POSTER SESSION BOOKLET

ICVSS 2024

International Computer Vision Summer School

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

University of Catania - University of Cambridge

International Computer Vision Summer School 2024

Computer Vision in the Age of Large Language Models

Sicily, 7 - 13 July 2024

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 2024.

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 2024

Roberto Cipolla
Sebastiano Battiato
Giovanni Maria Farinella

List of Posters [1]

1. COMPLEX 3D SCENES RETRIEVAL Abdari A, Falcon A., Serra G.

2. SYNTHESIZING DIVERSE COUNTERFACTUALS TO MITIGATE ASSOCIATIVE BIAS. Abdel Magid S., Wang J., Kafle K., Pfister H.

3. SIMCS: SIMULATION FOR DOMAIN INCREMENTAL ONLINE CONTINUAL SEGMENTATION Alfarra M., Cai Z, Bibi A., Ghanem B. , Müeller M.

4. COUNTGD: MULTI-MODAL OPEN-WORLD COUNTING Amini-Naieni N., Han T., Zisserman A.

5. DATADREAM: FEW-SHOT GUIDED DATASET GENERATION Kim J.M., Bader J., Alaniz S., Schmid C., Akata Z.

6. PHYSICAL UNDERSTANDING FROM THE SOUND OF POURING WATER Bagad P., Tapaswi M., Snoek C., Zisserman A.

7. MULTIMODAL-CONDITIONED LATENT DIFFUSION MODELS FOR FASHION IMAGE EDITING Baldrati A., Morelli D., Cornia M., Bertini M., Cucchiara R.

8. NEURAL PROCESSING OF TRI-PLANE HYBRID NEURAL FIELDS Cardace A., Zama Ramirez P., Ballerini F., Zhou A., Salti S., Di Stefano L.

9. MEASURING DEMENTIA BEHAVIOURS FROM DEPTH MAPS Ballester, I.

10. CONTRASTING DEEPFAKES DIFFUSION VIA CONTRASTIVE LEARNING AND GLOBAL-LOCAL SIMILARITIES Baraldi L., Cocchi F., Cornia M., Baraldi L., Nicolosi A., Cucchiara R.

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

11. TRAINING-FREE OPEN-VOCABULARY SEGMENTATION WITH OFFLINE DIFFUSION-AUGMENTED PROTOTYPE GENERATION Barsellotti L., Amoroso R., Cornia M., Baraldi L., Cucchiara R.

12. HYBRID FUNCTIONAL MAPS FOR CREASE-AWARE NON-ISOMETRIC SHAPE MATCHING Bastian L., Xie Y., Navab N., Lähner Z.

13. UNLEASHING THE POTENTIAL OF SYNTHETIC IMAGES: A STUDY ON HISTOPATHOLOGY IMAGE CLASSIFICATION Benito-del-Valle L., Alvarez-Gila A., Eguskiza I., Lopez-Saratxaga L.

14. THE DEVIL IS IN THE FINE-GRAINED DETAILS; EVALUATING OPEN-VOCABULARY OBJECT DETECTORS FOR FINE-GRAINED UNDERSTANDING Bianchi L., Carrara F., Messina N., Gennaro C., Falchi F.

15. HERO AND HERO-GPT:MULTI-MODAL APPROACHES ON MOBILE DEVICES FOR VISUAL-AWARE CONVERSATIONAL ASSISTANCE IN INDUSTRIAL DOMAINS Bonanno C., Strano L., Ragusa F., Furnari A., Farinella G.M.

16. BIOMEDICAL SHAPE RECONSTRUCTION VIA NEURAL DEFORMATION FIELDS Bongratz F., Rickmann A.-M., Wachinger C.

17. GAUSSIAN PANCAKES: GEOMETRICALLY-REGULARIZED 3D GAUSSIAN SPLATTING FOR REALISTIC ENDOSCOPIC RECONSTRUCTION Bonilla S, Zhang S, Psychogyios D, Stoyanov D, Vasconcelos F, & Bano S

18. PROMPT LEARNING FOR CONTINUAL WSI ANALYSIS Bontempo G., Lumetti L., Porrello A., Bolelli F., Calderara S., Ficarra E.

19. MUSES: THE MULTI-SENSOR SEMANTIC PERCEPTION DATASET Brodermann T., Bruggemann D., Sakaridis C., Ta K., Liagouris O., Corkill J., Van Gool L.

20. DO LANGUAGE MODELS UNDERSTAND MORALITY? Bulla L., Mongiovì M., Gangemi A.

21. SMART VIRTUAL RESPIRATORY WARD Burrows H., Maktabdar Oghaz M., Babu-Saheer L.

22. A SECOND-ORDER PERSPECTIVE ON COMPOSITIONALITY AND INCREMENTAL LEARNING Porrello A., Bonicelli L., Buzzega P., Millunzi M., Calderara S., Cucchiara R.

23. TRENDS, APPLICATIONS AND CHALLENGES IN HUMAN ATTENTION MODELLING Cartella G., Cuculo V., Cornia M., Cucchiara R.

24. SPAMMING LABELS: EFFICIENT ANNOTATIONS FOR THE TRACKERS OF TOMORROW Cetintas O., Meinhardt T., Brasó G., Leal-Taixé L.

25. A LIE GROUP APPROACH TO RIEMANNIAN BATCH NORMALIZATION Chen Z., Song Y., Liu Y., Sebe, N.

26. EMOTIONAL SPEECH DRIVEN 3D BODY ANIMATION VIA DISENTANGLED LATENT DIFFUSION Chhatre K., Danecek R., Athanasiou N., Becherini G., Peters C., Black M., Bolkart T.

27. UNI: UNLEARNING-BASED NEURAL INTERPRETATIONS Choi C.L., Duplessis A., Belongie S.

28. WEEDS AS ANOMALIES: ANOMALY DETECTION FOR WEED DETECTION IN AGRICULTURAL FIELDS Chong Y. L., Behley J., and Stachniss C.

29. VOLUMETRIC PRIMITIVES FOR MODELING AND RENDERING SCATTERING AND EMISSIVE MEDIA Condor , Jorge., Speierer, Sebastien., Bode, Lukas., Bozic, Aljaz., Green, Simon., Didyk, Piotr., Jarabo, Adrian.

30. TRANSFERRING DISENTANGLED REPRESENTATIONS Dapueto J.

31. VISION TRANSFORMERS NEED REGISTERS Darcet T., Oquab M., Mairal J., Bojanowski P.

32. SEMANTICALLY GUIDED REPRESENTATION LEARNING FOR ACTION ANTICIPATION Diko A., Avola D., Prenkaj B., Fontana F., Cinque L.

33. MEMFLOW: OPTICAL FLOW ESTIMATION AND PREDICTION WITH MEMORY Dong Q., Fu Y.

34. DON'T DROP YOUR SAMPLES! COHERENCE-AWARE TRAINING BENEFITS CONDITIONAL DIFFUSIONNAL Nicolas D., Victor B., Vicky K., David P.

35. SHAPE ESTIMATION OF SEMI- FLUID DEFORMABLE OBJECTS EL ASSAL O., M. MATTEO C., CIRON S., FOFI D.

36. INVISIBLE STITCH: GENERATING SMOOTH 3D SCENES WITH DEPTH INPAINTING Engstler P., Vedaldi A., Laina I., Rupprecht C.

37. CONTROLLING STYLE IN DIFFUSION MODELS THROUGH NOISE Everaert M.N., Süsstrunk S., Achanta R.

38. EXPLORING THE SYNERGY BETWEEN VISION-LANGUAGE PRE-TRAINING AND CHATGPT FOR ARTWORK CAPTIONING Castellano G., Fanelli N., Scaringi R., Vessio G.

39. OPENMASK3D: OPEN-VOCABULARY 3D INSTANCE SEGMENTATION Takmaz A., Fedele E., Sumner B., Pollefeys M., Tombari F., Engelmann F.

40. HOW TO PROBE: SIMPLE YET EFFECTIVE TECHNIQUES FOR IMPROVING POST-HOC EXPLANATIONS Gairola S., Böhle M., Locatello F., Schiele B.

41. GRAPHDREAMER: COMPOSITIONAL 3D SCENE SYNTHESIS FROM SCENE GRAPHS Gao G., Liu W., Chen A., Geiger A., Schölkopf B.

42. GEN2BALANCE USING GENERATIVE MODELS FOR LONG-TAIL VIDEO ACTION RECOGNITION Gatti P., Jenni S., Caba Heilbron F., Collomosse J., Damen D.

43. MPL: LIFTING 3D POSE FROM MULTI-VIEW 2D POSES Ghasemzadeh S.A., De Vleeschouwer C.

44. THE SHORT FILM DATASET (SFD): A BENCHMARK FOR STORY-LEVEL VIDEO UNDERSTANDING GHERMI R., Wang X., Kalogeiton V., Laptev I.

45. LARGE-SCALE GLOBAL LOCALIZATION USING TRANSFORMERS Gladkova M.

46. BAYES' RAYS: UNCERTAINTY QUANTIFICATION FOR NEURAL RADIANCE FIELDS Goli L., Reading C., Sellán S., Jacobson A., Tagliasacchi A.

47. LABEL EFFICIENT LEARNING PARADIGMS FOR SCALABLE ROBOT PERCEPTION Gosala N., Petek K., Besic B., Cattaneo D., Drews-Jr P., Kiran BR., Yogamani S., Burgard W., Valada A.

48. CALVING FRONTS AND WHERE TO FIND THEM – A BENCHMARK FRAMEWORK Gourmelon N., Seehaus T., Braun M., Maier A., Christlein V.

49. IMPARA: INDUSTRIAL MULTI-TASK PHOTO-REALISTIC ANNOTATED DATASET FOR ROBOTICS APPLICATIONS Govi E., Sapienza D., De Dominicis L., Capodieci N., Bertogna M.

50. BOOSTING UNSUPERVISED SEMANTIC SEGMENTATION WITH PRINCIPAL MASK PROPOSALS Hahn O., Araslanov N., Schaub-Meyer S., Roth S.

51. DIFFCD: A SYMMETRIC DIFFERENTIABLE CHAMFER DISTANCE FOR NEURAL IMPLICIT SURFACE FITTING Härenstam-Nielsen L., Sang L., Saroha A., Araslanov N., Cremers D.

52. ALL U-NEED IS U-NET(S) Helander R., Parvizi J.

53. SOAC: SPATIO-TEMPORAL OVERLAP-AWARE MULTI-SENSOR CALIBRATION USING NEURAL RADIANCE FIELDS Herau Q., Piasco N., Bennehar M., Roldão L., Tsishkou D., Migniot C., Vasseur P., Demonceaux C.

54. EVALUATING XAI METHODS Hesse R., Schaub-Meyer S., Roth S.

55. MULTI-MODAL GAZE FOLLOWING IN CONVERSATIONAL SCENARIOS Hou Y., Zhang Z., Horanyi N., Cheng Y., Chang H.J., University of Birmingham Moon, J., KETI

56. RECONSTRUCTING MULTI-BODY DYNAMICS FOR COMPOSITIONAL RE-SIMULATION HUANG S.,

57. 4M-21: AN ANY-TO-ANY VISION MODEL FOR TENS OF TASKS AND MODALITIES Kar O.F., Bachmann R., Mizrahi D., Garjani A., Gao M., Griffiths D., Hu J., Dehghan A., Zamir A.

58. PROTECTING DATA PRIVACY OF DIFFUSION MODELS Katsumata K., Nakayama H.

59. METTA: SINGLE-VIEW TO 3D TEXTURED MESH RECONSTRUCTION WITH TEST-TIME ADAPTATION Yu-Ji K., Ha H., Youwang K., Surh J., Ha H., Oh T.-H.

60. OPEN-VOCABULARY 3D SCENE GRAPHS Koch S, Vaskevicius N, Colosi M, Hermosilla P, Ropinski T

61. HOLONETS: SPECTRAL CONVOLUTIONS DO EXTEND TO DIRECTED GRAPHS Kokce C., Cremers D.

62. LEARNING GEOMETRY: ROTATION REPRESENTATION OF SYMMETRIC OBJECTS Kriegler A., Beleznai C., Gelautz M.

63. DFLOW: LEARNING TO SYNTHESIZE BETTER OPTICAL FLOW DATASETS VIA A DIFFERENTIABLE PIPELINE Byung-Ki K., Hyeon-Woo N., Kim J.-Y., Oh T.-H.

64. GALLOP: LEARNING GLOBAL AND LOCAL PROMPTS FOR VISION-LANGUAGE MODELS Lafon M., Ramzi E., Rambour C., Audebert N., Thome N.

65. FASTER THAN LIES: REAL-TIME DEEPFAKE DETECTION USING BINARY NEURAL NETWORKS Lanzino R., Fontana F., Diko A., Marini M. R., Cinque L.

66. GEOMETRY-AWARE PROJECTIVE MAPPING FOR UNBOUNDED NEURAL RADIANCE FIELDS Lee J., Jung H., Park J., Bae I., Jeon H.

67. ARE SYNTHETIC DATA USEFUL FOR EGOCENTRIC HAND-OBJECT INTERACTION DETECTION? Leonardi R., Furnari A., Ragusa F., Farinella G.M.

68. MAPPING EYE-TRACKING DATA ACROSS INDEPENDENT EGOCENTRIC VIDEO STREAMS LEON-CONTRERAS N., DIERKES K.

69. EGOGEN: AN EGOCENTRIC SYNTHETIC DATA GENERATOR Li G., Zhao K., Zhang S., Lyu X., Dusmanu M., Zhang Y., Pollefeys M., Tang S.

70. MULTI-IMAGE VISUAL QUESTION ANSWERING FOR UNSUPERVISED ANOMALY DETECTION Li J., Bercea C.I., Müller P., Felsner L., Kim S.,Rueckert D., Wiestler B., Schnabel J.A.

71. EXPANSION AND SHRINKAGE OF LOCALIZATION FOR WEAKLY-SUPERVISED SEMANTIC SEGMENTATION Li J., Jie Z., Wang X., Wei X., Ma L.

72. EVCAP: RETRIEVAL-AUGMENTED IMAGE CAPTIONING WITH EXTERNAL VISUAL-NAME MEMORY FOR OPEN-WORLD COMPREHENSION Li J., Vo D., Sugimoto A., Nakayama H.

73. VSTAR: GENERATIVE TEMPORAL NURSING FOR LONGER DYNAMIC VIDEO SYNTHESIS Li Y., Beluch W., Keuper M., Zhang D., Khoreva A.

74. OBJECT TRACKING WITH EVENTS AND FRAMES Li Z., Piga N., Pietro F., Iacono M., Glover A., Natale L., Bartolozzi C.

75. GAUSSIAN-SLAM: PHOTO-REALISTIC DENSE SLAM WITH GAUSSIAN SPLATTING Yugay V., Li Y., Gevers T., Oswald M.R.

76. UNCERTAINTY-GUIDED OPEN-SET SOURCE-FREE UNSUPERVISED DOMAIN ADAPTATION WITH TARGET-PRIVATE CLASS SEGREGATION Litrico M., Talon D., Battiato S., Del Bue A., Giuffrida M. V., Morerio P.

77. DEMOCRATIZING FINE-GRAINED VISUAL RECOGNITION WITH LARGE LANGUAGE MODELS Liu M. , Roy S. , Li W. , Zhong Z. , Sebe N., Ricci E.

78. LONG-TERM ACTION PREDICTION IN PROCEDURAL TASKS OF LONG DURATION Loureiro M.

79. DN-CBM: TASK AGNOSTIC CONCEPT BOTTLENECKS Rao S.*, Mahajan S.*, Böhle M., Schiele B.

80. DEEP LEARNING-BASED BRAIN SHAPE MORPHOLOGY ANALYSIS AND CORRELATES BETWEEN BRAIN AND FACIAL BIOMARKERS FOR THE DIAGNOSIS OF PSYCHOTIC DISORDERS Malé J., Martínez-Abadías N., Sevillano X.

81. DUAL-AXIS TILTING ROTOR QUAD-PLANE DESIGNED FOR PRECISION LANDING IN GUSTY MARITIME ENVIRONMENTS Mancinelli A., Smeur E., Remes B., De Croon G.

82. IS MULTIPLE OBJECT TRACKING A MATTER OF SPECIALIZATION? Mancusi G., Bernardi M., Panariello A., Porrello A., Calderara S., Cucchiara R.

83. ZERO SHOT NOISE2NOISE: EFFICIENT IMAGE DENOISING WITHOUT ANY DATA Mansour Y., Heckel R.

84. SEMI-SUPERVISED MULTI-TASK HYPERGRAPHS FOR EARTH UNDERSTANDING Marcu A., Pirvu M., Costea D., Haller E., Slusanschi E., Belbachir N., Sukthankar R., Leordeanu M.

85. ON GOOD PRACTICES FOR TASK-SPECIFIC DISTILLATION OF LARGE PRETRAINED VISUAL MODELS Marrie J., Arbel M., Mairal J., Larlus D.

86. EYES WIDE UNSHUT: UNSUPERVISED MISTAKE DETECTION IN EGOCENTRIC VIDEO BY DETECTING UNPREDICATABLE GAZE Mazzamuto M., Furnari A., Farinella G. M.

87. IMPROVING ZERO-SHOT GENERALIZATION OF PROMPT LEARNING VIA UNSUPERVISED KNOWLEDGE DISTILLATION Mistretta M., Baldrati A., Bertini M., Bagdanov A.

