deformation of certain biometric features for instance by widening the eyes or making the nose thinner. Popular filters are applied to face images on publicly available databases and we studied the effect of such filters which compromise the ability of automatic FR systems to recognize individuals and affect the estimation of other facial traits such as gender and weight.

**Contact:** mirabet@eurecom.fr

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 83

# LAFTER: LABEL-FREE TUNING OF ZERO-SHOT CLASSIFIERS USING LANGUAGE AND UNLABELED IMAGE COLLECTIONS

Mirza M. J., Leonid K., Lin W., Kozinski M., Possegger H., Feris R., Bischof H.

**Abstract:** Recently, large-scale pre-trained Vision and Language (VL) models have set a new state-of-the-art (SOTA) in zero-shot visual classification enabling open-vocabulary recognition of potentially unlimited set of categories defined as simple language prompts. However, despite these great advances, the performance of these zero-shot classifiers still falls short of the results of dedicated (closed category set) classifiers trained with supervised fine-tuning. In this paper we show, for the first time, how to reduce this gap without any labels and without any paired VL data, using an unlabeled image collection and a set of texts auto-generated using a Large Language Model (LLM) describing the categories of interest and effectively substituting labeled visual instances of those categories. Using our label-free approach, we are able to attain significant performance improvements over the zero-shot performance of the base VL model and other contemporary methods and baselines on a wide variety of datasets, demonstrating absolute improvement of up to 11.7% (3.8% on average) in the label-free setting. Moreover, despite our approach being label-free, we observe 1.3% average gains over leading few-shot prompting baselines that do use 5-shot supervision.

**Contact:** muhammad.mirza@icg.tugraz.at

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 84

# DUAL CONDITION DIFFUSION MODELS FOR OOD DETECTION

Mishra D, Zhao He, Saha P, Noble J.A

**Abstract:** We aim to detect Out-of-Distribution (OOD) samples in settings where In-distribution (ID) comprises distinct classes with local variations while the OOD and ID classes have high similarity globally. We introduce Dual Condition Diffusion Models (DCDM), where we condition diffusion models (DM) on ID class information (IDCC) and latent features of the input image (LIFC). We apply our model to separate heart views (ID) from other anatomies (OOD) in a free-hand fetal Ultrasound (US) video. Our model outperforms existing methods with a 12% improvement in accuracy.

**Contact:** divyanshu.mishra@eng.ox.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 85

# DEEP FEATURES FOR COLONOSCOPIC SLAM

Morlana, Javier

**Abstract:** Mapping and localization in endoluminal cavities from colonoscopies has to overcome significant shape and illumination changes. We present ColonMapper, the first system able to map the whole colon by combining a GeM global descriptor and LoFTR matching. ColonMapper is able to localize images from a different sequence within the built map. Besides, we also present DRAN, a unique deep feature extractor able to extract features for Image Retrieval, Reranking (by matching local features) and Feature-metric pose refinement. These two works will be combined in the future.

**Contact:** jmorlana@unizar.es

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 86

# ACCURATE 3D BODY SHAPE REGRESSION USING METRIC AND SEMANTIC ATTRIBUTES

Choutas Vasileios, Müller Lea, Huang Chun-Hao Paul, Tang Siyu, Tzionas Dimitrios, Black Michael J.

**Abstract:** Current 3D human mesh regressors fail to estimate body shape due to a lack of data. We exploit novel information in regressor training: (1) images and measurements of fashion models and (2) linguistic shape attributes for meshes and the model images. For shape evaluation, we collect HBW, a new dataset on which our method outperforms previous work. This is the first demonstration that 3D shape regression from images can be trained from easy-to-obtain measurements and linguistic shape attributes.

**Contact:** lea.mueller@tuebingen.mpg.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 87

# ASSEMBLYHANDS: TOWARDS EGOCENTRIC ACTIVITY UNDERSTANDING VIA 3D HAND POSE ESTIMATION

Ohkawa T., He K., Sener F., Hodan T., Tran L., and Keskin C.

**Abstract:** We present AssemblyHands, a large-scale benchmark dataset with accurate 3D hand pose annotations, to facilitate the study of egocentric activities with challenging hand-object interactions. The dataset includes synchronized egocentric and exocentric images sampled from the recent Assembly101 dataset, in which participants assemble and disassemble take-apart toys. Our study shows that having higher-quality hand poses directly improves the ability to recognize actions.

**Contact:** ohkawa-t@iis.u-tokyo.ac.jp

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 88

# LINE FEATURES: FROM DETECTION TO APPLICATIONS

Pautrat R., Suárez I., Lin J., Yu Y., Barath D., Oswald M., Larsson V., Pollefeys M.

**Abstract:** Many visual tasks such as SfM, SLAM, and 3D reconstruction, are based on feature points, which may fail in low-textured scenes and with repeated patterns. Lines are complementary features that can overcome such issues. We present here how to robustly extract lines from images, describe and match them across frames, and use them for visual applications. While lines suffer from additional challenges (occlusion, large extent, noisy endpoints), we show that they obtain state-of-the-art results.

**Contact:** remi.pautrat@inf.ethz.ch

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 89

# DOMAIN GENERALIZATION TO CLASSIFY 14TH-CENTURY ITALIAN PUNCHMARKS

Peaslee W. E., Cheng Y., Aviles-Rivero A. I., Wrapson L., Schönlieb C. B.

