Visual Genome: Connecting Language and Vision Using Annotations

International Journal of Computer Vision

123, 32-73

DOI: 10.1007/s11263-016-0981-7

Citation Report

#	Article	IF	Citations
1	Revisiting Visual Question Answering Baselines. Lecture Notes in Computer Science, 2016, , 727-739.	1.0	113
2	COCO Attributes: Attributes for People, Animals, and Objects. Lecture Notes in Computer Science, 2016, , 85-100.	1.0	39
3	Learning to Learn: Model Regression Networks for Easy Small Sample Learning. Lecture Notes in Computer Science, 2016, , 616-634.	1.0	140
4	Visual Relationship Detection with Language Priors. Lecture Notes in Computer Science, 2016, , 852-869.	1.0	444
5	SIGGRAPH Asia 2016. , 2016, , .		2
6	Semantic Extraction and Object Proposal for Video Search. Lecture Notes in Computer Science, 2017, , 475-479.	1.0	1
7	Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 2017, 163, 21-40.	3.0	199
8	Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 2017, 163, 3-20.	3.0	131
9	How deep learning extracts and learns leaf features for plant classification. Pattern Recognition, 2017, 71, 1-13.	5.1	379
10	Video Captioning via Sentence Augmentation and Spatio-Temporal Attention. Lecture Notes in Computer Science, 2017, , 269-286.	1.0	5
11	Attributes as Semantic Units Between Natural Language and Visual Recognition. Advances in Computer Vision and Pattern Recognition, 2017, , 301-330.	0.9	1
12	Enhanced Retrieval and Browsing in the IMOTION System. Lecture Notes in Computer Science, 2017, , 469-474.	1.0	13
13	Improving Visual Relationship Detection Using Semantic Modeling of Scene Descriptions. Lecture Notes in Computer Science, 2017, , 53-68.	1.0	28
14	Keyword-driven image captioning via Context-dependent Bilateral LSTM. , 2017, , .		6
15	Visual entity linking. , 2017, , .		4
16	Learning to Generate Descriptions of Visual Data Anchored in Spatial Relations. IEEE Computational Intelligence Magazine, 2017, 12, 29-42.	3.4	5
17	On support relations and semantic scene graphs. ISPRS Journal of Photogrammetry and Remote Sensing, 2017, 131, 15-25.	4.9	31
18	Image Forgery Detection Based on Semantic Image Understanding. Communications in Computer and Information Science, 2017, , 472-481.	0.4	1

#	Article	IF	Citations
19	Dense-Captioning Events in Videos. , 2017, , .		595
20	Inferring and Executing Programs for Visual Reasoning. , 2017, , .		239
21	Recurrent Visual Relationship Recognition with Triplet Unit. , 2017, , .		3
22	Tag Prediction at Flickr. , 2017, , .		3
23	Multi-Modal Knowledge Representation Learning via Webly-Supervised Relationships Mining. , 2017, , .		10
24	LoBIAG: A location-based collaborative image annotation game. , 2017, , .		2
25	Discriminative Bimodal Networks for Visual Localization and Detection with Natural Language Queries. , 2017, , .		31
26	Semantic Amodal Segmentation. , 2017, , .		81
27	Knowledge Acquisition for Visual Question Answering via Iterative Querying. , 2017, , .		29
28	Learning to Disambiguate by Asking Discriminative Questions. , 2017, , .		10
29	Deep Interactive Region Segmentation and Captioning. , 2017, , .		4
30	Generating Holistic 3D Scene Abstractions for Text-Based Image Retrieval. , 2017, , .		15
31	Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. , 2017, , .		105
32	Creation of a multi-paraphrase corpus based on various elementary operations. , 2017, , .		1
33	Video Visual Relation Detection. , 2017, , .		90
34	Image retrieval by dense caption reasoning. , 2017, , .		8
35	Deep Learningâ€"A New Era in Bridging the Semantic Gap. Intelligent Systems Reference Library, 2018, , 123-159.	1.0	4
36	A flexible testing environment for visual question answering with performance evaluation. Neurocomputing, 2018, 291, 128-135.	3.5	5