88. EGOCENTRIC VISION FOR ACTIVE ASSISTED LIVING Mucha W.

89. AFFORDANCES AND ATTENTION MODELS FOR SHORT-TERM OBJECT INTERACTION ANTICIPATION Mur-Labadia L., Martinez-Cantin R, Guerrero JJ, Farinella G.M., Furnari A.

90. MACHINE VISION IN INDUSTRIAL APPLICATIONS Neubauer M., Rueckert E.,

91. SCANTALK: 3D TALKING HEADS FROM UNREGISTERED SCANS Nocentini F., Besnier T., Ferrari C., Arguillere S., Berretti S., Daoudi M.

92. LOWFORMER: HARDWARE EFFICIENT DESIGN FOR CONVOLUTIONAL TRANSFORMER BACKBONE Nottebaum M., Dunnhofer M., Prof. Micheloni C.

93. SEGMENTING PUPILS WITH ELLIPTICAL FOURIER SERIES O'MAHONY F., DIERKES K., DREWS M.

94. FEW-SHOT UNSUPERVISED IMPLICIT NEURAL SHAPE REPRESENTATION LEARNING WITH SPATIAL ADVERSARIES Ouasfi A., Boukhayma A.

95. VISION-LANGUAGE MODELS FOR HOLISTIC OR DOMAIN MODELING Özsoy E., Pellegrini C., Keicher M., Czempiel T., Örnek P. E., Navab N.

96. LUCIDPPN: UNAMBIGUOUS PROTOTYPICAL PARTS NETWORK FOR USER-CENTRIC INTERPRETABLE COMPUTER VISION Pach M., Rymarczyk D., Lewandowska K., Tabor J., Zieliński B.

97. GENERATIVE TIMELINES FOR INSTRUCTED VISUAL ASSEMBLY Pardo A., Wang J., Ghanem B., Sivic J., Russell B., Caba F.

98. LONG-TERM TYPHOON TRAJECTORY PREDICTION: A PHYSICS-CONDITIONED APPROACH WITHOUT REANALYSIS DATA Park Y., Seo M., Kim D., Kim H., Choi S., Choi B., Ryu J., Son S., Jeon H., Choi Y.

99. GENERALIZED DEPTH PERCEPTION FROM EVERYDAY SENSORS Park J., Jeon H.

100. HARLEQUIN: COLOR-DRIVEN GENERATION OF SYNTHETIC DATA FOR REFERRING EXPRESSION COMPREHENSION Parolari L., Izzo E., Ballan L.

101. BREAKING THROUGH FAILURE BARRIERS OF STRUCTURE-FROM-MOTION Pataki Z.

102. VISION-LANGUAGE MODELS FOR MEDICAL IMAGE UNDERSTANDING Pellegrini C., Keicher M., Özsoy E., Busam B., Navab N.

103. ROBUST WIND TURBINE BLADE SEGMENTATION FROM RGB IMAGES IN THE WILD Pérez-Gonzalo R., Espersen A., Agudo A.

104. PREGO:ONLINE MISTAKE DETECTION IN PROCEDURAL EGOCENTRIC VIDEOS Flaborea A., D'Amely G., Plini L., Scofano L., De Matteis E., Furnari A., Farinella G., Galasso F.

105. DEEPFEATUREX NET: DEEP FEATURES EXTRACTORS BASED NETWORK FOR DISCRIMINATING SYNTHETIC FROM REAL IMAGES Pontorno O., Guarnera L., Battiato S.

106. MERGING AND SPLITTING DIFFUSION PATHS FOR SEMANTICALLY COHERENT PANORAMAS Quattrini F., Pippi V., Cascianelli S., Cucchiara R.

107. SYNCHRONIZATION IS ALL YOU NEED: EXOCENTRIC-TO-EGOCENTRIC TRANSFER FOR TEMPORAL ACTION SEGMENTATION WITH UNLABELED SYNCHRONIZED VIDEO PAIRS Quattrocchi C., Furnari A., Di Mauro D., Giuffrida M. V., Farinella G. M.

108. THE SIMPLER THE BETTER: AN ENTROPY-BASED IMPORTANCE METRIC TO REDUCE DEEP NEURAL NETWORKS' DEPTH Quétu V., Liao Z., Tartaglione E.

109. DETECTION OF IMAGE MANIPULATIONS USING DCT STATISTICS Ragaglia C., Guarnera F., Battiato S., Puglisi G.

110. ENIGMA-51: TOWARDS A FINE-GRAINED UNDERSTANDING OF HUMAN BEHAVIOR IN INDUSTRIAL SCENARIOS Ragusa F., Leonardi R., Mazzamuto M., Bonanno C., Scavo R., Furnari A., Farinella G. M.

111. BEYOND ARTIFACTS: IMPROVED DEEP LEARNING-POWERED MAR WITH DOMAIN TRANSFORMATION FOR RADIOTHERAPY PLANNING Rehman M.,Micheloni C.,Martinel N.,Spizzo R.,Avanzo M.

112. EASG: ACTION SCENE GRAPHS FOR LONG-FORM UNDERSTANDING OF EGOCENTRIC VIDEOS Rodin I., Furnari A., Min K., Tripathi S., Farinella G.M.

113. CROWDSIM++: UNIFYING CROWD NAVIGATION AND OBSTACLE AVOIDANCE Rosano M., Leocata D., Furnari A., Farinella G. M.

114. ICICLE: INTERPRETABLE CLASS INCREMENTAL CONTINUAL LEARNING Rymarczyk D., van de Weijer J., Zieliński B., Twardowski B.

115. FOLLOWING THE HUMAN THREAD IN SOCIAL NAVIGATION Scofano L.*, Sampieri A.*, Campari T.*, Sacco V., Spinelli I., Ballan L., Galasso F.

116. SCALING ANOMALY DETECTION TO WIDE IMAGES Samele S., Matteucci M.

117. TOWARDS SUSTAINABLE EV BATTERY RECYCLING: CV AND RL BASED BATTERY DISASSEMBLY Sarno F., Isoaho J., Özen Ö., Rätz R., Ochsenbein C.

118. TOWARDS AUTOMATED REGULATION OF JACOBAEA VULGARIS IN GRASSLAND USING DEEP NEURAL NETWORKS Schauer M., Hohl R., Vaupel D., Soethe N., Ghobadi S. E.

119. TWO EFFECTS, ONE TRIGGER: ON THE MODALITY GAP, OBJECT BIAS, AND INFORMATION IMBALANCE IN CONTRASTIVE VISION-LANGUAGE REPRESENTATION LEARNING Schrodi S., Hoffmann D.T., Argus M., Fischer V., Brox T.

120. DIFFERENTIABLE TASK GRAPH LEARNING: PROCEDURAL ACTIVITY REPRESENTATION AND ONLINE MISTAKE DETECTION FROM EGOCENTRIC VIDEOS Seminara L., Farinella G. M., Furnari A.

121. MAGIC: MODULAR AUTOENCODER FOR GENERALISABLE MODEL INVERSION WITH BIAS CORRECTIONS She Y., Atzberger C., Blake A., Gualandi A., Keshav S.

122. CLOSE IMITATION OF EXPERT RETOUCHING FOR BLACK-AND-WHITE PHOTOGRAPHY Shin S., Shin J., Bae J., Shim I., Jeon H.

123. LEARNING VIDEO-LANGUAGE MODELS WITH LIMITED SUPERVISION Shvetsova N., Kukleva A., Hong X., Rupprecht C., Schiele B., Kuehne H.

124. AR-TTA: A SIMPLE METHOD FOR REAL-WORLD CONTINUAL TEST-TIME ADAPTATION Sójka D., Cygert S., Twardowski B., Trzciński T.

125. COMPRESSED-LANGUAGE MODELS FOR UNDERSTANDING COMPRESSED FILE FORMATS: A JPEG EXPLORATION Pérez J., Pardo A., Soldan M., Itani M., Leon-Alcazar J., Ghanem B.

126. WHEN DOES PERCEPTUAL ALIGNMENT BENEFIT VISION REPRESENTATIONS? Sundaram S., Fu S

127. EXPLOITING ACTIVATION SPARSITY WITH DENSE TO DYNAMIC-K MIXTURE-OF-EXPERTS CONVERSION Szatkowski F., Wójcik B., Piórczyński M., Scardapane S.

128. THE ILLUSION OF REALITY: REAL-TIME REALISM IN AR/VR Tariq T.

129. LEARNING OBJECT STATES FROM ACTIONS VIA LARGE LANGUGE MODELS Tateno M., Yagi T., Furuta R., Sato Y.

130. FANTASTIC GAINS AND WHERE TO FIND THEM Roth K., Thede L., Koepke A. S., Vinyals O., Henaff O., Akata Z.

131. CLIMB3D: CONTINUAL LEARNING FOR IMBALANCED 3D INSTANCE SEGMENTATION Thengane V., Lahoud J., Khan S., Khan F., Li Y.

132. FROM CORRESPONDENCES TO POSE: NON-MINIMAL CERTIFIABLY OPTIMAL RELATIVE POSE WITHOUT DISAMBIGUATION Tirado-Garín J., Civera J.

133. PHYSICS-INFORMED MODELING OF DYNAMIC HUMANS AND THEIR INTERACTIONS TRIPATHI S., TZIONAS D., BLACK M.

134. RELATIVE POSE FROM 4 POINTS IN 3 VIEWS Tzamos C., Barath D., Sattler T., Kukelova Z.

135. WHAT DO MULTIMODAL MODELS KNOW ABOUT MOTION BLUR? Vartore A., Ortis A., Mahiddine A., Battiato S.

136. CONVEX DECOMPOSITION OF INDOOR SCENES Vavilala V., Forsyth D.

137. CAMERA POSE ESTIMATION FROM BOUNDING BOXES Vávra V., Sattler T., Kúkelová Z.

138. COMIX: A COMPREHENSIVE BENCHMARK FOR MULTITASK COMICS UNDERSTANDING Vivoli E., Bertini M., Karatzas D.

139. WHAT DO NEURAL NETWORKS LEARN IN IMAGE CLASSIFICATION? A FREQUENCY SHORTCUT PERSPECTIVE Wang S., Veldhuis R., Brune C., Strisciuglio N.

140. CNVID-3.5M: BUILD, FILTER, AND PRE-TRAIN THE LARGE-SCALE PUBLIC CHINESE VIDEO-TEXT DATASET Gan T., Wang Q., Dong X., Ren X., Nie L., Guo Q.

141. FLATTENING THE PARENT BIAS: HIERARCHICAL SEMANTIC SEGMENTATION IN THE POINCARÉ BALL Weber S., Zöngür B., Araslanov N., Cremers D.

142. T-DEED: TEMPORAL-DISCRIMINABILITY ENHANCER ENCODER-DECODER FOR PRECISE EVENT SPOTTING IN SPORTS VIDEOS Xarles A., Escalera S., Moeslund T., Clapés A.

143. DEMOCRATIZING HUMAN DIGITIZATION Xiu Y.

144. ENHANCING VIDEO SUPER-RESOLUTION VIA IMPLICIT RESAMPLING-BASED ALIGNMENT (CVPR2024 HIGHTLIGHT) Xu K., Yu Z., Wang X., Bi M, Yao A.

145. UNIMATCH: UNIFIED DENSE MATCHING FOR FLOW, STEREO, DEPTH AND GAUSSIAN SPLATTING Xu H., Pollefeys M., Geiger A.

146. KNOWLEDGE DECOMPOSITION AND REPLAY: A NOVEL CROSS-MODAL IMAGE-TEXT RETRIEVAL CONTINUAL LEARNING METHOD Yang R., Wang S., Zhang H., Xu S., Guo Y., Ye X., Hou B., Jiao L.

147. FIRE : FAST INVERSE RENDERING USING DIRECTIONAL AND SIGNED DISTANCE FUNCTIONS Yenamandra T., Tewari A., Yang N., Bernard F., Theobalt C., Cremers D.

148. IMAGE DEBLURRING GUIDED BY DEPTH MAP PROMPT Yi Z., Valsesia D., Bianchi T., Magli E

149. FEDERATED LEARNING FOR DECENTRALIZED MODEL TRAINING IN SKIN CANCER HISTOPATHOLOGY Zacharczuk J., Pisula J., Bozek K.

150. EFFICIENCY THROUGH SELF SUPERVISION IN GERIATRICS Zedda L.

151. ROAM: ROBUST AND OBJECT-AWARE MOTION GENERATION USING NEURAL POSE DESCRIPTORS Zhang W., Dabral R., Leimkühler T., Golyanik V., Habermann M., Theobalt C.

152. MASKED VIDEO AND BODY-WORN IMU AUTOENCODER FOR EGOCENTRIC ACTION RECOGNITION Zhang M., Huang Y., Liu R., Sato Y.

COMPLEX 3D SCENES RETRIEVAL

Abdari A, Falcon A., Serra G.

Abstract: In recent years, the study of complex 3D scenes has garnered significant attention from researchers across various fields. These scenes, characterized by their multiple areas and diverse objects, present unique challenges and opportunities. One notable application is 3D scene retrieval, where textual descriptions are also utilized in the retrieval process. During my Ph.D. program, I have focused extensively on complex 3D scenes Retrieval task, which I will present in greater detail.

Contact: abdari.ali@spes.uniud.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 1

SYNTHESIZING DIVERSE COUNTERFACTUALS TO MITIGATE ASSOCIATIVE BIAS.

Abdel Magid S., Wang J., Kafle K., Pfister H.

Abstract: We propose a framework to generate synthetic counterfactual images to create a diverse and balanced dataset that can be used to fine-tune CLIP. We leverage segmentation and inpainting models to place humans with diverse visual appearances in context. We show that CLIP trained on such data learns to disentangle the human appearance from the context of an image, i.e., what makes a doctor is not correlated to the person's visual appearance, like skin color or body type, but to the context.

Contact: sabdelmagid@g.harvard.edu

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 2

SIMCS: SIMULATION FOR DOMAIN INCREMENTAL ONLINE CONTINUAL SEGMENTATION

Alfarra M., Cai Z, Bibi A., Ghanem B. , Müeller M.

Abstract: Continual Learning is a step towards lifelong intelligence where models continuously learn from recently collected data without forgetting previous knowledge. Existing continual learning approaches mostly focus on image classification in the class-incremental setup with clear task boundaries and unlimited computational budget. This work explores the problem of Online Domain-Incremental Continual Segmentation (ODICS), where the model is continually trained over batches of densely labeled images from different domains, with limited computation and no information about the task boundaries. ODICS arises in many practical applications. In autonomous driving, this may correspond to the realistic scenario of training a segmentation model over time on a sequence of cities. We analyze several existing continual learning methods and show that they perform poorly in this setting despite working well in class-incremental segmentation. We propose SimCS, a parameter-free method complementary to existing ones that uses simulated data to regularize continual learning. Experiments show that SimCS provides consistent improvements when combined with different CL methods.

Contact: motasem.alfarra@kaust.edu.sa

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 3

COUNTGD: MULTI-MODAL OPEN-WORLD COUNTING

Amini-Naieni N., Han T., Zisserman A.

Abstract: This work aims to improve the generality and accuracy of open-vocabulary object counting in images. To improve the generality, we repurpose a detection foundation model (GroundingDINO) for counting and extend its capabilities by introducing modules to enable specifying the object to count by visual exemplars. In turn, this new capability - being able to specify the object by multiple-modalites (text and exemplars) - gives both greater flexibility and control, and improves the counting accuracy.

Contact: niki.amini-naieni@eng.ox.ac.uk

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 4

DATADREAM: FEW-SHOT GUIDED DATASET GENERATION

Kim J.M., Bader J., Alaniz S., Schmid C., Akata Z.

Abstract: Text-to-image diffusion models achieve SOTA image synthesis results, but they have yet to prove effective in downstream applications. We propose DataDream, which more faithfully represents the data distribution when guided by few-shot examples. It fine-tunes LoRA weights for the image generation model on few real images before generating training data. We demonstrate DataDream's efficacy through extensive experiments, surpassing SOTA classification with few-shot data on 9 /10 datasets.

Contact: jessica.bader@helmholtz-munich.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 5

PHYSICAL UNDERSTANDING FROM THE SOUND OF POURING WATER

Bagad P., Tapaswi M., Snoek C., Zisserman A.

Abstract: • Can we infer physical properties, e.g., container size, merely from the sound of pouring water in it? • Psychoacoustics shows humans can infer the time-to-fill, temperature of water, etc., surprisingly well [1, 2] • We present an analysis-by-synthesis approach to train a model to detect pitch in synthetic sounds of pouring • On real data, we demonstrate estimation of container height (MAE: 2.23cms), time-to-fill (MAE: 1.53s), etc., from sound alone

Contact: piyushnbagad11@gmail.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 6

MULTIMODAL-CONDITIONED LATENT DIFFUSION MODELS FOR FASHION IMAGE EDITING

Baldrati A., Morelli D., Cornia M., Bertini M., Cucchiara R.

Abstract: Fashion illustration is used by designers to communicate their vision and to bring the design idea from conceptualization to realization. In the context of fashion design, computer vision techniques have the potential to enhance the design process. To speed up the prototyping phase, we start a new computer vision task and propose a latent diffusion-based approach (Ti-MGD) to generate human-centric fashion guided by multimodal prompts, including text, garment sketches, and fabric textures.