**Abstract:** Decorative impressions in gold leaf—called punches—became particularly widespread in Italian artworks during the 14th and 15th centuries. Classifying punches can contribute to painting attribution and historical investigations, especially since punch tools were used for multiple paintings. Previous applications of supervised deep learning have high classification accuracy on previously seen artworks; we show that domain generalization methods can profoundly improve accuracy on unseen artworks.

**Contact:** wep25@cam.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 90

# LEARNING TO RECONSTRUCT AND UNDERSTAND THE 3D WORLD

Songyou Peng

**Abstract:** Numerous applications in AR/VR, robotics, and autonomous driving require accurate 3D reconstruction of large scenes, as well as high-level understanding of diverse content. We cover three works that leverage the power of deep learning to unlock these capabilities. ConvONet enables 3D reconstruction from point clouds for large-scale scenes, NICE-SLAM can optimize in an end-to-end manner to obtain dense 3D reconstruction and camera poses in large indoor scenes, and OpenScene is able to do all kinds of novel 3D scene understanding tasks with open vocabularies.

**Contact:** songyou.peng@inf.ethz.ch

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 91

# SUPER-RESOLUTION 3D HUMAN SHAPE FROM A SINGLE LOW-RESOLUTION IMAGE

Pesavento M., Volino M., Hilton A.

**Abstract:** We propose a novel framework to reconstruct super-resolution human shape from a single low-resolution input image. The approach overcomes limitations of existing approaches that reconstruct 3D human shape from a single image, which require high-resolution images together with auxiliary data such as surface normal or a parametric model to reconstruct high-detail shape. The proposed framework represents the reconstructed shape with a high-detail implicit function. Analogous to the objective of 2D image super-resolution, the approach learns the mapping from a low-resolution shape to its high-resolution counterpart and it is applied to reconstruct 3D shape detail from low-resolution images. The approach is trained end-to-end employing a novel loss function which estimates the information lost between a low and high-resolution representation of the same 3D surface shape. Evaluation for single image reconstruction of clothed people demonstrates that our method achieves high-detail surface reconstruction from low-resolution (256x256) images without auxiliary data. Extensive experiments show that the proposed approach can estimate super-resolution human geometries with a significantly higher level of detail than that obtained with previous approaches when applied to low-resolution images.

**Contact:** m.pesavento@surrey.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 92

# HANDWRITTEN TEXT GENERATION FROM VISUAL ARCHETYPES

Pippi V., Cascianelli S., Cucchiara R.

**Abstract:** Styled Handwritten Text Generation is a challenging task, especially in the case of unseen styles and new words, and even more when these latter contain long-tail characters, i.e., characters that are rarely seen in training. In our devised Visual Archetypes Transformer (VATr), we focus on obtaining an informative representation of the style and text via specific synthetic pretraining and a novel representation of the textual content as a sequence of Visual Archetypes, respectively.

**Contact:** vittorio.pippi@unimore.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 93

# INTERACTION IN MIXED REALITY ENVIRONMENTS

Pizzo M., Solari F., Chessa M.

**Abstract:** Mixed Reality combines real and virtual worlds, allowing for obstacle avoidance in Virtual Reality and enhancing interaction via passive haptics [1]. In this work, we propose a Mixed Reality system to enable coherent interaction with objects in the real environment that appear differently in the virtual one. The focus is on identifying chairs and allowing users to sit on virtual chairs.

**Contact:** marianna.pizzo@edu.unige.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 94

# SELF-SUPERVISED ACCIDENT ANTICIPATION FROM DASHCAM VIDEOS

Pjetri A., Abbondandolo D., Coimbra de Andrade D., Caprasecca S., Sambo F., Bagdanov A. D.

**Abstract:** Accident anticipation is the problem of predicting a dangerous event as early as possible. Typical methods in the literature approach the problem by building logic on top of object detectors (OD) and by using expensive labels like crash presence and crash timestamp. We propose a method that avoids OD by assuming danger level is non-decreasing in the moments before the event. It also uses less manual supervision by exploiting acceleration signal for crash timestamp detection.

**Contact:** aurel.pjetri@unifi.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 95

# MACHINE UNLEARNING FOR EXPLAINABILITY AND SUSTAINABILITY

Poppi S., Sarto S., Cornia M., Baraldi L., Cucchiara R.

**Abstract:** Machine learning's opacity poses ethical concerns, especially regarding the inability to be deleted from AI systems and the high costs associated with retraining. Privacy rights, like GDPR's right to be forgotten, and sustainability considerations demand a solution. Machine Unlearning (MU) offers a crucial remedy, enabling data removal from deep learning models and avoiding extensive retraining, ensuring privacy and cost-effectiveness.

**Contact:** samuele.poppi@unimore.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 96

# DINER: DEPTH-AWARE IMAGE-BASED NEURAL RADIANCE FIELDS

Prinzler Malte, Hilliges Otmar, Thies Justus

**Abstract:** Given a sparse set of RGB input views, we predict depth and feature maps to guide the reconstruction of a volumetric scene representation (NeRF) that allows us to render 3D objects under novel views. Specifically, we propose novel techniques to incorporate depth information into feature fusion and efficient scene sampling. We evaluate our method by synthesizing novel views, both for human heads and for general objects, and observe significantly improved qualitative results and increased perceptual metrics compared to previous methods.

**Contact:** malte.prinzler@tuebingen.mpg.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 97

# USE OF DEEP LEARNING FOR 3D FETAL CARDIAC MRI RECONSTRUCTION

Prokopenko D., Hammernik K., Roberts T., Lloyd D., Rueckert D., Hajnal J.