#	Article	IF	Citations
37	From image to language and back again. Natural Language Engineering, 2018, 24, 325-362.	2.1	0
38	Explicit ensemble attention learning for improving visual question answering. Pattern Recognition Letters, 2018, 111, 51-57.	2.6	16
39	Understanding visual scenes. Natural Language Engineering, 2018, 24, 441-465.	2.1	3
40	FVQA: Fact-Based Visual Question Answering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40, 2413-2427.	9.7	227
41	Progressive Stochastic Learning for Noisy Labels. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29, 5136-5148.	7.2	21
42	Multilabel Image Classification With Regional Latent Semantic Dependencies. IEEE Transactions on Multimedia, 2018, 20, 2801-2813.	5.2	133
43	Semantic Image Retrieval via Active Grounding of Visual Situations. , 2018, , .		6
44	Places: A 10 Million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40, 1452-1464.	9.7	1,833
45	RGB-D object detection and semantic segmentation for autonomous manipulation in clutter. International Journal of Robotics Research, 2018, 37, 437-451.	5.8	105
46	Robust Plackettâ€“Luce model for k-ary crowdsourced preferences. Machine Learning, 2018, 107, 675-702.	3.4	6
47	Face alignment recurrent network. Pattern Recognition, 2018, 74, 448-458.	5.1	18
48	Learning without prejudice: Avoiding bias in webly-supervised action recognition. Computer Vision and Image Understanding, 2018, 173, 24-32.	3.0	3
49	Flexible Sentence Analysis Model for Visual Question Answering Network. , 2018, , .		0
50	Cross-Dataset Adaptation for Visual Question Answering. , 2018, , .		10
51	Referring Relationships. , 2018, , .		64
52	R-VQA. , 2018, , .		49
53	Neural Motifs: Scene Graph Parsing with Global Context. , 2018, , .		506
54	Adapted GooLeNet for Visual Question Answering. , 2018, , .		3

#	Article	IF	Citations
55	Learning Convolutional Text Representations for Visual Question Answering. , 2018, , 594-602.		8
56	Paragraph Generation Network with Visual Relationship Detection. , 2018, , .		13
57	Relation Networks for Object Detection. , 2018, , .		801
58	Learning Answer Embeddings for Visual Question Answering. , 2018, , .		9
59	Towards Unconstrained Pointing Problem of Visual Question Answering: A Retrieval-based Method. , 2018, , .		2
60	Where and Who? Automatic Semantic-Aware Person Composition. , 2018, , .		33
61	Learning Representations Specialized in Spatial Knowledge: Leveraging Language and Vision. Transactions of the Association for Computational Linguistics, 2018, 6, 133-144.	3.2	5
62	Visual Manipulation Relationship Detection with Fully Connected CRFs for Autonomous Robotic Grasp. , 2018, , .		3
63	DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. , 2018, , .		1,294
64	Learning to Segment Every Thing. , 2018, , .		165
65	Zero-Shot Video Retrieval from a Query Phrase Including Multiple Concepts â€"Efforts and Challenges in TRECVID AVS Taskâ€". Journal of the Japan Society for Precision Engineering, 2018, 84, 983-990.	0.0	0
66	A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, 2018, 5, 180251.	2.4	84
67	Calligraphy: A Mobile Device Based Annotation Tool Supporting Learner Crowdsourcing. , 2018, , .		0
68	Fooling Vision and Language Models Despite Localization and Attention Mechanism. , 2018, , .		30
69	Visual Relationship Detection with Language prior and Softmax. , 2018, , .		6
70	Essay-Anchor Attentive Multi-Modal Bilinear Pooling for Textbook Question Answering. , 2018, , .		5
71	Identifying Exceptional Descriptions of People Using Topic Modeling and Subgroup Discovery. Lecture Notes in Computer Science, 2018, , 454-462.	1.0	2
72	Fast Parameter Adaptation for Few-shot Image Captioning and Visual Question Answering. , 2018, , .		25