Contact: alberto.baldrati@unifi.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 7

NEURAL PROCESSING OF TRI-PLANE HYBRID NEURAL FIELDS

Cardace A., Zama Ramirez P., Ballerini F., Zhou A., Salti S., Di Stefano L.

Abstract: Recent works explore the processing of neural fields for 3D tasks like classification and segmentation. Traditional methods using large MLPs struggle due to the high dimensionality and symmetries of the weight space, resulting in poorer performance than explicit representations. This paper shows that triplanes, a hybrid representation, can be effectively processed by standard deep learning machinery, while almost closing the performance gap between implicit and explicit representations.

Contact: francesco.ballerini4@unibo.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 8

MEASURING DEMENTIA BEHAVIOURS FROM DEPTH MAPS

Ballester, I.

Abstract: Human behaviour can be analysed from images, video or 3D representations such as point clouds. In dementia, tracking behavioural changes is key to assessing cognitive decline. To make our methods suitable for this purpose, we mitigate privacy concerns by using depth data. However, existing depth-based methods lack real-world applicability. Our work aims to bridge this gap by improving methods for different modalities derived from depth maps and by developing domain adaptation techniques.

Contact: irene.ballester@tuwien.ac.at

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 9

CONTRASTING DEEPFAKES DIFFUSION VIA CONTRASTIVE LEARNING AND GLOBAL-LOCAL SIMILARITIES

Baraldi L., Cocchi F., Cornia M., Baraldi L., Nicolosi A., Cucchiara R.

Abstract: Separating authentic content and AI-generated images is increasingly difficult. Solutions using foundation models like CLIP are not ideal for deepfake detection, lacking specialized training and local image features. We propose **{**Co}ntrastive **{**D}eepfake **{**E}mbeddings (**{**CoDE}), an embedding space tailored for deepfake detection, trained via contrastive learning with global-local similarities on an in-house dataset of 9.2 million generated images.

Contact: lorenzo.baraldi@phd.unipi.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 10

TRAINING-FREE OPEN-VOCABULARY SEGMENTATION WITH OFFLINE DIFFUSION-AUGMENT PROTOTYPE GENERATION

Barsellotti L., Amoroso R., Cornia M., Baraldi L., Cucchiara R.

Abstract: In Unsupervised Open-Vocabulary Segmentation (UOVS) the model is asked to segment an image according to a set of textual categories, without training on dense annotations. Previous works induce multimodal pixel-level alignment by training on image-caption pairs. However, they lack a direct source of supervision about the localization of concepts. We propose FreeDA, a training-free method that performs UOVS on the visual modality by retrieving a support set of generated synthetic references.

Contact: luca.barsellotti@unimore.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 11

HYBRID FUNCTIONAL MAPS FOR CREASE-AWARE NON-ISOMETRIC SHAPE MATCHING

Bastian L., Xie Y., Navab N., Lähner Z.

Abstract: Non-isometric shape correspondence is a fundamental challenge in computer vision. Traditional methods using Laplace-Beltrami operator (LBO) eigenmodes face limitations in characterizing fine details. We propose to combine the non-orthogonal extrinsic basis of eigenfunctions of an elastic thin-shell energy with the intrinsic ones of the LBO, creating a hybrid spectral space in which we construct functional maps. Our approach can be incorporated easily into existing functional map pipelines across varying applications and can handle complex deformations beyond isometries.

Contact: lennart.bastian@tum.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 12

UNLEASHING THE POTENTIAL OF SYNTHETIC IMAGES: A STUDY ON HISTOPATHOLOGY IMAGE CLASSIFICATION

Benito-del-Valle L., Alvarez-Gila A., Eguskiza I., Lopez-Saratxaga L.

Abstract: Histopathology image classification is key for accurate disease identification, but obtaining diverse annotated datasets is challenging due to the need for expert annotations and ethical constraints. Thus, we analysed methods to synthesize realistic histopathology images, focusing on selective data augmentation with various image selection techniques. Interestingly, high quality synthetic data does not guarantee an improvement on a downstream task. Instead, emphasis should be placed on the generative model type (GAN or DDPM) and its architecture (CNN or Transformer).

Contact: leire.benitodelvalle@tecnalia.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 13

THE DEVIL IS IN THE FINE-GRAINED DETAILS; EVALUATING OPEN-VOCABULARY OBJECT DETECTORS FOR FINE-GRAINED UNDERSTANDING

Bianchi L., Carrara F., Messina N., Gennaro C., Falchi F.

Abstract: Open-vocabulary object detectors (OVDs) excel at locating objects from arbitrary categories defined by free-text but struggle to recognize fine-grained attributes like materials. Existing OVD benchmarks do not assess performance on these properties. We introduce a benchmark with fine-grained captions and an evaluation protocol based on dynamic vocabulary generation to test whether models detect and assign the correct fine-grained description to objects in the presence of hard-negative classes.

Contact: lorenzo.bianchi@isti.cnr.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 14

HERO AND HERO-GPT:MULTI-MODAL APPROACHES ON MOBILE DEVICES FOR VISUAL-AWARE CONVERSATIONAL ASSISTANCE IN INDUSTRIAL DOMAINS

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

Abstract: We present HERO and HERO-GPT, two artificial assistants 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 systems in an industrial laboratory. In this setting, our systems 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 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 15

BIOMEDICAL SHAPE RECONSTRUCTION VIA NEURAL DEFORMATION FIELDS

Bongratz F., Rickmann A.-M., Wachinger C.

Abstract: Reconstructing anatomical shapes from 3D biomedical scans is essential for morphological analyses and for the computation of clinical biomarkers. To render the shapes locally comparable, corresponding surface points must be estimated precisely and matched to a template. To this end, we developed a geometric deep learning framework that learns to warp generic input templates to individual shape contours based on features extracted from magnetic resonance (MR) or computed tomography (CT) scans.

Contact: fabi.bongratz@tum.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 16

GAUSSIAN PANCAKES: GEOMETRICALLY-REGULARIZED 3D GAUSSIAN SPLATTING FOR REALISTIC ENDOSCOPIC RECONSTRUCTION

Bonilla S, Zhang S, Psychogyios D, Stoyanov D, Vasconcelos F, & Bano S

Abstract: Standard 3D reconstruction techniques struggle with reflective surfaces, lack of texture, and irregular illumination in endoscopic imaging, leading to inaccurate 3D reconstructions (Fig.1c). While often considered solved, 3D reconstruction within the human body remains challenging. Accurate texture rendering and 3D reconstructions are useful for downstream applications like computer aided diagnostics, virtual colonoscopy, and AR surgical training.We combined RNNSLAM1 for pose and depth estimation with 3D Gaussian Splatting2 enhanced by geometric and depth regularizations (Fig. 2). This approach aligns Gaussians with the colon surface (Fig.1a, 1b), improving texture rendering and surface accuracy while reducing floating artifacts. Our contributions include combining geometrically-regularized 3D Gaussian Splatting with RNNSLAM to enhance texture rendering and anatomical accuracy, integrating SLAM with 3D GS for robust 3D reconstructions, and optimizing the training and rendering process. Our method outperforms existing techniques with an 18% improvement in PSNR and a 16% enhancement in SSIM (Fig.4), delivering over 100 times faster rendering speeds and significantly shorter training times (Fig.3). This demonstrates the potential for real-time clinical applications.

Contact: sierra.bonilla.21@ucl.ac.uk

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 17

PROMPT LEARNING FOR CONTINUAL WSI ANALYSIS

Bontempo G., Lumetti L., Porrello A., Bolelli F., Calderara S., Ficarra E.

Abstract: Whole Slide Images (WSIs) are essential in histological diagnosis, offering high-resolution insights into cellular structures. Yet, the gigapixel scale of WSIs and the lack of pixel-level annotations pose substantial challenges. One of these is storing old data only to retrain a model whenever new data is available. Memorizing a vast WSI dataset poses challenges regarding storage capacity and potential privacy leaks if the data can be harmfully spoofed. For this purpose, this work proposes the first multi-resolution rehearsal-free continual learning framework explicitly designed for WSI analysis.

Contact: gianpaolo.bontempo@phd.unipi.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 18

MUSES: THE MULTI-SENSOR SEMANTIC PERCEPTION DATASET

Brodermann T., Bruggemann D., Sakaridis C., Ta K., Liagouris O., Corkill J., Van Gool L.

Abstract: We present the multi-modal MUSES dataset for driving under uncertainty and adverse conditions. MUSES includes 2500 images with diverse weather and illumination. Each image has high-quality 2D pixel-level panoptic and uncertainty annotations. It features calibrated and synchronized recordings from a frame camera, MEMS lidar, FMCW radar, HD event camera, and IMU/GNSS sensor, aiding sensor fusion for dense semantic understanding.

Contact: timbr@vision.ee.ethz.ch

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 19

DO LANGUAGE MODELS UNDERSTAND MORALITY?

Bulla L., Mongiovì M., Gangemi A.

Abstract: The increasing demand for tools able to understand human aspects necessitates the ability to detect and interpret moral values, ethics, and social norms in text. Our research delves into this challenge by extracting moral foundations from natural language documents by analyzing their normative, affective, cultural, and historical components. We propose a hybrid approach that leverages both the strengths of sub-symbolic machine learning and symbolic reasoning with Knowledge Graphs. This combination aims to achieve explainable moral reasoning, using the implicit common-sense knowledge acquired by current machine learning models, particularly Large Language Models (LLMs). Current supervised models often struggle in real-world scenarios due to overfitting specific training data distributions. This leads to performance degradation when testing on data with different distributions. To address this limitation, we propose leveraging state-of-the-art LLMs and Natural Language Inference (NLI) models trained on extensive common-sense data for unsupervised moral classification. Our methodology explores the effectiveness of different LLM sizes and prompt designs, considering both multi-label and binary classification for moral value detection. Moving forward, a critical focus will be on enhancing the explainability of our methods and model outputs. This involves making the underlying mechanisms driving the model's responses more transparent and interpretable for humans. This focus on explainability aligns with the broader goals of ethical AI development.

Contact: luana.bulla@phd.unict.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 20

SMART VIRTUAL RESPIRATORY WARD

Burrows H., Maktabdar Oghaz M., Babu-Saheer L.

Abstract: Lung sounds are an important variable for assessing Asthma and Chronic Obstructive Pulmonary Disease (COPD). Current diagnostic and monitoring methods involve regular appointments, bringing stress to patients and burdening healthcare services. Using Smart stethoscopes and Artificial Intelligence, lung sounds can be interpreted remotely, informing of patient condition at frequent intervals. This research investigates ConvNets and vision transformers for asthma and COPD detection from lung sounds.

Contact: hb643@pgr.aru.ac.uk

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 21

A SECOND-ORDER PERSPECTIVE ON COMPOSITIONALITY AND INCREMENTAL LEARNING

Porrello A., Bonicelli L., Buzzega P., Millunzi M., Calderara S., Cucchiara R.

Abstract: Identifying the conditions that promote compositionality when fine-tuning deep pre-trained models remains an open issue. We conduct a theoretical study leveraging the second-order Taylor approximation of the loss function, leading to two incremental learning algorithms designed to either train models individually or as a whole. Beyond SOTA incremental learning performance, these algorithms demonstrate unlearning and specialization capabilities.

Contact: pietro.buzzega@unimore.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 22

TRENDS, APPLICATIONS AND CHALLENGES IN HUMAN ATTENTION MODELLING

Cartella G., Cuculo V., Cornia M., Cucchiara R.

Abstract: Human attention modelling has proven to be useful for understanding the cognitive processes underlying visual exploration and providing support to AI models to solve problems in various domains, including image and video processing, vision-and-language applications, and language modelling. This work offers an overview of recent efforts to integrate human attention mechanisms into deep learning models and discusses future research directions, including potential applications to cultural heritage.

Contact: giuseppe.cartella@unimore.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 23

SPAMMING LABELS: EFFICIENT ANNOTATIONS FOR THE TRACKERS OF TOMORROW

Cetintas O., Meinhardt T., Brasó G., Leal-Taixé L.

Abstract: Increasing the annotation efficiency of trajectory annotations from videos will enable the next generation of data-hungry tracking algorithms to thrive on large-scale datasets. We introduce SPAM, a tracking data engine that provides high-quality labels with minimal human effort. SPAM uses a graph formulation to annotate detections and identities over time. Trackers trained on SPAM labels reach comparable performance to those trained on human annotations, requiring only 3-20% of the human effort.

Contact: orcun.cetintas@tum.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 24

A LIE GROUP APPROACH TO RIEMANNIAN BATCH NORMALIZATION

Chen Z., Song Y., Liu Y., Sebe, N.

Abstract: Manifold-valued measurements exist in numerous applications within computer vision and machine learning. Recent studies have extended Deep Neural Networks (DNNs) to manifolds, and concomitantly, normalization techniques have also been adapted to several manifolds, referred to as Riemannian normalization. Nonetheless, most of the existing Riemannian normalization methods have been derived in an ad hoc manner and only apply to specific manifolds. This paper establishes a unified framework for Riemannian Batch Normalization (RBN) techniques on Lie groups. Our framework offers the theoretical guarantee of controlling both the Riemannian mean and variance. Empirically, we focus on Symmetric Positive Definite (SPD) manifolds, which possess three distinct types of Lie group structures. Using the deformation concept, we generalize the existing Lie groups on SPD manifolds into three families of parameterized Lie groups. Specific normalization layers induced by these Lie groups are then proposed for SPD neural networks. We demonstrate the effectiveness of our approach through three sets of experiments: radar recognition, human action recognition, and electroencephalography (EEG) classification. The code is available at https://github.com/GitZH-Chen/LieBN.git.

Contact: ziheng_ch@163.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 25

EMOTIONAL SPEECH DRIVEN 3D BODY ANIMATION VIA DISENTANGLED LATENT DIFFUSION

Chhatre K., Danecek R., Athanasiou N., Becherini G., Peters C., Black M., Bolkart T.

Abstract: Existing methods for synthesizing 3D human gestures from speech lack explicit emotion modeling, resulting in animations without emotion control. To address this, we introduce AMUSE, an emotional speech-driven body animation model based on latent diffusion. AMUSE disentangles speech into content, emotion, and personal style, allowing precise control over emotion and style in gesture synthesis. Qualitative and quantitative evaluations demonstrate AMUSE's effectiveness in generating realistic and expressive gestures synchronized with speech content.

Contact: chhatre@kth.se

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 26

UNI: UNLEARNING-BASED NEURAL INTERPRETATIONS

Choi C.L., Duplessis A., Belongie S.

Abstract: We connect machine unlearning and path-based gradient interpretations to propose UNI. UNI computes (un)learnable and debiased baselines by perturbing inputs towards an unlearning direction of steepest ascent. UNI tackles notorious problems: 1. post-hoc attribution biases, 2. suboptimal baselines, 3. high-curvature output manifold.UNI eliminates all external assumptions except for the model's own predictive bias. Unlike static baselines, we avoid injecting texture or frequency assumptions that deviate from model behaviour. UNI computes more consistent path features; is debiased and invulnerable to perturbative attacks. Our findings point to unlearning as a promising avenue for generating faithful, efficient, robust interpretations.

Contact: ccl5a09@gmail.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 27

WEEDS AS ANOMALIES: ANOMALY DETECTION FOR WEED DETECTION IN AGRICULTURAL FIELDS

Chong Y. L., Behley J., and Stachniss C.

Abstract: Automatic weed detection is needed for autonomous weeding robots. Existing methods use semantic segmentation to detect weeds. Instead, we detect weeds as anomalies using anomaly detection. We can then detect weeds using unsupervised methods, and are more robust to the class imbalance (crop vs weed class).

Contact: linn.chong@uni-bonn.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 28

VOLUMETRIC PRIMITIVES FOR MODELING AND RENDERING SCATTERING AND EMISSIVE MEDIA

Condor , Jorge., Speierer, Sebastien., Bode, Lukas., Bozic, Aljaz., Green, Simon., Didyk, Piotr., Jarabo, Adrian.

Abstract: We propose a volumetric representation based on primitives to model scattering and emissive media. Accurate scene representations enabling efficient rendering are essential for many computer graphics applications. General and unified representations that can handle surface and volume-based representations simultaneously, allowing for physically accurate modeling, remain a research challenge. Inspired by recent methods for scene reconstruction that leverage mixtures of 3D Gaussians to model radiance fields, we formalize and generalize the modeling of scattering and emissive media using mixtures of simple kernel-based volumetric primitives. We introduce closed-form solutions for transmittance and free-flight distance sampling for 3D Gaussian kernels, and propose several optimizations to use our method efficiently within any off-the-shelf volumetric path tracer by leveraging ray tracing for efficiently query the medium. We demonstrate our method as an alternative to other forms of volume modeling (e.g. voxel grid-based representations) for forward and inverse rendering of scattering media. Furthermore, we adapt our method to the problem of radiance field optimization and rendering, and demonstrate comparable performance to the state of the art, while providing additional flexibility in terms of performance and usability.

Contact: jorge.condor@usi.ch

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 29

TRANSFERRING DISENTANGLED REPRESENTATIONS

Dapueto J.

Abstract: Developing meaningful and efficient representations that separate the fundamental structure of the data generation mechanism is crucial in representation learning. However, Disentangled Representation Learning (DRL), which aims to identify and disentangle underlying Factors of Variation (FoVs), is hardly applicable to real images. We investigate the possibility of leveraging synthetic data to learn general-purpose disentangled representations which we then transfer to real data.