**Abstract:** Dynamic fetal cardiac MRI requires high temporal and spatial resolution. Conventional two-stage acquisition methods are susceptible to motion and double scanning time. We explore and systematically evaluate a set of DNNs optimised to recover fully-sampled data, eliminating the need for two-stage approach and possibly replacing the reconstruction. We show that the best-performing model manages to recover detailed maternal anatomy but misses important dynamic properties of the fetal heart.

**Contact:** denis.prokopenko@kcl.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 98

# THERMAL IMAGE SYNTHESIS

Qazi Tayeba., Lall Brejesh.

**Abstract:** Development of DL applications for thermal data necessitates huge volume of thermal data. However, thermal data is scarce due to high-cost of thermal sensors. Therefore, synthetically creating thermal data will enable development of high-accuracy DL models for thermal applications. We adapt diffusion model for the task of thermal image synthesis.

**Contact:** bsz218186@iitd.ac.in

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 99

# PUT YOUR PPE ON: A TOOL FOR SYNTHETIC DATA GENERATION AND RELATED BENCHMARK IN CONSTRUCTION SITE SCENARIOS

Quattrocchi C., Di Mauro D., Furnari A., Lopes A., Moltisanti M., Farinella G. M.

**Abstract:** Machine Learning algorithms can enforce safety in construction sites by detecting if workers are wearing protective equipment, have fallen, or are too close to danger. However, large amounts of labeled data are required. We propose a pipeline to produce synthetic data to mitigate real data scarcity. Our benchmark tests the usefulness of the generated data for safety compliance, fall detection, and distance regression. Experiments show that synthetic data reduces the need for real data and achieves good performance.

**Contact:** camillo.quattrocchi@phd.unict.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 100

# STILLFAST: AN END-TO-END APPROACH FOR SHORT-TERM OBJECT INTERACTION ANTICIPATION

Ragusa F., Furnari A., Farinella G. M.

**Abstract:** We studied the short-term object interaction anticipation problem from the egocentric point of view, proposing a new end-to-end architecture named StillFast. Our approach simultaneously processes a still image and a video detecting and localizing next-active objects, predicting the verb which describes the future interaction and determining when the interaction will start. Experiments on the large-scale dataset EGO4D show that our method outperformed state-of-the-art approaches on the considered task. Project web page and code: https://iplab.dmi.unict.it/stillfast/.

**Contact:** francesco.ragusa@unict.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 101

# VISUAL DNA: REPRESENTING AND COMPARING IMAGES USING DISTRIBUTIONS OF NEURON ACTIVATIONS

Ramtoula B., Gadd M., Newman P., De Martini D.

**Abstract:** We propose using Distributions of Neuron Activations (DNAs) as a general-purpose tool to represent images and evaluate their differences. DNAs fit distributions to neuron activations in a pre-trained feature extractor. Comparing DNAs allows measuring dataset differences and can focus on custom attributes. DNAs are compact (¡15MB), and we demonstrate their value on diverse datasets and tasks such as conditional dataset comparison, synthetic image evaluation, and transfer learning prediction.

**Contact:** benjamin@robots.ox.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 102

# HIERARCHICAL AVERAGE PRECISION TRAINING FOR PERTINENT IMAGE RETRIEVAL

Ramzi E., Audebert N., Thome N., Rambour C., Bitot X.

**Abstract:** Image Retrieval (IR) is evaluated with binary metrics, e.g. Average Precision (AP)/Recall@k, that ignore errors severity. This work introduces the HAPPIER method for hierarchical (H-) IR, based on a new H-AP metric leveraging hierarchical labels to refine AP with errors importance. It optimizes H-AP with a smooth lower bound surrogate and a clustering loss that ensures consistent ordering. Experiments on 6 datasets show that HAPPIER outperforms SOTA methods on H-IR and is on par on standard IR.

**Contact:** elias.ramzi@lecnam.net

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 103

# USING EXPLANATIONS TO GUIDE MODELS

Rao S., Böhle M., Parchami-Araghi A., Schiele B.

**Abstract:** Deep networks often base their decisions on spurious features, which hurts generalization. Guiding models by optimizing their explanations to focus on the right regions could mitigate this, but has so far been limited to relatively simple datasets using costly fine-grained annotations. In our work, we study the effectiveness of using coarse bounding box annotations. We conduct an in-depth evaluation on large-scale datasets, and show its effectiveness with coarse, noisy, and limited annotations.

**Contact:** srao@mpi-inf.mpg.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 104

# TOWARDS DATA-DRIVEN REGULARIZATION IN ABDOMINAL IMAGE REGISTRATION

Reithmeir A.,Zimmer V.A., Schnabel J.A.

**Abstract:** In medical image registration, regularization of the deformation is used to obtain a physically-plausible solution. The majority of the state-of-the-art methods use hand-crafted and homogeneous regularizers. For abdominal registration, such regularizers might not be suitable due to the complex multi-organ motion. Instead, spatially-adaptive and data-driven regularizers should be explored.

**Contact:** anna.reithmeir@tum.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 105

# VOLRECON: VOLUME RENDERING OF SIGNED RAY DISTANCE FUNCTIONS FOR GENERALIZABLE MULTI-VIEW RECONSTRUCTION

Ren Y., Wang F., Zhang T., Pollefeys M., Süsstrunk S.