#	Article	IF	Citations
73	Engagement Learning: Expanding Visual Knowledge by Engaging Online Participants. , 2018, , .		2
74	Csrs-Siat: A Benchmark Remote Sensing Dataset to Semantic-Enabled and Cross-Scales Scene Recognition. , 2018, , .		4
75	Detailed Sentence Generation Architecture for Image Semantics Description. Lecture Notes in Computer Science, 2018, , 423-432.	1.0	6
76	Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering. , 2018, , .		208
77	Learning Visual Knowledge Memory Networks for Visual Question Answering. , 2018, , .		41
78	VizWiz Grand Challenge: Answering Visual Questions from Blind People. , 2018, , .		204
79	Focal Visual-Text Attention for Visual Question Answering. , 2018, , .		67
80	Discriminability Objective for Training Descriptive Captions. , 2018, , .		104
81	Relevance Estimation with Multiple Information Sources on Search Engine Result Pages. , 2018, , .		14
82	Big Data Analytics. Lecture Notes in Computer Science, 2018, , .	1.0	2
83	Image Generation from Scene Graphs. , 2018, , .		425
84	Visual Relationship Detection Based on Local Feature and Context Feature. , 2018, , .		3
85	Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. , 2018, , .		85
86	Enhancing Visual Question Answering Using Dropout. , 2018, , .		4
87	Deep Image Understanding Using Multilayered Contexts. Mathematical Problems in Engineering, 2018, 2018, 1-11.	0.6	6
88	Evaluation of Multiple Approaches for Visual Question Reasoning. Communications in Computer and Information Science, 2018, , 208-216.	0.4	0
89	DOCK: Detecting Objects by Transferring Common-Sense Knowledge. Lecture Notes in Computer Science, 2018, , 506-522.	1.0	16
90	Factorizable Net: An Efficient Subgraph-Based Framework for Scene Graph Generation. Lecture Notes in Computer Science, 2018, , 346-363.	1.0	147

#	Article	IF	Citations
91	Visual Spatial Attention Network for Relationship Detection. , 2018, , .		19
92	Context-based sketch classification. , 2018, , .		8
93	Discovering Connotations as Labels for Weakly Supervised Image-Sentence Data. , 2018, , .		0
94	Towards Automatic Report Generation in Spine Radiology Using Weakly Supervised Framework. Lecture Notes in Computer Science, 2018, , 185-193.	1.0	18
95	Bottle Detection in the Wild Using Low-Altitude Unmanned Aerial Vehicles. , 2018, , .		22
96	Image interpretation above and below the object level. Interface Focus, 2018, 8, 20180020.	1.5	4
97	Question Answering Mediated by Visual Clues and Knowledge Graphs. , 2018, , .		1
98	Object-Based Reasoning in VQA. , 2018, , .		13
99	Visual Question Answering With a Hybrid Convolution Recurrent Model. , 2018, , .		2
100	Object sequences: encoding categorical and spatial information for a yes/no visual question answering task. IET Computer Vision, 2018, 12, 1141-1150.	1.3	2
101	Learning to Detect Human-Object Interactions. , 2018, , .		256
102	Image-Text Surgery: Efficient Concept Learning in Image Captioning by Generating Pseudopairs. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29, 5910-5921.	7.2	17
103	An image conveys a message: A brief survey on image description generation. , 2018, , .		6
104	Few-Example Object Detection with Model Communication. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41, 1641-1654.	9.7	104
105	A survey on deep neural network-based image captioning. Visual Computer, 2019, 35, 445-470.	2.5	57
106	Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. International Journal of Hospitality Management, 2019, 76, 111-121.	5.3	327
107	Survey of deep learning and architectures for visual captioningâ€"transitioning between media and natural languages. Multimedia Tools and Applications, 2019, 78, 32187-32237.	2.6	14
108	Semantic Situation Extraction from Satellite Image Based on Neural Networks. Lecture Notes in Computer Science, 2019, , 295-307.	1.0	0