Contact: jacopo.dapueto@edu.unige.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 30

VISION TRANSFORMERS NEED REGISTERS

Darcet T., Oquab M., Mairal J., Bojanowski P.

Abstract: In this work we: - Characterize artifacts in ViT feature / attention maps - Propose a natural fix by adding new tokens (“registers”) - Validate that we fix the attention maps of supervised, text-supervised and self-supervised models

Contact: timothee.darcet@gmail.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 31

SEMANTICALLY GUIDED REPRESENTATION LEARNING FOR ACTION ANTICIPATION

Diko A., Avola D., Prenkaj B., Fontana F., Cinque L.

Abstract: Predicting future events from incomplete observations is challenging due to inherent uncertainty and the complex relations between actions. We propose S-GEAR, a framework focusing on learning action representations that understand inter-action semanticity based on prototypical action patterns and contextual co-occurrences. S-GEAR learns visual action prototypes and uses language models to structure their geometric relationships, inducing semantic interconnectivity as inferred from language.

Contact: diko@di.uniroma1.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 32

MEMFLOW: OPTICAL FLOW ESTIMATION AND PREDICTION WITH MEMORY

Dong Q., Fu Y.

Abstract: Classical optical flow estimation uses two frames as input, whilst some recent methods consider multiple frames to explicitly model long-range information. The former ones limit their ability to fully leverage temporal coherence along the video sequence; and the latter ones incur heavy computational overhead, typically not possible for real-time flow estimation. Some multi-frame-based approaches even necessitate unseen future frames for current estimation, compromising real-time applicability in safety critical scenarios. To this end, we present MemFlow, a realtime method for optical flow estimation and prediction with memory. Our method enables memory read-out and update modules for aggregating historical motion information in real-time. Furthermore, we integrate resolution-adaptive re-scaling to accommodate diverse video resolutions. Besides, our approach seamlessly extends to the future prediction of optical flow based on past observations. Codes and models are available at: https://dqiaole.github.io/MemFlow/ .

Contact: qldong18@fudan.edu.cn

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 33

DON'T DROP YOUR SAMPLES! COHERENCE-AWARE TRAINING BENEFITS CONDITIONAL DIFFUSIONNAL

Nicolas D., Victor B., Vicky K., David P.

Abstract: Conditional diffusion models are powerful generative models that can leverage various types of conditional information. However, in many real-world scenarios, conditional information may be noisy or unreliable due to human annotation errors or weak alignment. In this paper, we propose the Coherence-Aware Diffusion (CAD), a novel method that integrates coherence in conditional information into diffusion models, allowing them to learn from noisy annotations without discarding data.

Contact: nicolas.dufourn@gmail.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 34

SHAPE ESTIMATION OF SEMI- FLUID DEFORMABLE OBJECTS

EL ASSAL O., M. MATTEO C., CIRON S., FOFI D.

Abstract: In this poster we treat the problematic of estimating the shape of fluid deformable objects. We adopt a framework that automatically generates the ground truth labels used to train a generative network that estimates the shape of the weld pool. In welding application, understanding the shape of the weld pool is crucial to provide a clear understanding of the process, for later purposes such as process control and analysis.

Contact: omar.el-assal@alstomgroup.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 35

INVISIBLE STITCH: GENERATING SMOOTH 3D SCENES WITH DEPTH INPAINTING

Engstler P., Vedaldi A., Laina I., Rupprecht C.

Abstract: In the well-established field of 3D scene generation, most prior work generates scenes by iteratively stitching newly generated frames with existing geometry. They depend on monocular depth estimators to lift the generated images into 3D, ignoring the geometry of the existing scene. We thus introduce a novel depth completion model, which achieves greater geometric consistency according to our proposed benchmarking scheme for scene generation methods, fully rooted in ground truth geometry.

Contact: paule@robots.ox.ac.uk

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 36

CONTROLLING STYLE IN DIFFUSION MODELS THROUGH NOISE

Everaert M.N., Süsstrunk S., Achanta R.

Abstract: We observe that the style of images generated by Stable Diffusion is tied to the initial noise. Thus, we propose a method to adapt Stable Diffusion to various styles using style-specific noise during fine-tuning (ICCV23). We subsequently explain that white noise added during training preserves low-frequency (LF) content, and the model then learns to maintain the LF of the initial noise. Controlling this initial noise allows to generate images with desired styles without finetuning (WACV24).

Contact: martin.everaert@epfl.ch

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 37

EXPLORING THE SYNERGY BETWEEN VISION-LANGUAGE PRETRAINING AND CHATGPT FOR ARTWORK CAPTIONING

Castellano G., Fanelli N., Scaringi R., Vessio G.

Abstract: Generating accurate and informative captions for artworks is challenging due to the need to understand artistic intent, historical context, and complex visual elements. This study introduces a new dataset for artwork captioning generated using prompt engineering techniques and ChatGPT, and refined with CLIPScore to filter out noise. By fine-tuning GIT-Base and employing multi-task learning to predict artwork metadata, the proposed approach generates visually accurate and informative captions that surpass the ground truth.

Contact: nicola.fanelli@uniba.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 38

OPENMASK3D: OPEN-VOCABULARY 3D INSTANCE SEGMENTATION

Takmaz A., Fedele E., Sumner B., Pollefeys M., Tombari F., Engelmann F.

Abstract: We propose OpenMask3D, the first method for open-vocabulary 3D instance segmentation. Guided by predicted class-agnostic 3D instance masks, our model aggregates per-mask features via multi-view fusion of CLIP-based image embeddings. These task-agnostic embeddings enable our method to identify object instances based on novel, open-vocabulary queries describing object properties such as semantics, geometry, affordances, and materials.

Contact: efedele@ethz.ch

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 39

HOW TO PROBE: SIMPLE YET EFFECTIVE TECHNIQUES FOR IMPROVING POST-HOC EXPLANATIONS

Gairola S., Böhle M., Locatello F., Schiele B.

Abstract: Post-hoc attribution methods explain DNNs without considering training specifics. However, recent studies show that training paradigms significantly affect explanation quality. With diverse pre-training techniques emerging, understanding their impact on explanation methods is crucial. We systematically study various visual pre-training frameworks and find that small changes in the classification layer of models can greatly enhance explanation quality, more so than the pre-training scheme itself.

Contact: sgairola@mpi-inf.mpg.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 40

GRAPHDREAMER: COMPOSITIONAL 3D SCENE SYNTHESIS FROM SCENE GRAPHS

Gao G., Liu W., Chen A., Geiger A., Schölkopf B.

Abstract: TL;DR: GraphDreamer grounds 3D generation in scene graphs, which are semantically more accurate, and generates compositional 3D scenes by jointly supervising 3D objects and their mutual relationships via a 2D diffusion model.

Contact: gege.gao@inf.ethz.ch

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 41

GEN2BALANCE USING GENERATIVE MODELS FOR LONG-TAIL VIDEO ACTION RECOGNITION

Gatti P., Jenni S., Caba Heilbron F., Collomosse J., Damen D.

Abstract: With recent advancements in image and video generation, we explore whether synthetically generated data can improve long-tail video action recognition. We propose two long-tail versions of action recognition datasets, UCF-LT and Kinetics 400-LT, and employ ControlNet to generate data for balancing the long tail. Using VideoMAE as a baseline, our results on UCF-LT demonstrate improved overall performance, as well as on the few-shot and tail classes.

Contact: prajwal.gatti@bristol.ac.uk

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 42

MPL: LIFTING 3D POSE FROM MULTI-VIEW 2D POSES

Ghasemzadeh S.A., De Vleeschouwer C.

Abstract: Estimating 3D poses from 2D images is challenging due to occlusions. While multi-view setups improve accuracy, they often fail to generalize to real-world cases due to reliance on multi-view images paired with 3D poses, which are scarce and expensive to obtain. Thus, we propose decoupling 2D pose estimation and 3D pose lifting using a transformer-based network called Multi-view 3D Pose Lifter. This approach uses available 2D datasets and synthetic 2D-3D pose pairs, enhancing the generalization.

Contact: seyed.ghasemzadeh@uclouvain.be

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 43

THE SHORT FILM DATASET (SFD): A BENCHMARK FOR STORY-LEVEL VIDEO UNDERSTANDING

GHERMI R., Wang X., Kalogeiton V., Laptev I.

Abstract: We propose the Short Film Dataset (SFD), a novel benchmark for long-form video understanding on amateur movies. While most datasets are confined to short videos with limited events and narrow narratives, we focus on long-term story-oriented video tasks in the form of multiple-choice and open-ended question answering, with videos lasting 13 minutes on average. The dataset comprises 1,078 publicly available short films, 243 hours of content, and 4,885 questions. Additionally, SFD offer minimum data contamination issues, as short movies are unknown to recent LLMs.

Contact: ridouaneg@gmail.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 44

LARGE-SCALE GLOBAL LOCALIZATION USING TRANSFORMERS

Gladkova M.

Abstract: Place recognition is an essential component of an autonomous system. While GPS can provide fairly accurate global location, it operates under favourable outdoor conditions with direct signal to a satellite, thus failing in spaces such as tunnels and indoor environments. In this poster, we present two works, CASSPR [1] and VXP [2], that aim to utilize the power of transformer architectures for achieving SOTA performances on LiDAR single-scan and image-LiDAR cross-modal localization. We advocate shared latent spaces in challenging conditions of extremely sparse scans and night-day settings.

Contact: mariia.gladkova@tum.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 45

BAYES' RAYS: UNCERTAINTY QUANTIFICATION FOR NEURAL RADIANCE FIELDS

Goli L., Reading C., Sellán S., Jacobson A., Tagliasacchi A.

Abstract: What is the problem? 3D Reconstruction from a few 2D images has inherent geometric ambiguity causing uncertainty. Why does it matter? Detecting uncertainty helps with tasks like Outlier (floater) Removal and Next Best View planning for scene capture. What did we do? Get spatial uncertainty of any pre-trained NeRF in a minute and remove artifacts in real-time!

Contact: lily.goli@mail.utoronto.ca

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 46

LABEL EFFICIENT LEARNING PARADIGMS FOR SCALABLE ROBOT PERCEPTION

Gosala N., Petek K., Besic B., Cattaneo D., Drews-Jr P., Kiran BR., Yogamani S., Burgard W., Valada A.

Abstract: Existing scene understanding approaches are fully-supervised and label intensive which hinders model scalability. My research circumvents the scalability bottleneck using novel label-efficient learning paradigms. I propose geometry-aware designs to simplify model training, exploit spatio-temporal consistency for geometry reasoning, and decompose models for targeted optimization. Resulting frameworks can thus leverage their simple, yet rich, designs to perform complex tasks with limited data.

Contact: gosalan@cs.uni-freiburg.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 47

CALVING FRONTS AND WHERE TO FIND THEM – A BENCHMARK FRAMEWORK

Gourmelon N., Seehaus T., Braun M., Maier A., Christlein V.

Abstract: To improve climate predictions, climate models need to be calibrated using the position of marine-terminating glaciers‘ calving fronts – the edge between ocean and glacier. Monitoring of this position can be automated by deep learning models. To enable the development of such models, we present the „Calving fronts and where to find them“ (CaFFe) benchmark for glacier calving front extraction from synthetic aperture radar (SAR) imagery [1]. Model performances of published studies are provided.

Contact: nora.gourmelon@fau.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 48

IMPARA: INDUSTRIAL MULTI-TASK PHOTO-REALISTIC ANNOTATED DATASET FOR ROBOTICS APPLICATIONS

Govi E., Sapienza D., De Dominicis L., Capodieci N., Bertogna M.

Abstract: We introduce IMPARA, (Industrial Multi-task Photo-realistic Annotated dataset for Robotics Applications), a new simulated, highly realistic dataset specifically designed for industrial use cases. The aim of this dataset is to address some of the challenges persisting in the field of robotic industrial automation. Its novelty lies in several aspects: the generated frames captures unique material properties of the objects, including cluttered scenes accurately accounting for rigid-body physics.

Contact: elena.govi@unimore.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 49

BOOSTING UNSUPERVISED SEMANTIC SEGMENTATION WITH PRINCIPAL MASK PROPOSALS

Hahn O., Araslanov N., Schaub-Meyer S., Roth S.

Abstract: Unsupervised semantic segmentation aims to automatically partition images into semantically meaningful regions by identifying global categories within an image corpus without any form of annotation. Building upon recent advances in self-supervised representation learning, we focus on how to leverage these large pre-trained models for the downstream task of unsupervised segmentation. We present PriMaPs - Principal Mask Proposals - decomposing images into semantically meaningful masks based on their feature representation. This allows us to realize unsupervised semantic segmentation by fitting class prototypes to PriMaPs with a stochastic expectation-maximization algorithm, PriMaPs-EM. Despite its conceptual simplicity, PriMaPs-EM leads to competitive results across various pre-trained backbone models, including DINO and DINOv2, and across datasets, such as Cityscapes, COCO-Stuff, and Potsdam-3. Importantly, PriMaPs-EM is able to boost results when applied orthogonally to current state-of-the-art unsupervised semantic segmentation pipelines.

Contact: oliver.hahn@visinf.tu-darmstadt.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 50

DIFFCD: A SYMMETRIC DIFFERENTIABLE CHAMFER DISTANCE FOR NEURAL IMPLICIT SURFACE FITTING

Härenstam-Nielsen L., Sang L., Saroha A., Araslanov N., Cremers D.

Abstract: The predominant approach for fitting neural SDFs to point clouds corresponds to minimizing the one-sided Chamfer Distance between the point cloud and the implicit surface. Due to the non-symmetric nature of the loss, such methods are susceptible to spurious surfaces. We analyze the SIREN approach to dealing with spurious surfaces and reveal that it corresponds to regularizing surface area, leading to over-smoothing. We instead propose to minimize the full symmetric Chamfer distance.

Contact: Linus.nielsen@tum.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 51

ALL U-NEED IS U-NET(S)

Helander R., Parvizi J.

Abstract: Radiotherapy treatment planning is the process of designing treatments that maximize tumor control while minimizing damage to healthy tissues, see Fig. 2. A recent way to speed up this tedious process is to train an ML model to predict the dose distribution within a patient [1]. This work evaluates a cascaded architecture for the dose prediction task on breast cancer patients and demonstrates that it is superior to a non-cascaded architecture, yielding a 7.8% mean test set loss reduction.

Contact: rashel@raysearchlabs.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 52

SOAC: SPATIO-TEMPORAL OVERLAP-AWARE MULTI-SENSOR CALIBRATION USING NEURAL RADIANCE FIELDS

Herau Q., Piasco N., Bennehar M., Roldão L., Tsishkou D., Migniot C., Vasseur P., Demonceaux C.

Abstract: In autonomous driving, using multiple sensors with different modalities is crucial for precision and stability. Accurate calibration of these sensors is essential to unify their data. This paper leverages implicit scene representation to represent various sensor modalities in a shared volumetric space. By partitioning the scene based on visible parts for each sensor and focusing on overlapping areas, we enhance calibration robustness and accuracy.

Contact: quentin.herau@gmail.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 53

EVALUATING XAI METHODS

Hesse R., Schaub-Meyer S., Roth S.

Abstract: Evaluating explainable AI methods for computer vision is challenging due to the lack of ground-truth explanations. To address this, we first introduce a synthetic dataset that enables image-space interventions to measure changes in the model output, thereby providing ”ground-truth” explanations. In our second approach, we extend this concept to real-world datasets by training and evaluating models on images with deleted patches.

Contact: robin.hesse@visinf.tu-darmstadt.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 54

MULTI-MODAL GAZE FOLLOWING IN CONVERSATIONAL SCENARIOS

Hou Y., Zhang Z., Horanyi N., Cheng Y., Chang H.J., University of Birmingham
Moon, J., KETI

Abstract: Gaze following estimates gaze targets of in-scene person by understanding human behavior and scene information. Existing methods usually analyze scene images for gaze following. However, compared with visual images, audio also provides crucial cues for determining human behavior. This suggests that we can further improve gaze following considering audio cues. In this paper, we explore gaze following tasks in conversational scenarios. We propose a novel multi-modal gaze following framework based on our observation" audiences tend to focus on the speaker". We first leverage the correlation between audio and lips, and classify speakers and listeners in a scene. We then use the identity information to enhance scene images and propose a gaze candidate estimation network. The network estimates gaze candidates from enhanced scene images and we use MLP to match subjects with candidates as classification tasks. Existing gaze following datasets focus on visual images while ignore audios. To evaluate our method, we collect a conversational dataset, VideoGazeSpeech (VGS), which is the first gaze following dataset including images and audio. Our method significantly outperforms existing methods in VGS datasets. The visualization result also prove the advantage of audio cues in gaze following tasks. Our work will inspire more researches in multi-modal gaze following estimation.

Contact: yxh029@student.bham.ac.uk

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 55

RECONSTRUCTING MULTI-BODY DYNAMICS FOR COMPOSITIONAL RE-SIMULATION

HUANG S.,

Abstract: We model multi-body dynamics for authentic re-simulation. Starting with PREDATOR for pair-wise registration of static scenes, PCAccumulation and LivingScenes extend to dynamic scenarios, with LivingScenes integrating surface reconstruction objectives for registration. For high-fidelity re-simulation, NFL offers a sensor model and implicit scene representation. DyNFL advances this to dynamic scenes with compositional neural fields, ensuring detailed and flexible re-simulation.