**Abstract:** Multi-view reconstruction estimates the dense geometry of a scene, given several posed images. Previous generalizable implicit surface reconstruction methods build geometry volume to interpolate SDF features but are limited by the volume resolution. We propose to combine the global shape prior with local features to reconstruct with fine details.

**Contact:** yufan.ren@epfl.ch

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 106

# PLANT: EXPLAINABLE PLANNING TRANSFORMERS VIA OBJECT-LEVEL REPRESENTATIONS

Renz K., Chitta K., Mercea O.B., Koepke A.S., Akata Z., Geiger A.

**Abstract:** Planning an optimal route in a complex environment requires efficient reasoning about the surrounding scene. In this paper, we propose PlanT, a novel approach for that uses a standard transformer architecture. PlanT is based on imitation learning with a compact object-level input representation. Combining PlanT with an off-the-shelf perception module provides a sensor-based driving system that is more than 10 points better in terms of driving score than the existing state of the art.

**Contact:** Katrin.renz@uni-tuebingen.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 107

# MOVING: A MODULAR AND FLEXIBLE PLATFORM FOR EMBODIED VISUAL NAVIGATION

Rosano M., Ragusa F., Furnari A., Farinella G.M.

**Abstract:** We present MOVING, a flexible and modular hardware and software platform for visual mapping (SLAM) and navigation in the real world. The platform is easily detachable and can be installed on any robot with minimal adaptation required, enabling the reuse of the same robotic software regardless of the robot employed. The proposed platform can represent a solution to speed up the design and testing of new software for autonomous navigation systems, minimizing deployment time in the real world.

**Contact:** marco.rosano@unict.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 108

# EGOCENTRIC AUDITORY ATTENTION LOCALIZATION IN CONVERSATIONS

Ryan F., Jiang H., Shukla A., Rehg J.M., Ithapu V.

**Abstract:** In everyday settings like restaurants and parties, people converse in the presence of multiple speakers. We introduce the task of Selective Auditory Attention Localization (SAAL), which uses video and audio from a worn headset to determine who the wearer is listening to. We propose a multimodal spatiotemporal model that uses egocentric video and multichannel audio to predict the heatmap of the camera wearer's auditory attention. We evaluate on a challenging multi-speaker conversation dataset.

**Contact:** fiona.rcs@gmail.com

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 109

# TIME DOES TELL, A LOT: SELF-SUPERVISED LEARNING OF DENSE IMAGE REPRESENTATIONS

Mohammadreza Salehi, Efstratios Gavves, Cees G. M. Snoek, Yuki M. Asano

**Abstract:** Dense self-supervised learning (SSL) is gaining traction in unsupervised segmentation and pretraining, but the untapped potential of temporal video information limits its effectiveness. To overcome this, an approach called time-tuning is introduced. It incorporates temporal consistency by fine-tuning pretrained image models using an SSL temporal-alignment clustering loss on unlabelled videos. This unlocks the latent potential of temporal video data, significantly enhancing SSL performance.

**Contact:** s.salehidehnavi@uva.nl

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 110

# NLX-GPT: A MODEL FOR NATURAL LANGUAGE EXPLANATIONS IN VISION AND VISION-LANGUAGE TASKS

Sammani, Fawaz., Mukherjee Tanmoy., Deligiannis Nikos

**Abstract:** Natural language explanation (NLE) models describe black box decision-making using human-friendly language. Current NLE models separate task and explanation processes, disentangling the reasoning process. They are also memory expensive. We propose NLX-GPT, a compact language model that predicts and explains the answer simultaneously. With no region proposals or task model, it achieves better scores, fewer parameters, and is 15x faster than state-of-the-art models. We also introduce two new evaluation measures: explain-predict and retrieval-based attack.

**Contact:** fawaz.sammani@vub.be

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 111

# IMPLICIT SHAPE COMPLETION VIA ADVERSARIAL SHAPE PRIORS

Saroha Abhishek, Eisenberger Marvin, Yenamandra Tarun, Cremers Daniel

**Abstract:** We present a novel neural implicit method for partial shape completion. To that end, we combine a conditional Deep-SDF architecture with learned, adversarial shape priors. More specifically, our network converts partial inputs into a global latent code and then recovers the full geometry via an implicit, signed distance generator. Additionally, we train a PointNet++ discriminator that impels the generator to produce plausible, globally consistent reconstructions. In our experiments, we demonstrate state-of-the-art performance for completing partial shapes, considering both man-made objects (e.g. airplanes, chairs, ...) and deformable shape categories (human bodies).

**Contact:** abhishek.saroha@tum.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 112

# COMPUTATIONAL DESIGN OF FABRICABLE GEOMETRIC PATTERNS

Scandurra E., Callieri M., Giorgi D., Laccone F., Malomo L., Cignoni P.

**Abstract:** We design tileable and fabricable patterns that have predictable mechanical performance and can be assembled for creating surface tessellations. A regular recursive subdivision of the planar space generates different geometric configurations for candidate patterns, which we then refine to improve robustness. We simulate the pattern mechanical response targeting out-of-plane bending. The result is a library of diverse metamaterial patterns that support the direct design of mechanical structures.

**Contact:** elena.scandurra@phd.unipi.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 113

# DISTRACTING DOWNPOUR: ADVERSARIAL WEATHER ATTACKS FOR MOTION ESTIMATION

Schmalfuss J., Mehl L., Bruhn A.