Contact: shenhuan@ethz.ch

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 56

4M-21: AN ANY-TO-ANY VISION MODEL FOR TENS OF TASKS AND MODALITIES

Kar O.F., Bachmann R., Mizrahi D., Garjani A., Gao M., Griffiths D., Hu J., Dehghan A., Zamir A.

Abstract: We perceive the world through different modalities. Each provides a distinct view of the same physical reality and when combined, they allow us to better understand our world. In this work, we train an any-to-any vision foundation model by scaling the model to 3B parameters trained on 1T tokens from 21 distinct modalities via masked pre-training. The resulting model, 4M-21, has strong capabilities such as zero-shot solving several vision tasks, controllable multi-modal generation, retrieval, fusion, and transferability to downstream tasks. Code and models are released.

Contact: oguzhan.kar@epfl.ch

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 57

PROTECTING DATA PRIVACY OF DIFFUSION MODELS

Katsumata K., Nakayama H.

Abstract: This study examines the vulnerability of diffusion models to data leakage, focusing on the impact of hard negative sampling. Diffusion models, which transform noise into images, are susceptible to Membership Inference Attacks and data extraction. By filtering out low-loss samples during training, we aim to reduce this risk. Experiments show decreased vulnerability with visual quality loss. Future work includes applying these methods to text-to-image models and enhancing privacy during training.

Contact: raven@sfc.wide.ad.jp

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 58

METTA: SINGLE-VIEW TO 3D TEXTURED MESH RECONSTRUCTION WITH TEST-TIME ADAPTATION

Yu-Ji K., Ha H., Youwang K., Surh J., Ha H., Oh T.-H.

Abstract: One popular approach to reconstructing 3D from a single-view image is learning-based methods, which suffer challenges when test cases are unfamiliar with training data (Out-of-distribution; OoD). To adapt for unseen samples in test time, we propose MeTTA, a test-time adaptation (TTA) exploiting generative prior. We design joint optimization of 3D geometry, appearance, and pose to handle OoD cases. MeTTA achieves high-fidelity geometry and realistic physically based rendering (PBR) textures.

Contact: ugkim@postech.ac.kr

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 59

OPEN-VOCABULARY 3D SCENE GRAPHS

Koch S, Vaskevicius N, Colosi M, Hermosilla P, Ropinski T

Abstract: In 3D scene understanding, 3D scene graphs provide a structured and semantically meaningful representation by modeling scene entities and their relationships, useful for tasks in AR/VR and robotics. However current approaches are limited by closed vocabularies during training, reducing expressiveness. Open3DSG is the first method to predict explicit open-vocabulary object classes and open-set relationships enabling the prediction of rare and specific objects and relationships in 3D scene graphs.

Contact: kochsebastian98@gmail.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 60

HOLONETS: SPECTRAL CONVOLUTIONS DO EXTEND TO DIRECTED GRAPHS

Kokce C., Cremers D.

Abstract: Within the graph learning community conventional wisdom dictates that spectral convolutional networks may only be deployed on undirected graphs: Only there is the existence of a suitable graph Fourier transform guaranteed, so that information may be translated between spatial- and spectral domains. Using ideas from complex analysis, we here show the traditional reliance on the graph Fourier transform to be superfluous and extend spectral convolutions to directed graphs.

Contact: christian.koke@tum.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 61

LEARNING GEOMETRY: ROTATION REPRESENTATION OF SYMMETRIC OBJECTS

Kriegler A., Beleznai C., Gelautz M.

Abstract: We propose a representation usable for 6D pose estimation of symmetric objects. In existing works, this problem is usually circumvented, via 1.) evaluating with a metric that is symmetry-insensitive (e.g. VSSD), or 2.) modifying the architecture. In contrast, our representation is both unique and continuous for given symmetries, enabling symmetry-sensitive evaluation for basic architectures. We evaluate our approach on T-LESS.

Contact: andreas.kriegler@tuwien.ac.at

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 62

DFLOW: LEARNING TO SYNTHESIZE BETTER OPTICAL FLOW DATASETS VIA A DIFFERENTIABLE PIPELINE

Byung-Ki K., Hyeon-Woo N., Kim J.-Y., Oh T.-H.

Abstract: Studies of synthetic optical flow datasets aim to identify properties that improve learning-based optical flow estimation. Manual identification of these properties requires impractical large-scale trial-and-error experiments. To address this, we propose DFlow, a differentiable optical flow data generation pipeline with a loss function. DFlow efficiently synthesizes datasets for target domains without cumbersome trials. It uses neural networks to approximate and compare datasets instead of explicit pairwise comparison.

Contact: qudrlskfk@postech.ac.kr

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 63

GALLOP: LEARNING GLOBAL AND LOCAL PROMPTS FOR VISION-LANGUAGE MODELS

Lafon M., Ramzi E., Rambour C., Audebert N., Thome N.

Abstract: Recent prompt learning methods for few-shots adaptation [1, 2, 3] of vision-language models (VLMs) often trade accuracy for robustness (e.g. domain generalization or OOD detection). We introduce Global-Local Prompts (GalLoP), which learns diverse prompts from global and local features using an enhanced vision-text local alignment and a sparsity strategy to focus only on relevant local features. GalLoP surpasses previous prompt learning methods for VLMs in accuracy and robustness.

Contact: marc.lafon@lecnam.net

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 64

FASTER THAN LIES: REAL-TIME DEEPFAKE DETECTION USING BINARY NEURAL NETWORKS

Lanzino R., Fontana F., Diko A., Marini M. R., Cinque L.

Abstract: Deepfake detection counters the spread of deep-generated media that undermines online trust. We introduce a novel approach using Binary Neural Networks (BNNs) [1] for fast inference with minimal accuracy loss, incorporating FFT and LBP to detect manipulation. Evaluations on COCOFake [2], DFFD [3], and CIFAKE [4] datasets show state-of-the-art performance with up to a 20x reduction in FLOPs. This work paves the way for efficient deepfake detection research.

Contact: romeo.lanzino@uniroma1.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 65

GEOMETRY-AWARE PROJECTIVE MAPPING FOR UNBOUNDED NEURAL RADIANCE FIELDS

Lee J., Jung H., Park J., Bae I., Jeon H.

Abstract: Neural radiance fields (NeRFs) uses certain mapping functions to represent for an unbounded scene where cameras point in any direction and contents exist at any distance, yet they either work in object-centric scenes or focus on objects close to the camera. This work designs an adaptive mapping function and its ray parameterization for unbounded scenes in neural radiance fields (NeRFs) using projection functions.

Contact: juno@gm.gist.ac.kr

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 66

ARE SYNTHETIC DATA USEFUL FOR EGOCENTRIC HAND-OBJECT INTERACTION DETECTION?

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

Abstract: We investigate the effectiveness of synthetic data in enhancing egocentric hand-object interaction detection. Via extensive experiments and comparative analyses on three egocentric datasets, EPIC-KITCHENS VISOR, EgoHOS, and ENIGMA-51, our findings reveal how to exploit synthetic data for the HOI detection task when real labeled data are scarce or unavailable. Specifically, by leveraging only 10% of real labeled data, we achieve improvements in Overall AP compared to baselines trained exclusively on real data of: +5.67% on EPIC-KITCHENS VISOR, +8.24% on EgoHOS, and +11.69% on ENIGMA-51. Our analysis is supported by a novel data generation pipeline and the newly introduced HOI-Synth benchmark which augments existing datasets with synthetic images of hand-object interactions automatically labeled with hand-object contact states, bounding boxes, and pixel-wise segmentation masks.

Contact: rosario.leonardi@phd.unict.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 67

MAPPING EYE-TRACKING DATA ACROSS INDEPENDENT EGOCENTRIC VIDEO STREAMS

LEON-CONTRERAS N., DIERKES K.

Abstract: Head-mounted eye trackers such as Pupil Labs' Neon report gaze direction in the pixel space of an integrated egocentric scene camera. For use-cases requiring higher image quality and faster frames rates, we present a calibration-free pipeline for mapping gaze directions into an external egocentric action camera. Our method leverages optic flow and DL-based feature matching for the estimation of local mapping functions. We explore stereo calibration for adding epipolar constraints.

Contact: sof@pupil-labs.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 68

EGOGEN: AN EGOCENTRIC SYNTHETIC DATA GENERATOR

Li G., Zhao K., Zhang S., Lyu X., Dusmanu M., Zhang Y., Pollefeys M., Tang S.

Abstract: Understanding the world in first-person view is fundamental in Augmented Reality (AR). This immersive perspective brings dramatic visual changes and unique challenges compared to third-person views. Synthetic data has empowered third-person-view vision models, but its application to embodied egocentric perception tasks remains largely unexplored. A critical challenge lies in simulating natural human movements and behaviors that effectively steer the embodied cameras to capture a faithful egocentric representation of the 3D world. To address this challenge, we introduce EgoGen, a new synthetic data generator that can produce accurate and rich ground-truth training data for egocentric perception tasks. At the heart of EgoGen is a novel human motion synthesis model that directly leverages egocentric visual inputs of a virtual human to sense the 3D environment. Combined with collision-avoiding motion primitives and a two-stage reinforcement learning approach, our motion synthesis model offers a closed-loop solution where the embodied perception and movement of the virtual human are seamlessly coupled. Compared to previous works, our model eliminates the need for a pre-defined global path, and is directly applicable to dynamic environments. Combined with our easy-to-use and scalable data generation pipeline, we demonstrate EgoGen's efficacy in three tasks: mapping and localization for head-mounted cameras, egocentric camera tracking, and human mesh recovery from egocentric views. EgoGen will be fully open-sourced, offering a practical solution for creating realistic egocentric training data and aiming to serve as a useful tool for egocentric computer vision research.

Contact: gen.li@inf.ethz.ch

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 69

MULTI-IMAGE VISUAL QUESTION ANSWERING FOR UNSUPERVISED ANOMALY DETECTION

Li J., Bercea C.I., Müller P., Felsner L., Kim S.,Rueckert D., Wiestler B., Schnabel J.A.

Abstract: Unsupervised anomaly detection identifies pathological areas by comparing images with pseudo-healthy reconstructions, but interpreting anomaly maps is challenging due to a lack of explanations. We propose, to the best of our knowledge, the first framework using language models for unsupervised anomaly detection with a Visual Question Answering (VQA) dataset. Our multi-image VQA framework, enhanced by a Knowledge Q-Former module, effectively answers questions and improves pathology detection using anomaly maps.

Contact: june.li@tum.de

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 70

EXPANSION AND SHRINKAGE OF LOCALIZATION FOR WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

Li J., Jie Z., Wang X., Wei X., Ma L.

Abstract: 1. CAM generates incomplete and inaccurate pseudo labels. 2. Existing methods rely on regular DCNNs, but (1) challenging to obtain high-quality pseudo labels. (2) time-consuming due to iterative CAM maps fusion. 3. Loss Minimization leads to high-accuracy classification but (1) lead to local most-discriminative activations. (2) include many noisy backgrounds. (3) results in incomplete and in accurate localization maps.

Contact: jinlong.li@unitn.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 71

EVCAP: RETRIEVAL-AUGMENTED IMAGE CAPTIONING WITH EXTERNAL VISUAL-NAME MEMORY FOR OPEN-WORLD COMPREHENSION

Li J., Vo D., Sugimoto A., Nakayama H.

Abstract: Large language models (LLMs)-based image captioning has the capability of describing objects not explicitly observed in training data, but frequently occurring novel objects require up-to-date knowledge for open-world comprehension. We introduce a highly effective retrieval-augmented image captioning method that prompts LLMs with object names retrieved from External Visual–name memory (EVCap). We build ever-changing object knowledge memory, enabling us to (i) update the memory at a minimal cost and (ii) effortlessly augment LLMs with retrieved object names by utilizing a lightweight and fast-to-train model. EVCap outperforms other methods on benchmarks and synthetic commonsense-violating data.

Contact: li@nlab.ci.i.u-tokyo.ac.jp

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 72

VSTAR: GENERATIVE TEMPORAL NURSING FOR LONGER DYNAMIC VIDEO SYNTHESIS

Li Y., Beluch W., Keuper M., Zhang D., Khoreva A.

Abstract: VSTAR enables pretrained text-to-video models for longer video generation with dynamic visual evolution in a single inference pass. Through a controlled analysis, we identify that temporal attention (TA) manipulation can effectively influence the video dynamics, which leads us to a simply yet effective TA regularization technique. We further visualize TA across different T2V models, and provide valuable insights for model training with enhanced video dynamics.

Contact: yumeng.li@de.bosch.com

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 73

OBJECT TRACKING WITH EVENTS AND FRAMES

Li Z., Piga N., Pietro F., Iacono M., Glover A., Natale L., Bartolozzi C.

Abstract: Robust object pose tracking plays an important role in robot manipulation, but it is still an open issue for quickly moving targets as motion blur and low frequency detection can reduce pose estimation accuracy even for STOA RGB-D-based methods. An event-camera is a low-latency vision sensor that can act complementary to RGB-D. Specifically, its sub-millisecond temporal resolution can be exploited to correct for pose estimation inaccuracies due to low frequency RGB-D based detection. We propose a dual Kalman filter: the first filter estimates an object's velocity from the spatio-temporal patterns of "events", the second filter fuses the tracked object velocity with a low-frequency object pose estimated from a deep neural network using RGB-D data. The full system outputs high frequency, accurate object poses also for fast moving objects. The proposed method works towards low-power robotics by replacing high-cost GPU-based optical flow used in prior work with event-cameras that inherently extract the required signal without costly processing.

Contact: zhichao.li@iit.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 74

GAUSSIAN-SLAM: PHOTO-REALISTIC DENSE SLAM WITH GAUSSIAN SPLATTING

Yugay V., Li Y., Gevers T., Oswald M.R.

Abstract: We present a dense SLAM method that uses 3D Gaussians as the scene representation, enabling high-quality rendering and large-scale reconstruction from real-world inputs. We propose a novel strategy for seeding Gaussians of newly explored areas and their effective online optimization that is independent of the scene size, thus scalable to larger scenes.

Contact: y.li6@uva.nl

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 75

UNCERTAINTY-GUIDED OPEN-SET SOURCE-FREE UNSUPERVISED DOMAIN ADAPTATION WITH TARGET-PRIVATE CLASS SEGREGATION

Litrico M., Talon D., Battiato S., Del Bue A., Giuffrida M. V., Morerio P.

Abstract: Standard Unsupervised Domain Adaptation (UDA) aims to transfer knowledge from a labeled source domain to an unlabeled target but usually requires simultaneous access to both source and target data. Moreover, UDA approaches commonly assume that source and target domains share the same labels space. Yet, these two assumptions are hardly satisfied in real-world scenarios. This paper considers the more challenging Source-Free Open-set Domain Adaptation (SF-OSDA) setting, where both assumptions are dropped. We propose a novel approach for SF-OSDA that exploits the granularity of target-private categories by segregating their samples into multiple unknown classes. Starting from an initial clustering-based assignment, our method progressively improves the segregation of target-private samples by refining their pseudo-labels with the guide of an uncertainty-based sample selection module. Additionally, we propose a novel contrastive loss, named NL-InfoNCELoss, that, integrating negative learning into self-supervised contrastive learning, enhances the model robustness to noisy pseudo-labels. Extensive experiments on benchmark datasets demonstrate the superiority of the proposed method over existing approaches, establishing new state-of-the-art performance. Notably, additional analyses show that our method is able to learn the underlying semantics of novel classes, opening the possibility to perform novel class discovery.

Contact: mattia.litrico@phd.unict.it

Presentation Type: Poster

Date: Monday 8 July 2024

Time: 21:30

Poster Session: 1

Poster Number: 76

DEMOCRATIZING FINE-GRAINED VISUAL RECOGNITION WITH LARGE LANGUAGE MODELS

Liu M. , Roy S. , Li W. , Zhong Z. , Sebe N., Ricci E.

Abstract: Identifying subordinate-level categories from images is a longstanding task in computer vision and is referred to as fine-grained visual recognition (FGVR). We propose Fine-grained Semantic Category Reasoning (FineR) system that internally leverages the world knowledge of large language models (LLMs) as a proxy in order to reason about fine-grained category names, showing promise in working in the wild and in new domains where gathering expert annotation is arduous.

Contact: mingxuan.liu@unitn.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 77

LONG-TERM ACTION PREDICTION IN PROCEDURAL TASKS OF LONG DURATION

Loureiro M.

Abstract: Long-term action prediction involves predicting sequences of future actions based on visual features from past events. Current transformer-based approaches are inefficient for long sequences, limiting their application to short video segments, typically only a few minutes long. We propose a model that addresses this limitation by incorporating a compression module, which stores relevant information in a compressed form to effectively leverage information from longer past sequences.

Contact: malourei@pa.uc3m.es

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 78

DN-CBM: TASK AGNOSTIC CONCEPT BOTTLENECKS

Rao S.*, Mahajan S.*, Böhle M., Schiele B.

Abstract: Concept Bottleneck Models (CBMs) induce interpretability into deep networks by first mapping the image to human understandable concepts, and then linearly combining them for classification. However, current CBMs are task-specific, necessitating querying LLMs for concepts and training CBMs separately. We propose DN-CBM, an intuitive and alternative paradigm to first discover concepts learnt by the model, and then name them. We use this to build task-agnostic and performant CBMs in an automated manner, which is surprisingly effective at discovering meaningful concepts.