**Abstract:** To test the robustness of motion estimation with optical flow under advers(arial) weather conditions, we present a differentiable particle-rendering to generate time-consistent and naturalistic snow, rain, sparks and fog. While moving particles already interfere with the motion prediction, the predictions become severely distorted when the weather is adversarially optimized. However, adding random weather to the training improves not only the weather-resilience, but also the generalization.

**Contact:** jenny.schmalfuss@vis.uni-stuttgart.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 114

# LIFELONG LOOP CLOSURE DETECTION

Scucchia M.

**Abstract:** Loop Closure Detection (LCD) [1] is the ability of robots to recognize already visited places and it is fundamental to enabling long-term autonomous navigation. This is not an easy task; in fact, places can change their appearance depending on several factors. To solve this problem, we need robots capable of learning in a continual fashion. The goal of my research activity is to improve the loop closure detection, powering robots with lifelong autonomous navigation capabilities.

**Contact:** matteo.scucchia2@unibo.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 115

# AN AR-BASED TOOL FOR ACQUISITION AND AUTOMATIC LABELING OF HUMAN-OBJECT INTERACTIONS FROM FIRST PERSON VISION

Seminara L., Ragusa F., Leonardi R., Farinella G. M., Furnari A.

**Abstract:** We focus on detecting Human-Object Interactions (HOI) from images acquired through the use of wearable devices. Detecting interactions enables object-specific assistance and enhances worker safety. Existing egocentric HOI detection methods require domain-specific data for fine-tuning. To reduce labeling costs, we developed a tool utilizing Microsoft HoloLens2's capabilities. We collected and labeled an industrial dataset with our tool, showing the improved performance of EHOI recognition models.

**Contact:** seminara.luigi@studium.unict.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 116

# MATERIALISTIC: SELECTING SIMILAR MATERIALS IN IMAGES

Sharma P., Philip J., Gharbi M., Freeman B., Durand F., Deschaintre V.

**Abstract:** We present a method capable of selecting the regions of a photograph exhibiting the same material as an artist-chosen area. Our proposed approach is robust to shading, specular highlights, and cast shadows, enabling selection in real images. We leverage DINO features coupled with a proposed Cross-Similarity module and an MLP head to extract material similarities in an image. Our model is trained on a synthetic image dataset and yet generalizes well to real-world images.

**Contact:** prafull@mit.edu

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 117

# GAZENERF: 3D-AWARE GAZE REDIRECTION WITH NEURAL RADIANCE FIELDS

Ruzzi A., Shi X., Wang X., Li G., De Mello S., Chang H. J., Zhang X., Hilliges O.

**Abstract:** We propose GazeNeRF, a 3D-aware method for the task of gaze redirection. We build on the intuition that the face region and eyeballs are separate 3D structures that move in a coordinated yet independent fashion. Our method leverages neural radiance fields and proposes a two-stream architecture that predicts volumetric features for the face and eye regions separately. Rigidly transforming the eye features via a 3D rotation matrix provides fine-grained control over the desired gaze angle.

**Contact:** X.Shi-3@tudelft.nl

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 118

# VIEWSET DIFFUSION: 0 TO N IMAGE-CONDITIONED 3D GENERATION FROM 2D DATA

Szymanowicz S., Rupprecht C., Vedaldi A.

**Abstract:** We tackle the inherent ambiguity in single-view 3D reconstruction by treating it as a conditional generation task. Training 3D generative models is challenging because often 3D ground truth data is not available. We address the issue of data availability by training a Diffusion Model to jointly denoise a multi-view image set. Our design of architecture and training scheme allows our model to perform 3D generation and ambiguity-aware single-view reconstruction in a feed-forward manner.

**Contact:** szymanowiczs@gmail.com

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 119

# ANOMALY DETECTION IN INDUSTRIAL PROCESSES

Taylor Alex

**Abstract:** I work in applying anomaly detection to industrial processes. The current de facto anomaly detection dataset, MVTec [1], is quasi-solved and does not represent challenging industrial processes due to containing images of homogeneous nature (consistent lighting, geometry, angles). Therefore, a new dataset is released, Industrial Defects in the Wild (IDW). The dataset has six classes and is far more challenging than MVTec. It contains images taken from video feeds of real industrial inspection processes. Alongside this, an anomaly detection wrapper is released, which contains 13 algorithms, making it the largest currently available. These algorithms are the top algorithms according to the PapersWithCode website. The wrapper has a focus on ease of use and allows fair benchmarking of these algorithms. A novel metric, Proportion Localized (PL), is created, which reports the portion of defects found by the anomaly detection algorithm. This is more interpretable than AUC and AUPRO and mitigates the issue of pixel-by-pixel noise by matching at the bounding box level. Finally, a system has been devised that directly allows anomaly detection algorithms to make bounding box predictions. Certain problems in industry require that an anomaly detection algorithm makes discrete decisions regarding anomalies and their positions. Therefore, bounding box outputs are desired as opposed to heatmaps. A novel architecture and training system have been designed that fulfill this task.

**Contact:** adjt20@bath.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 120

# FAST AND EASY REGISTRATION OF MICROSCOPY TIME-LAPSE DATA WITH NAPARI-ROI REGISTRATION PLUGIN

Giorgia Tortora (1), Alessia Candeo (2), Andrea Bassi (3), Stefano Buratti (4), Matteo Grenzi (5), Alex Costa (6)

**Abstract:** Registration and analysis of time-lapse data are required for the study of all dynamic biological processes. We present our easy-to-use plugin napari-roi-registration starting from the case study which led to its creation. The plugin allows simultaneous registration and subsequent analysis of multiple regions of interest (ROIs) in time-lapse data. Even though it was conceived for time-lapse widefield light microscopes data, it can potentially be used for any stack of frames.

**Contact:** giorgia.tortora@polimi.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 121

# INCORPORATING DIAGNOSTIC PRIOR WITH SEGMENTATION: A DEEP LEARNING PIPELINE FOR THE AUTOMATIC CLASSIFICATION OF AUTOIMMUNE BULLOUS SKIN DISEASES

Pastore V P, Touijer L, Capurro N, Cozzani E, Gasparini G, Parodi A, Odone F

**Abstract:** Autoimmune Bullous Diseases are a group of skin disorders defined by the development of autoantibodies against structural proteins of the epidermis and the basal membrane. The diagnosis requires ImmunoFluorescence and is currently limited to highly specialized centers. We propose an automatic classification framework based on a two-branches CNN, to classify the two main IF patterns, incorporating diagnostic prior with image segmentation, to increase prediction accuracy and medical expert trust.

**Contact:** larbi.touijer@edu.unige.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 122

# SPARF: NEURAL RADIANCE FIELDS FROM SPARSE AND NOISY POSES

Truong Prune, Rakotosaona Marie-Julie, Manhardt Fabian, Tombari Federico

**Abstract:** Neural Radiance Field (NeRF) has recently emerged as a powerful representation to synthesize photorealistic novel views. While showing impressive performance, it relies on the availability of dense input views with highly accurate camera poses, thus limiting its application in real-world scenarios. In this work, we introduce Sparse Pose Adjusting Radiance Field (SPARF), to address the challenge of novel-view synthesis given only few wide-baseline input images (as low as 3) with noisy camera poses. Our approach exploits multi-view geometry constraints in order to jointly learn the NeRF and refine the camera poses. By relying on pixel matches extracted between the input views, our multi-view correspondence objective enforces the optimized scene and camera poses to converge to a global and geometrically accurate solution. Our depth consistency loss further encourages the reconstructed scene to be consistent from any viewpoint. Our approach sets a new state of the art in the sparse-view regime on multiple challenging datasets.

**Contact:** prune.truong@vision.ee.ethz.ch

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 123

# AUTOMATED MONITORING OF IVD DEVICE ANALYSIS PROCESSES USING MACHINE LEARNING AND COMPUTER VISION TECHNIQUES

Tufo G.

**Abstract:** This study presents an automated monitoring system which analyze In Vitro Diagnostics (IVD) device processes in real-time. It captures images of elements involved in aspiration and dispensing of liquids, namely tips, during critical moments of the diagnostic session. The acquired images are processed using Machine Learning and Computer Vision algorithms, allowing continuous monitoring of the diagnostic process. The study focuses on the critical operation of gripping tips during the session.

**Contact:** giulia.tufo@uniroma1.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 124

# ANOMALY DETECTION FOR VISUAL INSPECTION

Ugwu C. I.

**Abstract:** Anomaly detection is a long-standing research field. Some anomalies appear in an object's geometry or density-based property. Although most companies already have advanced vision systems capable of capturing 3D data measurements, there is a research gap in 3D anomaly detection. This may also be due to the lack of publicly available datasets designed specifically for this task. It is my goal to contribute more to this field through my research.

**Contact:** cugwu@unibz.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 125

# SCADE: NERFS FROM SPACE CARVING WITH AMBIGUITY-AWARE DEPTH ESTIMATES

Uy Mikaela Angelina, Martin-Brualla Ricardo, Guibas Leonidas, Li Ke

**Abstract:** Neural radiance fields (NeRFs) have enabled high fidelity 3D reconstruction from multiple 2D input views. However, a well-known drawback of NeRFs is the less-than-ideal performance under a small number of views, due to insufficient constraints enforced by volumetric rendering. To address this issue, we introduce SCADE, a novel technique that improves NeRF reconstruction quality on sparse, unconstrained input views for in-the-wild indoor scenes. To constrain NeRF reconstruction, we leverage geometric priors in the form of per-view depth estimates produced with state-of-the-art monocular depth estimation models, which can generalize across scenes. A key challenge is that monocular depth estimation is an ill-posed problem, with inherent ambiguities. To handle this issue, we propose a new method that learns to predict, for each view, a continuous, multimodal distribution of depth estimates using conditional Implicit Maximum Likelihood Estimation (cIMLE). In order to disambiguate exploiting multiple views, we introduce an original space carving loss that guides the NeRF representation to fuse multiple hypothesized depth maps from each view and distill from them a common geometry that is consistent with all views. Experiments show that our approach enables higher fidelity novel view synthesis from sparse views.

**Contact:** mikacuy@stanford.edu

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 126

# ALIGNMIXUP: IMPROVING REPRESENTATIONS BY INTERPOLATING ALIGNED FEATURES

Venkataramanan Shashanka, Kijak Ewa, Amsaleg Laurent, and Avrithis Yannis

**Abstract:** Mixup is a powerful data augmentation method that interpolates between two or more examples in the input or feature space and between the corresponding target labels. However, how to best interpolate images is not well defined. Recent mixup methods overlay or cut-and-paste two or more objects into one image, which needs care in selecting regions. Mixup has also been connected to autoencoders, because often autoencoders generate an image that continuously deforms into another. However, such images are typically of low quality.