Contact: swmahaja@mpi-inf.mpg.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 79

DEEP LEARNING-BASED BRAIN SHAPE MORPHOLOGY ANALYSIS AND CORRELATES BETWEEN BRAIN AND FACIAL BIOMARKERS FOR THE DIAGNOSIS OF PSYCHOTIC DISORDERS

Malé J., Martínez-Abadías N., Sevillano X.

Abstract: Diagnosis and prognosis of genetic and developmental disorders require complex, time-consuming analysis, traditionally involving expert neurologists and radiologists manually annotating pre-defined landmarks. Recent advancements focus on automating reconstruction, segmentation, registration, and landmark detection, significantly reducing manual effort. Innovations in automatic biomarker definition are enhancing accuracy and efficiency in identifying and managing these disorders.

Contact: jordi.male@salle.url.edu

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 80

DUAL-AXIS TILTING ROTOR QUAD-PLANE DESIGNED FOR PRECISION LANDING IN GUSTY MARITIME ENVIRONMENTS

Mancinelli A., Smeur E., Remes B., De Croon G.

Abstract: The goal of the work is to develop a Vertical Takeoff and Landing hybrid Unmanned Aerial Vehicle (UAV) capable of autonomously taking off from and landing on a moving ship. To accomplish this task, a novel quad-plane design was conceived. Equipped with four dual-axis tilting rotors, the vehicle achieves full 6 Degrees of Freedom (DOF) control in the hovering configuration. This propulsion solution also enhances wind resistance capability by directly countering linear acceleration disturbances.

Contact: a.mancinelli@tudelft.nl

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 81

IS MULTIPLE OBJECT TRACKING A MATTER OF SPECIALIZATION?

Mancusi G., Bernardi M., Panariello A., Porrello A., Calderara S., Cucchiara R.

Abstract: End-to-end transformer-based trackers excel on human datasets but struggle in diverse scenarios due to poor domain generalization, requiring costly fine-tuning. We introduce PASTA, which uses Parameter-Efficient Fine-Tuning (PEFT) and Modular Deep Learning (MDL). By training specialized PEFT modules for key attributes, our framework enhances generalization. Experiments show specialized trackers outperform a monolithic one, raising the question: ”Is MOT about specialization or generalization?”

Contact: gianluca.mancusi@unimore.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 82

ZERO SHOT NOISE2NOISE: EFFICIENT IMAGE DENOISING WITHOUT ANY DATA

Mansour Y., Heckel R.

Abstract: Current dataset free methods are either computationally expensive, require a noise model, or have inadequate image quality. We show that a simple 2 layer network, without any training data or noise model, can achieve high quality image denoising at low computational cost. Experiments on artificial, real world camera, and microscope noise show that our method termed Zero Shot Noise2Noise often outperforms existing methods, while significantly reducing compute.

Contact: y.mansour@tum.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 83

SEMI-SUPERVISED MULTI-TASK HYPERGRAPHS FOR EARTH UNDERSTANDING

Marcu A., Pirvu M., Costea D., Haller E., Slusanschi E., Belbachir N., Sukthankar R., Leordeanu M.

Abstract: There are many ways to interpret the world, some highly interdependent. We introduce a multi-task hypergraph method for scene understanding in UAV flights and Earth Observations. Each node is an interpretation layer, and each hyperedge, modeled by a lightweight neural network, predicts an output node. Nodes can be inputs and outputs in different hyperedges, creating robust pseudolabels. Using semi-supervised learning, we show improvements, adapting to data shifts and recovering missing data.

Contact: alymarcu91@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 84

ON GOOD PRACTICES FOR TASK-SPECIFIC DISTILLATION OF LARGE PRETRAINED VISUAL MODELS

Marrie J., Arbel M., Mairal J., Larlus D.

Abstract: Large pretrained visual models exhibit remarkable generalization across diverse recognition tasks. Yet, real-world applications often demand compact models tailored to specific problems. These can be learned by distilling knowledge from larger models. In this paper, we show that the excellent robustness and versatility of recent pretrained models challenge common practices established in the literature, calling for a new set of optimal guidelines for task-specific distillation.

Contact: juliette.marrie@inria.fr

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 85

EYES WIDE UNSHUT: UNSUPERVISED MISTAKE DETECTION IN EGOCENTRIC VIDEO BY DETECTING UNPREDICATABLE GAZE

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

Abstract: In this paper, we address the challenge of unsupervised mistake detection in egocentric video through the analysis of gaze signals, a critical component for advancing user assistance in smart glasses. Traditional supervised methods, reliant on manually labeled mistakes, suffer from domain-dependence and scalability issues. This research introduces an unsupervised method for detecting mistakes in videos of human activities, overcoming the challenges of domain-specific requirements and the necessity for annotated data. By analyzing unusual gaze patterns that signal user disorientation during tasks, we propose a gaze completion model that forecasts eye gaze trajectories from incomplete inputs. The difference between the anticipated and observed gaze paths acts as an indicator for identifying errors. Our method is validated on the EPIC-Tent and EGTEA Gaze+ datasets, showing its superiority compared to current supervised and unsupervised techniques.

Contact: michele.mazzamuto@phd.unict.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 86

IMPROVING ZERO-SHOT GENERALIZATION OF PROMPT LEARNING VIA UNSUPERVISED KNOWLEDGE DISTILLATION

Mistretta M., Baldrati A., Bertini M., Bagdanov A.

Abstract: Vision-Language Models (VLMs) demonstrate remarkable zero-shot generalization to unseen tasks, but when is not enough, they do not adapt well to new distribution with limited data. Prompt learning is emerging as a parameter-efficient method for adapting VLMs but typical requires annotated samples. We propose Knowledge Distillation for Prompt Learning (KDPL), a label-free approach that can be seamlessly integrated into existing prompt learning techniques improving zero-shot generalization.

Contact: marco.mistretta@unifi.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 87

EGOCENTRIC VISION FOR ACTIVE ASSISTED LIVING

Mucha W.

Abstract: The focus of this research is the use of wearable cameras to understand and improve health and lifestyle from egocentric videos. As many health-related activities involve the use of our hands, this work aims to improve egocentric hand pose estimation and action recognition. The successful application of these methods to real-world AAL tasks requires domain adaptation research. Finally, to demonstrate the applicability of this work, we develop a method for hand rehabilitation of stroke patients.

Contact: wiktor.mucha@tuwien.ac.at

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 88

AFFORDANCES AND ATTENTION MODELS FOR SHORT-TERM OBJECT INTERACTION ANTICIPATION

Mur-Labadia L., Martinez-Cantin R, Guerrero JJ, Farinella G.M., Furnari A.

Abstract: Short-Term object interaction Anticipation (STA) involves detecting the next-active objects, interaction noun and verb categories, and time to contact from egocentric video. We propose STA-former, an attention based architecture with frame-guided temporal pooling, dual image-video attention, and multi-scale feature fusion. We also introduce an environment affordance model and interaction hotspot prediction. Results show up to +45% mAP improvement on Ego4D scoring 2nd on the CVPR 24' challenge

Contact: lmur@unizar.es

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 89

MACHINE VISION IN INDUSTRIAL APPLICATIONS

Neubauer M., Rueckert E.,

Abstract: The metal processing industry faces a significant challenge due to its dependence on high-quality metal scrap. At present, shredded metals are exported because of their low quality, and higher-quality scrap needs to be imported. Our goal is to enhance the quality of steel scrap to minimize resource loss during the recycling of metal composite waste and reduce Co2 emissions.

Contact: melanie.neubauer@unileoben.ac.at

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 90

SCANTALK: 3D TALKING HEADS FROM UNREGISTERED SCANS

Nocentini F., Besnier T., Ferrari C., Arguillere S., Berretti S., Daoudi M.

Abstract: Existing Speech-driven 3D talking heads methods are constrained by animating faces with fixed topologies, wherein point-wise correspondence is established, and the number and order of points remains consistent across all identities the model can animate. In this work, we present ScanTalk, a novel framework capable of animating 3D faces in arbitrary topologies including scanned data. Our approach relies on the DiffusionNet architecture to overcome the fixed topology constraint, offering promising avenues for more flexible and realistic 3D animations. By leveraging the power of DiffusionNet, ScanTalk not only adapts to diverse facial structures but also maintains fidelity when dealing with scanned data, thereby enhancing the authenticity and versatility of generated 3D talking heads. While our primary objective is to develop a generic method free from topological constraints, all state-of-the-art methodologies are bound by such limitations.

Contact: federico.nocentini@unifi.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 91

LOWFORMER: HARDWARE EFFICIENT DESIGN FOR CONVOLUTIONAL TRANSFORMER BACKBONE

Nottebaum M., Dunnhofer M., Prof. Micheloni C.

Abstract: Research in efficient backbones evolves into models that are a mixture of convolutions and transformer blocks. A smart combination of both, architecturewise and component-wise is mandatory to excel in the speed-accuracy trade-off. Most publications focus on maximizing accuracy and minimizing MACs (multiply accumulate), however MACs often don't represent well how fast a model actually is due to factors such as memory hierarchy, a universal concept that defines execution time on all kind of computing devices.We analyzed common modules and architectural design choices for backbones not in terms of MACs, but rather in actual execution time on several GPUs and a CPU. We applied the conclusions taken from that to design a new family of backbone networks that maximizes the speed-accuracy trade-off. Additionally we introduce a slimmed-down version of Multi-head Attention, that aligns with these findings and reduces the problem of quadratic complexity in terms of resolution that persists for traditional attention. Our new architecture achieves a remarkable speedup in terms of throughput on CPU and GPU, compared to current state-of-the-art models, while achieving a similar or better accuracy.

Contact: nottebaum.moritz@spes.uniud.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 92

SEGMENTING PUPILS WITH ELLIPTICAL FOURIER SERIES

O'MAHONY F., DIERKES K., DREWS M.

Abstract: Accurate segmentation of pupils in near-eye images is paramount for contour-based gaze-estimation methods. Based on the observation that pupils appear as deformed ellipses in these images, we introduce a new deep-learning segmentation method detecting pupil contours as low-order elliptical Fourier series. Analysing our results, we discuss pitfalls of the area-based IOU metric commonly used in the segmentation domain and develop a perimeter-based evaluation metric instead.

Contact: flx@pupil-labs.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 93

FEW-SHOT UNSUPERVISED IMPLICIT NEURAL SHAPE REPRESENTATION LEARNING WITH SPATIAL ADVERSARIES

Ouasfi A., Boukhayma A.

Abstract: Learning SDFs from sparse 3D point clouds in the absence of ground truth supervision remains a very challenging task. While recent methods rely on smoothness priors to regularize the learning, our method introduces a regularization term that leverages adversarial samples around the shape to improve the learned SDFs. We illustrate the efficacy of our proposed method, highlighting its capacity to improve SDF learning with respect to baselines and the state-of-the-art using synthetic and real data.

Contact: amine1ouasfi@hotmail.fr

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 94

VISION-LANGUAGE MODELS FOR HOLISTIC OR DOMAIN MODELING

Özsoy E., Pellegrini C., Keicher M., Czempiel T., Örnek P. E., Navab N.

Abstract: Surgical procedures take place in highly complex operating rooms (OR), involving medical staff, patients, devices and their interactions. We believe Vision-Language Models open a up a pathway for automated, comprehensive and semantic understanding and modeling of the OR domain through semantic scene graphs (SSG). Our work paves the way for enabling more efficient and precise decision-making during surgical procedures, and ultimately improving patient safety and surgical outcomes.

Contact: ege.oezsoy@tum.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 95

LUCIDPPN: UNAMBIGUOUS PROTOTYPICAL PARTS NETWORK FOR USER-CENTRIC INTERPRETABLE COMPUTER VISION

Pach M., Rymarczyk D., Lewandowska K., Tabor J., Zieliński B.

Abstract: We introduce LucidPPN, a novel prototypical parts network that separates color from other visual features. It also identifies prototypical parts corresponding to semantic parts of classified objects. Our experiments demonstrate that LucidPPN has comparable accuracy to baseline methods and generates less ambiguous prototypical parts, enhancing user understanding.

Contact: mateusz.pach@student.uj.edu.pl

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 96

GENERATIVE TIMELINES FOR INSTRUCTED VISUAL ASSEMBLY

Pardo A., Wang J., Ghanem B., Sivic J., Russell B., Caba F.

Abstract: The objective of this work is to manipulate visual timelines (e.g. a video) through natural language instructions, making complex timeline editing tasks accessible to non-expert or potentially even disabled users. We call this task Instructed visual assembly. This task is challenging as it requires (i) identifying relevant visual content in the input timeline as well as retrieving relevant visual content in a given input (video) collection, (ii) understanding the input natural language instruction, and (iii) performing the desired edits of the input visual timeline to produce an output timeline. To address these challenges, we propose the Timeline Assembler, a generative model trained to perform instructed visual assembly tasks. The contributions of this work are three-fold. First, we develop a large multimodal language model, which is designed to process visual content, compactly represent timelines and accurately interpret timeline editing instructions. Second, we introduce a novel method for automatically generating datasets for visual assembly tasks, enabling efficient training of our model without the need for human-labeled data. Third, we validate our approach by creating two novel datasets for image and video assembly, demonstrating that the Timeline Assembler substantially outperforms established baseline models, including the recent GPT-4o, in accurately executing complex assembly instructions across various real-world inspired scenarios.

Contact: alejandro.pardo@kaust.edu.sa

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 97

LONG-TERM TYPHOON TRAJECTORY PREDICTION: A PHYSICS-CONDITIONED APPROACH WITHOUT REANALYSIS DATA

Park Y., Seo M., Kim D., Kim H., Choi S., Choi B., Ryu J., Son S., Jeon H., Choi Y.

Abstract: As climate change escalates, accurate typhoon prediction becomes vital. Traditional models are comprehensive but slow, relying on outdated reanalysis data. We propose a new model using real-time Unified Model (UM) data, providing forecasts every 6 hours for up to 72 hours, outperforming existing methods. Our PHYSICS TRACK dataset includes ERA5 reanalysis, typhoon best-track, and UM forecast data to enhance prediction accuracy.

Contact: youngjae.park@gm.gist.ac.kr

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 98

GENERALIZED DEPTH PERCEPTION FROM EVERYDAY SENSORS

Park J., Jeon H.

Abstract: Depth estimation remains a critical challenge in computer vision, essential for robust perception across diverse scenes and sensors. Traditional methods, both optimization-based and learning-based, rely heavily on extensive pixel-wise labeled data and frequently encounter scale issues and systematic biases like density, sensing patterns, and scan range, which hinder their real-world applicability due to poor generalization across different sensors and environments. In response, we introduce the concept of Universal Depth Completion (UniDC) and a baseline architecture that utilizes a foundation model for monocular depth estimation to develop a comprehensive understanding of 3D structures, thereby enhancing adaptability and generalization. To address depth measurement biases, we implement a novel depth prompt module that disentangles input modalities such as images and depth, allowing for prompt engineering to adapt feature representation to new depth distributions and mitigate biases from various sensor types or scene configurations. Furthermore, we employ a pixel-wise affinity map generated from the foundation model to adjust depth information from arbitrary sensors to the monocular estimates. By embedding the learned features into hyperbolic space, which builds implicit hierarchical structures of 3D data, our model achieves strong adaptation to various depth sensors with minimal labeled data and provides absolute scale depth maps free from depth scan range constraints, promising significant improvements in the versatility and effectiveness of depth estimation techniques in the wild.

Contact: jinhwipark@gm.gist.ac.kr

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 99

HARLEQUIN: COLOR-DRIVEN GENERATION OF SYNTHETIC DATA FOR REFERRING EXPRESSION COMPREHENSION

Parolari L., Izzo E., Ballan L.

Abstract: Context. Referring Expression Comprehension (REC) aims to identify an object in a scene described by text. Existing dataset are manually collected and annotated.

Aim. To collect new datasets leveraging generative AI, enabling diversity and controllability in data.

Method. We generate a new dataset in two steps: (1) process existing data to create variations in the annotations; (2) generate an image using altered annotations as guidance.

Result. A new dataset for pre-training REC models.

Contact: luca.parolari@studenti.unipd.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 100

BREAKING THROUGH FAILURE BARRIERS OF STRUCTURE-FROM-MOTION

Pataki Z.

Abstract: Common sources of failure in Structure-from-Motion (SfM): Images across diverse conditions Scenes with sparse views and little visual overlap Image with Low baseline/parallax We cover two works that together facilitate accurate reconstruction in these scenarios. iCDC trains feature that are robust to diverse conditions (1) MonoSfM leverages data-driven priors to break traditional SfM requirements (2+3)

Contact: zador.pataki@inf.ethz.ch

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 101

VISION-LANGUAGE MODELS FOR MEDICAL IMAGE UNDERSTANDING

Pellegrini C., Keicher M., Özsoy E., Busam B., Navab N.

Abstract: In radiology, the integration of vision-language models (VLMs) offers transformative potential for medical image analysis. By combining visual data with advanced language understanding, these models can enhance automated diagnosis, structured reporting, and interactive clinical assistance while benefiting from internal LLM knowledge and external knowledge sources. The natural language interface further can enhance trust in clinical AI systems by providing intuitive explanations and enabling a collaborative diagnostic process.

Contact: chantal.pellegrini@tum.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 102

ROBUST WIND TURBINE BLADE SEGMENTATION FROM RGB IMAGES IN THE WILD

Pérez-Gonzalo R., Espersen A., Agudo A.

Abstract: The wind industry's growth demands automated data-driven maintenance solutions such as automatic blade region identification. We propose a novel segmentation algorithm that strengthens the U-Net by a tailored loss, pooling the focal and contiguity losses. Post-processing steps are proposed to ensure a reliable, generic, robust and efficient algorithm. Our approach fills holes enclosed by blade pixels and image boundaries, and corrects misclassified pixels with an on-the-fly random forest.

Contact: rperez@iri.upc.edu

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 103

PREGO:ONLINE MISTAKE DETECTION IN PROCEDURAL EGOCENTRIC VIDEOS

Flaborea A., D'Amely G., Plini L., Scofano L., De Matteis E., Furnari A., Farinella G., Galasso F.

Abstract: Identifying procedural errors in egocentric videos is crucial for real-time mistake detection in fields like manufacturing. We propose PREGO, the first online one-class classification model for mistake detection in procedural egocentric videos, leveraging online action recognition and symbolic reasoning. We evaluate PREGO on Assembly101 and Epic-tent specifically adapted for online procedural mistake detection.

Contact: leonardo.plini@uniroma1.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 104

DEEPFEATUREX NET: DEEP FEATURES EXTRACTORS BASED NETWORK FOR DISCRIMINATING SYNTHETIC FROM REAL IMAGES

Pontorno O., Guarnera L., Battiato S.

Abstract: Deepfakes, synthetic media created by deep learning algorithms, are a major challenge in Digital Forensics. Current methods struggle to generalize, performing poorly with new, unseen architectures. In this work we propose a novel approach that uses three blocks, called ”Base Models”, to extract features from specific image classes (Diffusion Model-generated, GAN-generated, or real) with unbalanced datasets. These features are combined to determine the image’s origin.

Contact: opontorno@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 105

MERGING AND SPLITTING DIFFUSION PATHS FOR SEMANTICALLY COHERENT PANORAMAS

Quattrini F., Pippi V., Cascianelli S., Cucchiara R.

Abstract: Diffusion models are the State-of-the-Art for text-to-image generation, and increasing research effort has been dedicated to adapting the inference process of pretrained diffusion models to achieve zero-shot capabilities. An example is the generation of long images, which has been tackled in recent works by combining strided diffusions over overlapping latent features. This yield perceptually aligned but semantically incoherent panoramas. We propose the Merge-Attend-Diffuse operator (MAD), pluggable into different types of diffusion models featuring attention operations, and an inference-time strategy for generating perceptually and semantically coherent panoramas.

Contact: fabio.quattrini@unimore.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 106

SYNCHRONIZATION IS ALL YOU NEED: EXOCENTRI TO-EGOCENTRIC TRANSFER FOR TEMPORAL ACTION SEGMENTATION WITH UNLABELED SYNCHRONIZED VIDEO PAIRS

Quattrocchi C., Furnari A., Di Mauro D., Giuffrida M. V., Farinella G. M.

Abstract: We consider the problem of transferring a temporal action segmentation system initially designed for exocentric (fixed) cameras to an egocentric scenario, where wearable cameras capture video data. The conventional supervised approach requires the collection and labeling of a new set of egocentric videos to adapt the model, which is costly and time-consuming. Instead, we propose a novel methodology which performs the adaptation leveraging existing labeled exocentric videos and a new set of unlabeled, synchronized exocentric-egocentric video pairs, for which temporal action segmentation annotations do not need to be collected. We implement the proposed methodology with an approach based on knowledge distillation, which we investigate both at the feature and Temporal Action Segmentation model level. Experiments on Assembly101 and EgoExo4D demonstrate the effectiveness of the proposed method against classic unsupervised domain adaptation and temporal alignment approaches. Without bells and whistles, our best model performs on par with supervised approaches trained on labeled egocentric data, without ever seeing a single egocentric label, achieving a +15.99 improvement in the edit score (28.59 vs 12.60) on the Assembly101 dataset compared to a baseline model trained solely on exocentric data. In similar settings, our method also improves edit score by +3.32 on the challenging EgoExo4D benchmark.

Contact: camillo.quattrocchi@phd.unict.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 107

THE SIMPLER THE BETTER: AN ENTROPY-BASED IMPORTANCE METRIC TO REDUCE DEEP NEURAL NETWORKS' DEPTH

Quétu V., Liao Z., Tartaglione E.

Abstract: While DNNs are highly effective at solving complex tasks, large pre-trained models are commonly employed even to solve consistently simpler downstream tasks, which do not necessarily require a large model's complexity. We propose an efficiency strategy that leverages prior knowledge transferred by large models. Simple but effective, we propose a method relying on an Entropy-bASed Importance mEtRic (EASIER) to reduce the depth of over-parametrized DNNs, which reduces their computational burden.

Contact: victor.quetu@telecom-paris.fr

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 108

DETECTION OF IMAGE MANIPULATIONS USING DCT STATISTICS

Ragaglia C., Guarnera F., Battiato S., Puglisi G.

Abstract: In my first PhD course year, me and my research team developed novel algorithms for detecting image manipulations such as cropping, resizing, rotation, and shifting using Discrete Cosine Transform (DCT) statistics. The algorithms analyze the frequency components of images to identify manipulations by comparing DCT coefficients. First method is about Cropping Detection: it uses DCT statistics and a SVM classifier to detect cropping in images by identifying the original resolution based on distinctive patterns in the β values of AC distributions. Second method focuses on Transformations Detection: this approach detects image transformations such as scaling, rotation, and shifting by applying DCT to image blocks, computing histograms of DCT coefficients, and measuring Euclidean distances to identify frequency detail divergences.

Contact: claudio.ragaglia@phd.unict.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 109

ENIGMA-51: TOWARDS A FINE-GRAINED UNDERSTANDING OF HUMAN BEHAVIOR IN INDUSTRIAL SCENARIOS

Ragusa F., Leonardi R., Mazzamuto M., Bonanno C., Scavo R., Furnari A., Farinella G. M.

Abstract: ENIGMA-51 is a new egocentric dataset acquired in an industrial scenario by 19 subjects who followed instructions to complete the repair of electrical boards using industrial tools (e.g., electric screwdriver) and equipments (e.g., oscilloscope). The 51 egocentric video sequences are densely annotated with a rich set of labels that enable the systematic study of human behavior in the industrial domain. We provide benchmarks on four tasks related to human behavior: 1) untrimmed temporal detection of human-object interactions, 2) egocentric human-object interaction detection, 3) short-term object interaction anticipation and 4) natural language understanding of intents and entities. Baseline results show that the ENIGMA-51 dataset poses a challenging benchmark to study human behavior in industrial scenarios. We publicly release the dataset at: https://iplab.dmi.unict.it/ENIGMA-51/

Contact: francesco.ragusa@unict.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 110

BEYOND ARTIFACTS: IMPROVED DEEP LEARNING-POWERED MAR WITH DOMAIN TRANSFORMATION FOR RADIOTHERAPY PLANNING

Rehman M.,Micheloni C.,Martinel N.,Spizzo R.,Avanzo M.

Abstract: For planning radiotherapy treatments of head and neck cancers, Computed Tomography (CT) scans are typically employed. However, the quality of standard CT scans, generated using kilo-Voltage (kV) X-ray tube potentials, is often compromised by streak artifacts caused by metallic implants, such as dental fillings. To address this issue, some radiotherapy devices offer Mega-Voltage CT (MVCT) for daily patient setup verification. Due to the higher energy X-rays used in MVCT, these scans are largely free from artifacts, making them more suitable for accurate radiotherapy treatment planning. Our methodology integrates the attention-gate mechanism within a Residual Convolutional Neural Network (CNN), facilitating the selective filtration of important features critical for precise image reconstruction. By leveraging attentive feature fusion, we effectively reconstruct damaged regions within kVCT scans, thereby mitigating the impact of artifacts on diagnostic accuracy. This novel approach demonstrates significant success, achieving a mean Peak Signal-to-Noise Ratio (PSNR) of 31.249 dB across the entire patient volume and 27.753 dB in artifact-affected regions. Notably, the PSNR calculation focuses exclusively on the region of interest, excluding the background. Additionally, we explored domain transformation from the kVCT domain to the MVCT domain, which holds considerable promise for enhancing radiotherapy calibration and treatment planning. This transformative technique underscores the potential for improved clinical outcomes through superior image quality and artifact reduction.

Contact: rehman.mubashara@spes.uniud.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 111

EASG: ACTION SCENE GRAPHS FOR LONG-FORM UNDERSTANDING OF EGOCENTRIC VIDEOS

Rodin I., Furnari A., Min K., Tripathi S., Farinella G.M.

Abstract: Egocentric Action Scene Graphs (EASGs) is a new representation for long-form understanding of egocentric videos. EASGs provide a temporally evolving graph-based description of the camera wearer's actions, interacted objects, their relationships, and how actions unfold in time. We extend the Ego4D dataset with manually labeled EASGs. Along with the dataset, we establish the benchmark for EASG generation and show the effectiveness of EASGs for the downstream long-form video understanding tasks.

Contact: ivan.rodin@unict.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 112

CROWDSIM++: UNIFYING CROWD NAVIGATION AND OBSTACLE AVOIDANCE

Rosano M., Leocata D., Furnari A., Farinella G. M.

Abstract: This is the abstract

Contact: marco.rosano@unict.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 113

ICICLE: INTERPRETABLE CLASS INCREMENTAL CONTINUAL LEARNING

Rymarczyk D., van de Weijer J., Zieliński B., Twardowski B.

Abstract: Continual learning enables incremental learning of new tasks without forgetting those previously learned.

However, continual learning poses new challenges for interpretability, as the rationale behind model predictions may change over time, leading to interpretability concept drift. We address this problem by Interpretable Class-InCremental LEarning (ICICLE). It consists of: interpretability regularization, proximity-based initialization, and task recency bias compensation.

Contact: dawid.rymarczyk@uj.edu.pl

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 114

FOLLOWING THE HUMAN THREAD IN SOCIAL NAVIGATION

Scofano L.*, Sampieri A.*, Campari T.*, Sacco V., Spinelli I., Ballan L., Galasso F.

Abstract: Social Dynamics Adaptation model (SDA) is the first model to infer social dynamics based on the robot's state-action history. We propose a two-stage Reinforcement Learning framework: the first learns to encode human trajectories into social dynamics and learns a conditioned motion policy. In the second stage, the trained policy operates without direct access to trajectories. Tested on the novel Habitat 3.0 platform, SDA sets a novel state of the art performance in Social Navigation.

Contact: sacco@di.uniroma1.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 115

SCALING ANOMALY DETECTION TO WIDE IMAGES

Samele S., Matteucci M.

Abstract: There are many challenges in making anomaly detection algorithms on product images deployable in real-world industrial scenarios. Our work aims to develop an algorithm that can effectively scale to larger, more complex objects. We present a method, SADSeM (Scaling Anomaly Detection with Segmentation Models), based on classic convolutional neural networks for segmentation, such as Mask-RCNN. Because of these models' ability to learn and internally represent a given object's structure, we can build a pipeline leveraging both the segmentation maps produced by the model and its embeddings to perform unsupervised anomaly detection. As the segmentation task is effectively solved by these models independently from the size of the image, we achieve state-of-the-art anomaly detection and segmentation performances in industrial product images, scaling to higher-resolution images with more effectiveness than competitors.

Contact: stefano.samele@polimi.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 116

TOWARDS SUSTAINABLE EV BATTERY RECYCLING: CV AND RL BASED BATTERY DISASSEMBLY

Sarno F., Isoaho J., Özen Ö., Rätz R., Ochsenbein C.

Abstract: Automatic disassembly of electric vehicle (EV) batteries is critical due to safety concerns for human operators and challenges posed by a lack of training data for automated systems. This research introduces a novel approach combining computer vision (CV) and reinforcement learning (RL) to address these issues. We demonstrate (1) the effective detection and segmentation of EV battery components using neural networks trained on synthetic data and (2) the feasibility of using RL to learn policies for interacting with the battery. Our findings show promising results for safer and more efficient battery disassembly.

Contact: sarnof96@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 117

TOWARDS AUTOMATED REGULATION OF JACOBAEA VULGARIS IN GRASSLAND USING DEEP NEURAL NETWORKS

Schauer M., Hohl R., Vaupel D., Soethe N., Ghobadi S. E.

Abstract: The highly poisonous ragwort (Jacobaea Vulgaris) is increasingly spreading, posing significant risks to agriculture, livestock, and nature conservation due to toxic pyrrolizidine alkaloids (PAs). Current manual control methods, like plucking, are labor-intensive and time-consuming. We introduce a workflow for automated regulation of ragwort, consisting of deep learning-based monitoring and controlling. The goal is to detect and control ragwort early before reseeding, focusing on detecting green leaf rosettes on a meadow.

Contact: moritz.schauer@mni.thm.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 118

TWO EFFECTS, ONE TRIGGER: ON THE MODALITY GAP, OBJECT BIAS, AND INFORMATION IMBALANCE IN CONTRASTIVE VISION-LANGUAGE REPRESENTATION LEARNING

Schrodi S., Hoffmann D.T., Argus M., Fischer V., Brox T.

Abstract: Contrastive vision-language models have achieved great success on some tasks, but have performed surprisingly poorly on others. Previous work has attributed these problems to the modality gap or a bias towards objects. We investigate both and find that only a few embedding dimensions drive the modality gap, object bias does not lead to poorer performance on other factors such as attributes, and information imbalance between modalities is the driving factor for both.

Contact: schrodi@cs.uni-freiburg.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 119

DIFFERENTIABLE TASK GRAPH LEARNING: PROCEDURAL ACTIVITY REPRESENTATION AND ONLINE MISTAKE DETECTION FROM EGOCENTRIC VIDEOS

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

Abstract: Procedural activities are sequences of key-steps aimed at achieving specific goals. They are crucial to build intelligent agents able to assist users effectively. In this context, task graphs have emerged as a human-understandable representation of procedural activities, encoding a partial ordering over the key-steps. While previous works generally relied on hand-crafted procedures to extract task graphs from videos, in this paper, we propose an approach based on direct maximum likelihood optimization of edges' weights, which allows gradient-based learning of task graphs and can be naturally plugged into neural network architectures. Experiments on the CaptainCook4D dataset demonstrate the ability of our approach to predict accurate task graphs from the observation of action sequences, with an improvement of +16.7

- Compressed File Formats (CFFs): standards for compressing information and encoding it into a byte structure. Compressed-Language Models (CLMs): language models - operating on raw byte streams from CFFs.

Question: Can CLMs understand files compressed by CFFs?

Short answer: Yes, they can!

Long answer: - Our evidence suggests CLMs operating on JPEG files can (1) Recognize file properties. (2) Handle files with anomalies. (3) Generate new files.

Contact: mattia.soldan@kaust.edu.sa

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 125

WHEN DOES PERCEPTUAL ALIGNMENT BENEFIT VISION REPRESENTATIONS?

Sundaram S., Fu S

Abstract: We find that finetuning state-of-the-art vision models on human similarity judgments improves upon the original backbones across many downstream tasks, including counting, semantic segmentation, depth estimation, instance retrieval, and retrieval-augmented generation. Performance is largely preserved on other tasks, including specialized domains such as medical imaging. Our results suggest that injecting an inductive bias about human perception into vision models can benefit representations.

Contact: shobhita@mit.edu

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 126

EXPLOITING ACTIVATION SPARSITY WITH DENSE TO DYNAMIC-K MIXTURE-OF-EXPERTS CONVERSION

Szatkowski F., Wójcik B., Piórczyński M., Scardapane S.

Abstract: Activations of transformer models are highly sparse, so their inference cost can be reduced by conversion into an equivalent Mixture-of-Experts(MoE). We investigate the impact of sparsity on the conversion and show that higher sparsity leads to better efficiency. We also introduce dynamic-k expert selection that adjusts the number of executed experts on a per-token basis. Our method, Dense to Dynamic-k Mixture-of-Experts (D2DMoE), outperforms existing approaches on common NLP and vision tasks, allowing us to save up to 60% of inference cost.

Contact: fszatkowski96@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 127

THE ILLUSION OF REALITY: REAL-TIME REALISM IN AR/VR

Tariq T.

Abstract: I do not add a traditional abstract to my posters, I add easy to understand quick glance text and points, so I am posting them below.

GOAL: PASSING THE VISUAL TURING TEST

"Realizing the dream of real-time VR/AR that is perceptually indistinguishable from how we see the visual world through our eyes"

THE DIMENSIONS OF VISUAL REALISM

Realistic Luminance: Capturing the Dynamic Range of the visual world

Spatial Realism: Spatial Equivalence with the visual world

Realistic Motion: Equivalence of Movement Detection and Perceptual Quantification

Realistic 3D space: Equivalence in the perception of Depth and Distances

Contact: taimoor.tariq@usi.ch

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 128

LEARNING OBJECT STATES FROM ACTIONS VIA LARGE LANGUGE MODELS

Tateno M., Yagi T., Furuta R., Sato Y.

Abstract: Temporal localization of object states (e.g., cracked egg) in videos is crucial but challenging due to ambiguity and variety, leading to a lack of training data. We propose to extract the object state information from action information included in narrations using LLMs. Our context-aware framework generates object state pseudo-labels from narrations to train a classification model. Our model shows significant improvement in the newly collected datasets over strong vision-language models.

Contact: masatate@iis.u-tokyo.ac.jp

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 129

FANTASTIC GAINS AND WHERE TO FIND THEM

Roth K., Thede L., Koepke A. S., Vinyals O., Henaff O., Akata Z.