In this work, we revisit mixup from the deformation perspective and introduce AlignMixup, where we geometrically align two images in the feature space. The correspondences allow us to interpolate between two sets of features, while keeping the locations of one set. Interestingly, this retains mostly the geometry or pose of one image and the appearance or texture of the other. We also show that an autoencoder can still improve representation learning under mixup, without the classifier ever seeing decoded images. AlignMixup outperforms state-of-theart mixup methods on five different benchmarks.

**Contact:** shashanka.venkataramanan@inria.fr

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 127

# COVR: LEARNING COMPOSED VIDEO RETRIEVAL FROM WEB VIDEO CAPTIONS

Ventura L., Yang A., Schmid C., Varol G.

**Abstract:** Composed Image Retrieval (CoIR) has recently gained popularity as a task that considers both text and image queries together, to search for relevant images in a database. Most CoIR approaches require manually annotated datasets, containing image-text-image triplets, where the text describes a modification from the query image to the target image. However, manual curation of CoIR triplets is expensive and prevents scalability. In this work, we instead propose a scalable automatic dataset creation methodology that generates triplets given video-caption pairs. To this end, we mine paired videos with a similar caption from a large database, and leverage a large language model to generate the corresponding modification text. We automatically construct our WebVid-CoVR dataset by applying this procedure to the large WebVid2M collection, resulting in 1.6M triplets. Moreover, we introduce a new benchmark for composed video retrieval (CoVR) and contribute a manually annotated evaluation set, along with baseline results. We further show that training a CoVR model on our dataset transfers well to CoIR, improving the state of the art in the zero-shot setup on both the CIRR and FashionIQ benchmarks. Our code, datasets, and models will be made publicly available

**Contact:** lucas.ventura.r@gmail.com

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 128

# POP-3D: OPEN-VOCABULARY 3D OCCUPANCY PREDICTION FROM IMAGES

Vobecky A., Siméoni O., Hurych D., Gidaris S., Bursuc A., Pérez P., Sivic J.

**Abstract:** We propose a method for predicting open-vocabulary 3D semantic voxel occupancy maps from 2D images. We address the challenges of 2D-3D ambiguity and lack of annotated 3D data. Our contributions include a novel architecture and tri-modal self-supervised learning with images, language, and LiDAR point clouds using a strong pre-trained vision-language model. We demonstrate the effectiveness of our model on open vocabulary tasks, including zero-shot 3D semantic segmentation and text-based retrieval.

**Contact:** a.vobecky@gmail.com

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 129

# DISCO: DISENTANGLED CONTROL FOR HUMAN DANCE GENERATION

Wang T., Li L., Lin K., Lin C., Yang Z., Zhang H., Liu Z., Wang L.

**Abstract:** Generative AI remains challenging especially in the generation of human-centric content such as dance synthesis in real-world dance scenarios. In this paper, we highlight the real-world dance generation with three important properties: (i) Faithfulness; (ii) Generalizability; (iii) Compositionality. To address these challenges, we introduce a novel approach, DisCo, which includes the disentangled control to improve the faithfulness and compositionality of dance synthesis, and an effective human attribute pre-training for better generalizability to unseen humans.

**Contact:** TAN317@e.ntu.edu.sg

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 130

# POSEDIFFUSION: SOLVING POSE ESTIMATION VIA DIFFUSION-AIDED BUNDLE ADJUSTMENT

Wang J., Rupprecht C., Novotny D.

**Abstract:** We formulate the SfM problem inside a probabilistic diffusion framework. This novel view of the old problem (1) mirrors the iterative procedure of classic bundle adjustment and (2) allows a seamless integration of geometric constraints from epipolar geometry. Our approach significantly improves over the classic and learned methods on two real-world datasets, especially in difficult scenarios such as sparse views. We observe that our method can generalize across datasets without further training.

**Contact:** jianyuan@robots.ox.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 131

# SYMNP: LEARNING SYMMETRIES BETWEEN NEURAL POINTS

Wewer C., Ilg E., Schiele B., Lenssen J.

**Abstract:** 3D object reconstruction requires to balance prior and flexibility in order to infer unseen regions while perfectly fitting the observations. We present SymNP, a method to learn category-level symmetries within neural point radiance fields. By learning how information is shared between coherent neural points, our approach derives novel views from as few as a single input view. SymNP outperforms previous methods in reconstructing symmetric unseen regions, while providing semantic correspondences.

**Contact:** cwewer@mpi-inf.mpg.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 132

# LEARNING DEPTH FROM FOCUS IN THE WILD

Won C., Jeon H.

**Abstract:** For better photography, recent commercial cameras including smartphones have used a burst mode to take multiple images within a short time. These features lead us to examine depth from focus (DfF). Our method differs from relevant state-of-the-art works with unique features. Our method allows depth maps to be inferred in an end-to-end manner even with image alignment. And we introduce two key modules: a sharp region detection and an efficient downsampling, which address challenges of DfF.

**Contact:** cywon1997@gm.gist.ac.kr

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 133

# SEGMENTING MOVING OBJECTS VIA AN OBJECT-CENTRIC LAYERED REPRESENTATION

Xie J., Xie W., Zisserman A.