Abstract: Training deep networks involves various design choices, leading models to learn unique features. We explore transferring complementary knowledge between pretrained models without performance loss, even from weaker models. Using ImageNet-trained models, we reveal standard knowledge distillation's limits and propose a robust, unsupervised transfer method through data partitioning, demonstrating successful transfer across diverse models.

Contact: Lukas.Thede@uni-tuebingen.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 130

CLIMB3D: CONTINUAL LEARNING FOR IMBALANCED 3D INSTANCE SEGMENTATION

Thengane V., Lahoud J., Khan S., Khan F., Li Y.

Abstract: 3D instance segmentation identifies objects in physical space but often overlooks realism. Real-world environments evolve, necessitating continual learning with class imbalance. We propose three realistic scenarios based on object frequency, semantics, and random grouping to enhance understanding. Our framework uses exemplar replay and knowledge distillation, with a component considering object occurrence frequency. Evaluation on ScanNet200 shows superior performance, with improvements in respective scenarios.

Contact: v.thengane@surrey.ac.uk

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 131

FROM CORRESPONDENCES TO POSE: NON-MINIMAL CERTIFIABLY OPTIMAL RELATIVE POSE WITHOUT DISAMBIGUATION

Tirado-Garín J., Civera J.

Abstract: This work addresses the four-fold ambiguity in calibrated epipolar geometry: Our method directly estimates the relative camera pose between two images using only 2D correspondences instead of just the essential matrix.

Contact: jtiradogarin@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 132

PHYSICS-INFORMED MODELING OF DYNAMIC HUMANS AND THEIR INTERACTIONS

TRIPATHI S., TZIONAS D., BLACK M.

Abstract: Understanding how humans use physical contact to interact with the world is key to enabling human-centric artificial intelligence. Existing methods estimate 3D bodies in the camera coordinates, often producing physically-implausible poses and interactions. Similarly, existing human motion generation models suffer from physical and biomechanical artifacts when viewed in the context of the scene. This is because most methods ignore the fact that people move in a scene, interact with it, and receive physical support by contacting it. We go beyond existing methods by introducing IPMAN, a 3D human pose estimation method that leverages ground support and novel intuitive physics (IP) terms to estimate physically plausible 3D bodies from a single color image. IPMAN's IP terms, however, only apply to static 3D poses. In HUMOS, we extend IPMAN to propose general IP terms that are effective in generating dynamic human motions conditioned on body shape. While IPMAN and HUMOS accurately model interactions with the ground, humans routinely interact with diverse objects and scenes. To enable physical reasoning for general human-object interactions, our final work DECO infers dense vertex-level 3D contacts on the full human body. Thus, our research represents a step towards realistic modeling of humans and their interactions in complex real-world scenarios.

Contact: shashank.tripathi123@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 133

RELATIVE POSE FROM 4 POINTS IN 3 VIEWS

Tzamos C., Barath D., Sattler T., Kukelova Z.

Abstract: We study challenging problems of estimating the relative pose of three cameras and propose novel efficient solutions to the difficult configuration of four points in three calibrated views, known as the 4p3v problem, and to the unsolved configuration of four points in three views with unknown equal focal length, the 4p3vf problem. Our solutions rely on the idea of generating one or two virtual correspondences in two views by using the locations of the four correspondences in the three views.

Contact: tzamos.charalampos@fel.cvut.cz

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 134

WHAT DO MULTIMODAL MODELS KNOW ABOUT MOTION BLUR?

Vartore A., Ortis A., Mahiddine A., Battiato S.

Abstract: Motion deblurring is an ill-posed problem. To regularise it previous work uses either learnt or hand-crafted priors or additional inputs. Vision and language models learn semantic representations that have proven useful to downstream tasks in NLP and more recently computer vision. In this poster we explore the understanding of motion blur of such models and apply it to the task of deblurring.

Contact: andre.vartore@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 135

CONVEX DECOMPOSITION OF INDOOR SCENES

Vavilala V., Forsyth D.

Abstract: We present a method to parse indoor scenes into simple primitives with good normals, depth, and segmentation. A neural network accepts an RGBD image and predicts a good start point for convex parameters; a subsequent refinement step directly optimizes the parameters with respect to the training losses. Further, we can introduce negative primitives, utilizing CSG and improving quality. Finally, ensembling networks dramatically improves fits. Our primitives can be useful in real image editing.

Contact: vv16@illinois.edu

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 136

CAMERA POSE ESTIMATION FROM BOUNDING BOXES

Vávra V., Sattler T., Kúkelová Z.

Abstract: Visual localization typically estimates the camera pose from matches between 2D pixel positions and 3D points. This can be memory intensive and lead to privacy risks. As an interesting alternative, we investigate multiple strategies based on converting bounding box correspondences to point correspondences and propose a novel and simple 2-point absolute camera pose solver, DP2P, that exploits the fact that the depths of the objects can be approximated from the sizes of their bounding boxes.

Contact: vavravac@fel.cvut.cz

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 137

COMIX: A COMPREHENSIVE BENCHMARK FOR MULTITASK COMICS UNDERSTANDING

Vivoli E., Bertini M., Karatzas D.

Abstract: The comic field is advancing quickly, but current evaluation metrics and datasets often lag behind, confined to small or single-style sets. We introduce CoMix, a new benchmark designed to assess the multi-task capabilities of comic analysis models. To counter the dominance of manga-style data, we have included a diverse set of American comic-style books. This initiative sets a new standard for comprehensive comic analysis, providing a robust benchmark for widespread community evaluation.

Contact: emanuele.vivoli@unifi.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 138

WHAT DO NEURAL NETWORKS LEARN IN IMAGE CLASSIFICATION? A FREQUENCY SHORTCUT PERSPECTIVE

Wang S., Veldhuis R., Brune C., Strisciuglio N.

Abstract: Frequency analysis is useful for understanding the mechanisms of representation learning in neural networks (NNs). Most research in this area focuses on the learning dynamics of NNs for regression tasks, while little for classification. This study empirically investigates the latter and expands the understanding of frequency shortcuts. First, we perform experiments on synthetic datasets, designed to have a bias in different frequency bands. Our results demonstrate that NNs tend to find simple solutions for classification, and what they learn first during training depends on the most distinctive frequency characteristics, which can be either low- or high-frequencies. Second, we confirm this phenomenon on natural images. We propose a metric to measure class-wise frequency characteristics and a method to identify frequency shortcuts. The results show that frequency shortcuts can be texture-based or shape-based, depending on what best simplifies the objective. Third, we validate the transferability of frequency shortcuts on out-of-distribution (OOD) test sets. Our results suggest that frequency shortcuts can be transferred across datasets and cannot be fully avoided by larger model capacity and data augmentation.

Contact: s.wang-2@utwente.nl

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 139

CNVID-3.5M: BUILD, FILTER, AND PRE-TRAIN THE LARGE-SCALE PUBLIC CHINESE VIDEO-TEXT DATASET

Gan T., Wang Q., Dong X., Ren X., Nie L., Guo Q.

Abstract: Owing to well-designed large-scale video-text datasets, recent years have witnessed tremendous progress in videotext pre-training. However, existing large-scale video-text datasets are mostly English-only. Though there are certain methods studying the Chinese video-text pre-training, they pre-train their models on private datasets whose videos and text are unavailable. This lack of large-scale public datasets and benchmarks in Chinese hampers the research and downstream applications of Chinese video-text pretraining. Towards this end, we release and benchmark CNVid-3.5M, a large-scale public cross-modal dataset containing over 3.5M Chinese video-text pairs. We summarize our contributions by three verbs, i.e., “Build”, “Filter”, and “Pre-train”: 1) To build a public Chinese video-text dataset, we collect over 4.5M videos from the Chinese websites. 2) To improve the data quality, we propose a novel method to filter out 1M weakly-paired videos, resulting in the CNVid-3.5M dataset. And 3) we benchmark CNVid3.5M with three mainstream pixel-level pre-training architectures. At last, we propose the Hard Sample Curriculum Learning strategy to promote the pre-training performance. To the best of our knowledge, CNVid-3.5M is the largest public video-text dataset in Chinese, and we provide the first pixel-level benchmarks for Chinese video-text pretraining.

Contact: qing.wang@qmul.ac.uk

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 140

FLATTENING THE PARENT BIAS: HIERARCHICAL SEMANTIC SEGMENTATION IN THE POINCARÉ BALL

Weber S., Zöngür B., Araslanov N., Cremers D.

Abstract: Although recent work on semantic segmentation shows outstanding accuracy by leveraging parent logits in the training process, we show that this progress is unrelated to semantic taxonomies. In contrast, flat classifiers, where parents are inferred from children, generalize much better. However, we reveal that Euclidean representations may suffer from a parent bias. Our analysis suggests that hyperbolic representation allows for mitigating this bias, and outperforms hierarchical approach in terms of accuracy and calibration, whatever the domain shift.

Contact: sim.weber@tum.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 141

T-DEED: TEMPORAL-DISCRIMINABILITY ENHANCER ENCODER-DECODER FOR PRECISE EVENT SPOTTING IN SPORTS VIDEOS

Xarles A., Escalera S., Moeslund T., Clapés A.

Abstract: Here we introduce T-DEED, a Temporal-Discriminability Enhancer Encoder-Decoder for Precise Event Spotting in sports videos. T-DEED addresses multiple challenges in the task, including the need for discriminability among frame representations, high output temporal resolution to maintain prediction precision, and the necessity to capture information at different temporal scales to handle events with varying dynamics. Its specialized architecture achieves SOTA performance on the FigureSkating and FineDiving datasets.

Contact: arturxe@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 142

DEMOCRATIZING HUMAN DIGITIZATION

Xiu Y.

Abstract: Reconstruction is a kind of Conditional Generation

Contact: yuliang.xiu@tuebingen.mpg.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 143

ENHANCING VIDEO SUPER-RESOLUTION VIA IMPLICIT RESAMPLING-BASED ALIGNMENT (CVPR2024 HIGHTLIGHT)

Xu K., Yu Z., Wang X., Bi M, Yao A.

Abstract: In video super-resolution, it is common to use a frame-wise alignment to support the propagation of information over time. The role of alignment is well-studied for low-level enhancement in video, but existing works overlook a critical step – resampling. We show through extensive experiments that for alignment to be effective, the resampling should preserve the reference frequency spectrum while minimizing spatial distortions. However, most existing works simply use a default choice of bilinear interpolation for resampling even though bilinear interpolation has a smoothing effect and hinders super-resolution. From these observations, we propose an implicit resampling-based alignment. The sampling positions are encoded by a sinusoidal positional encoding, while the value is estimated with a coordinate network and a window-based cross-attention. We show that bilinear interpolation inherently attenuates high-frequency information while an MLP-based coordinate network can approximate more frequencies. Experiments on synthetic and real-world datasets show that alignment with our proposed implicit resampling enhances the performance of state-of-the-art frameworks with minimal impact on both compute and parameters.

Contact: kxu@u.nus.edu

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 144

UNIMATCH: UNIFIED DENSE MATCHING FOR FLOW, STEREO, DEPTH AND GAUSSIAN SPLATTING

Xu H., Pollefeys M., Geiger A.

Abstract: We present UniMatch, a unified dense feature matching framework for four correspondence and 3D reconstruction tasks: optical flow, stereo matching, depth estimation, and feed-forward 3D Gaussian Splatting. We solve these tasks efficiently with a unified model by simply comparing feature similarities, where the features are obtained with a cross-view-aware Transformer. We achieve state-of-the-art performance on 12 standard datasets for four tasks while being simpler and more efficient than previous task-specific methods.

Contact: haofei.xu@inf.ethz.ch

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 145

KNOWLEDGE DECOMPOSITION AND REPLAY: A NOVEL CROSS-MODAL IMAGE-TEXT RETRIEVAL CONTINUAL LEARNING METHOD

Yang R., Wang S., Zhang H., Xu S., Guo Y., Ye X., Hou B., Jiao L.

Abstract: To enable machines to mimic human cognitive abilities and alleviate the catastrophic forgetting problem in cross-modal image-text retrieval (CMITR), this paper proposes a novel continual learning method, Knowledge Decomposition and Replay (KDR), which emulates the process of knowledge decomposition and replay exhibited by humans in complex and changing environments. KDR has two components: a feature Decomposition-based CMITR Model (DCM) and a cross-task Generic Knowledge Replay strategy (GKR). DCM decomposes text and image features into task-specific and generic knowledge features, mimicking the human cognitive process of knowledge decomposition. Specifically, it employs a generic knowledge features extraction module for all tasks and a task-specific module for each task with a few trainable fully connected layers. Similarly, GKR emulates the human behavior of knowledge replay by utilizing the image-text similarity matrix output from the old task model with inputting the previous samples to induce the learning of the image-text similarity matrix output from the current task model with inputting the previous samples, using knowledge distillation technology. To demonstrate the effect of KDR, we adapted a continual learning dataset Seq-COCO from MSCOCO. Extensive experiments on Seq-COCO showed that KDR reduces catastrophic forgetting and consolidates general knowledge, improving the model's learning ability in CMITR.

Contact: r_yang@stu.xidian.edu.cn

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 146

FIRE : FAST INVERSE RENDERING USING DIRECTIONAL AND SIGNED DISTANCE FUNCTIONS

Yenamandra T., Tewari A., Yang N., Bernard F., Theobalt C., Cremers D.

Abstract: We propose a novel neural scene representation, DDF defined on the unit sphere, for rendering images from our SDF model during inference with 1 forward pass through the model. We present an algorithm to reconstruct 3D shapes from single view depth maps using our DDF and SDF models, which is 15.5× per iteration faster than competing methods.

Contact: tarun.yenamandra@tum.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 147

IMAGE DEBLURRING GUIDED BY DEPTH MAP PROMPT

Yi Z., Valsesia D., Bianchi T., Magli E

Abstract: Dataset As an ill-posed linear inverse problem, deblurring is inherently challenging due to the unknown blur kernel and noise. Utilizing multimodal information, especially depth maps, in the deblurring process holds significant promise. In this work, we introduce a novel adapter structure, which integrates seamlessly with existing deblurring architectures, such as Restormer[1] and NAFNet[2], enabling the models to utilize depth information effectively. Through the prompt learning on the deblur-depth paired dataset, the adapter-deblurred models show superior performance on the validation dataset.

Contact: ziyao.yi@polito.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 148

FEDERATED LEARNING FOR DECENTRALIZED MODEL TRAINING IN SKIN CANCER HISTOPATHOLOGY

Zacharczuk J., Pisula J., Bozek K.

Abstract: Artificial intelligence can predict the effectiveness of treatment based on histopathology images. Clinical centers often lack enough data to train robust models. In the medical field, where patient privacy is crucial, data processing is challenging. Federated Learning (FL) addresses this problem by training distributed models. This study compares FL techniques to predict cutaneous Squamous Cell Carcinoma (cSCC) progression using Whole Slide Images (WSI) from multiple clinical centers.

Contact: zacharczuk.jakub2k@gmail.com

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 149

EFFICIENCY THROUGH SELF SUPERVISION IN GERIATRICS

Zedda L.

Abstract: Radiomics, the high-throughput extraction of quantitative imaging features from medical images, has shown great promise in augmenting clinical efforts to improve disease diagnosis and response prediction. The success of radiomics models is often limited by small dataset sizes. Self-supervised learning (SSL) techniques have emerged as a potential solution to this challenge. In our research, we investigate the use of a masked autoencoder (MAE) within a geriatric radiomics classification context.

Contact: luca.zedda@unica.it

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 150

ROAM: ROBUST AND OBJECT-AWARE MOTION GENERATION USING NEURAL POSE DESCRIPTORS

Zhang W., Dabral R., Leimkühler T., Golyanik V., Habermann M., Theobalt C.

Abstract: We present a character-scene interaction framework which generates realistic motions from walking to sitting and lying by firstly optimising a goal pose. Our system is trained on only one object per action in MoCap dataset and robust to unseen objects from the same category.

Contact: wzhang@mpi-inf.mpg.de

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 151

MASKED VIDEO AND BODY-WORN IMU AUTOENCODER FOR EGOCENTRIC ACTION RECOGNITION

Zhang M., Huang Y., Liu R., Sato Y.

Abstract: Compared with visual signals, Inertial Measurement Units (IMUs) placed on human limbs can capture accurate motion signals while being robust to lighting variation and occlusion. While these characteristics are intuitively valuable to help egocentric action recognition, the potential of IMUs remains under-explored. In this work, we present a novel method for action recognition that integrates motion data from body-worn IMUs with egocentric video. Due to the scarcity of labeled multimodal data, we design an MAE-based self-supervised pretraining method, obtaining strong multi-modal representation via modeling the natural correlation between visual and motion signals. To model the complex relation of multiple IMU devices placed across the body, we exploit the collaborative dynamics in multiple IMU devices and propose to embed the relative motion features of human joints into a graph structure. Experiments show our method can achieve state-of-the-art performance on multiple public datasets. The effectiveness of our MAE-based pretraining and graph-based IMU modeling are further validated by experiments in more challenging scenarios, including partially missing IMU devices and video quality corruption, promoting more flexible usages in the real world.

Contact: mfzhang@iis.u-tokyo.ac.jp

Presentation Type: Poster

Date: Tuesday 9 July 2024

Time: 21:30

Poster Session: 2

Poster Number: 152