**Abstract:** The paper aims to discover, track and segment multiple moving objects in videos. To achieve this, we propose an object-centric segmentation model with a depth-ordered layer representation. We also introduce a scalable pipeline for generating multi-object synthetic data to train the model, eliminating the need for labour-intensive annotations. Our model is evaluated on standard video segmentation benchmarks, demonstrating state-of-the-art performance among existing human-label-free methods.

**Contact:** jyx@robots.ox.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 134

# FINDING AND NAVIGATING TO HUMANS IN COMPLEX ENVIRONMENTS FOR ASSISTIVE TASKS

Yaar A., Furnari A., Rosano M., Härmä A., Farinella GM

**Abstract:** Locating and navigating to humans in unseen environments is a major challenge for intelligent agents and social robots. Effective exploration and navigation strategies are necessary to locate the human performing various activities. In this paper, we propose an approach that uses state-of-the-art components to allow the agent to explore the environment, identify the human's location on the map, and approach them while maintaining a safe distance. We conducted experiments using the Gibson dataset and the Habitat AI simulator, where the proposed approach achieves promising results.

**Contact:** asfandasif93@gmail.com

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 135

# VID2SEQ: LARGE-SCALE PRETRAINING OF A VISUAL LANGUAGE MODEL FOR DENSE VIDEO CAPTIONING

Yang Antoine, Nagrani Arsha, Seo Paul, Miech Antoine, Pont-Tuset Jordi, Laptev Ivan, Sivic Josef, Schmid Cordelia

**Abstract:** Vid2Seq is a visual language model for dense video captioning, that predicts captions and timestamps for all events in an untrimmed video by generating a single sequence of tokens. Vid2Seq can be effectively pretrained on unlabeled narrated videos, by reformulating transcribed speech sentences and their corresponding boundaries as pseudo event captions and boundaries. The resulting Vid2Seq model pretrained on the YT-Temporal-1B dataset improves the SoTA on a variety of video captioning tasks.

**Contact:** antoineyang3@gmail.com

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 136

# TEXTMANIA: ENRICHING VISUAL FEATURES BY TEXT-DRIVEN MANIFOLD AUGMENTATION

Ye-Bin M., Kim J., Kim H., Son K., Oh T.-H.

**Abstract:** We propose TextManiA, a text-driven manifold augmentation methods that semantically enriches visual feature spaces, regardless of data distribution. Our method augments visual data with intra-class semantic perturbation by exploiting easy-to-understand visually mimetic words, i.e., attributes. To this end, we bridge between the text representation and a target visual feature space, and propose an efficient vector augmentation. Our experiments demonstrate that TextManiA is powerful in scarce samples with class imbalance as well as even distribution.

**Contact:** ybmoon@postech.ac.kr

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 137

# BEYOND PIXELS

Yousef Yeganeh

**Abstract:** Pixels are the common way to represent images, but they may not capture the full complexity of data, especially when we have prior knowledge about the image modality. For example, some images are acquired in the frequency domain, or some tasks require global information to process the images. In such cases, we can use other spaces, such as the frequency domain or the attention domain, to enrich the data representation. We explore this idea in two of our papers that use dual-encoder networks for image segmentation. Image segmentation is a task that assigns a label to each pixel in an image. The first paper, Y-Net, uses an FFC encoder and a CNN encoder to combine features from the frequency domain and the pixel domain. In DIAMANT paper, we use two CNN encoders to combine features from the attention domain and the pixel domain. We show that both papers achieve remarkable results, especially in segmenting challenging regions such as fluid pockets that are not compact and distributed through image.

**Contact:** y.yeganeh@tum.de

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 138

# SYNTHESIZING DIVERSE HUMAN MOTIONS IN 3D INDOOR SCENES

Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, Siyu Tang

**Abstract:** We present a novel method for populating 3D indoor scenes with virtual humans that can navigate the environment and interact with objects in a realistic manner. We propose a reinforcement learning based approach to learn policy networks that predict latent variables of a powerful generative motion model trained on large-scale motion capture datasets as actions to control humans to interact with scenes. Our method can synthesize realistic and diverse human-object interactions (e.g., sitting on a chair and then getting up) for out-of-distribution test scenarios and sequential performance of interaction event chains.

**Contact:** kaifeng.zhao@inf.ethz.ch

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 139

# EXPLOITING STABLE GRASPS FOR HAND-OBJECT RECONSTRUCTION IN EGOCENTRIC VIDEOS

Zhu Z., Damen D.

**Abstract:** We address in-the-wild hand-object reconstruction in egocentric videos. We curate a dataset of 914 sequences with stable grasp annotation, featuring 232 instances of 6 object categories. Assuming knowledge of the general object category, we reconstruct hand-object stable grasps using segmentation masks.

**Contact:** zhifan.zhu@bristol.ac.uk

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 140

# AI-BASED MONITORING SYSTEM FOR ENHACING INDUSTRIAL PROCESSES: A FOCUS ON VIALS INSPECTION

Zribi M., Pagliuca P.

**Abstract:** The specific focus of this study is to introduce an AI-based monitoring system within a production chain involved in manufacturing plastic consumables for analytical laboratories, specifically targeting the control of vials containing an anticoagulant substance. Two models will be proposed: one model capable of identifying the presence of the anticoagulant regardless of the analyzed vial, and a second model capable of recognizing both the vial type and the presence of the anticoagulant.

**Contact:** meriam.zribi@uniroma1.it

**Presentation Type:** Poster

**Date:** Tuesday 11 July 2023

**Time:** 21:30

**Poster Session:** 2

**Poster Number:** 141