#	Article	IF	Citations
109	Design Study to Prevent Mold Delamination for Overmolded Lead Frame Package. , 2019, , .		0
110	Long Activity Video Understanding Using Functional Object-Oriented Network. IEEE Transactions on Multimedia, 2019, 21, 1813-1824.	5.2	14
111	The Focus-Aspect-Value Model for Explainable Prediction of Subjective Visual Interpretation. , 2019, , .		2
112	PlanIT. ACM Transactions on Graphics, 2019, 38, 1-15.	4.9	87
113	Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. Nature Communications, 2019, 10, 3111.	5.8	65
114	BTDP. ACM Transactions on Multimedia Computing, Communications and Applications, 2019, 15, 1-21.	3.0	3
115	Multi-source Multi-level Attention Networks for Visual Question Answering. ACM Transactions on Multimedia Computing, Communications and Applications, 2019, 15, 1-20.	3.0	14
116	Search Result Reranking with Visual and Structure Information Sources. ACM Transactions on Information Systems, 2019, 37, 1-38.	3.8	4
117	Multivariate Attention Network for Image Captioning. Lecture Notes in Computer Science, 2019, , 587-602.	1.0	0
118	DRAU: Dual Recurrent Attention Units for Visual Question Answering. Computer Vision and Image Understanding, 2019, 185, 24-30.	3.0	28
119	BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, 8102-8109.	3.6	105
120	KVQA: Knowledge-Aware Visual Question Answering. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, 8876-8884.	3.6	52
121	Learning Fragment Self-Attention Embeddings for Image-Text Matching. , 2019, , .		74
122	Visual Relation Detection with Multi-Level Attention. , 2019, , .		14
123	Visual Relationship Detection with Relative Location Mining. , 2019, , .		12
124	Triple attention network for sentimental visual question answering. Computer Vision and Image Understanding, 2019, 189, 102829.	3.0	5
125	Perceptual Visual Reasoning with Knowledge Propagation. , 2019, , .		12
126	CRA-Net. , 2019, , .		33

#	Article	IF	Citations
127	Erasing-based Attention Learning for Visual Question Answering. , 2019, , .		16
128	Survey on Multi-Output Learning. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31, 1-21.	7.2	107
129	Neural Storyboard Artist. , 2019, , .		15
130	Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search. , 2019, , .		52
131	Multiple Feature Integration for Classification of Thoracic Disease in Chest Radiography. Applied Sciences (Switzerland), 2019, 9, 4130.	1.3	58
132	Who, Where, and What to Wear?. , 2019, , .		19
133	Spatial-Temporal Variation of Bacterial Communities in Sediments in Lake Chaohu, a Large, Shallow Eutrophic Lake in China. International Journal of Environmental Research and Public Health, 2019, 16, 3966.	1.2	17
134	Zero-Shot Object Detection for Indoor Robots. , 2019, , .		2
135	Hierarchical Visual Relationship Detection. , 2019, , .		6
136	Visual Relationship Recognition via Language and Position Guided Attention. , 2019, , .		5
137	Semantic Relation Detection between Construction Entities to Support Safe Human-Robot Collaboration in Construction. , 2019, , .		6
138	Large-Scale Training Framework for Video Annotation. , 2019, , .		1
139	Visual Dialog with Targeted Objects. , 2019, , .		2
140	Concept-Aware Web Image Compression Based on Crowdsourced Salient Object Detection. , 2019, , .		1
141	Few-Shot Image and Sentence Matching via Gated Visual-Semantic Embedding. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, 8489-8496.	3.6	16
142	Popup. , 2019, , .		8
143	Recurrent convolutional video captioning with global and local attention. Neurocomputing, 2019, 370, 118-127.	3.5	23
144	Quantifying and Alleviating the Language Prior Problem in Visual Question Answering. , 2019, , .		24

#	Article	IF	Citations
145	Deep Image Captioning: An Overview. , 2019, , .		19
146	Serving deep neural networks at the cloud edge for vision applications on mobile platforms. , 2019, , .		20
147	Annotating Objects and Relations in User-Generated Videos. , 2019, , .		78
148	Adversarial Adaptation of Scene Graph Models for Understanding Civic Issues. , 2019, , .		7
149	Large-Scale Visual Relationship Understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, 9185-9194.	3.6	72
150	V3C1 Dataset. , 2019, , .		45
151	Towards Automated Ship Detection and Category Recognition from High-Resolution Aerial Images. Remote Sensing, 2019, 11, 1901.	1.8	32
152	Improving visual question answering using dropout and enhanced question encoder. Pattern Recognition, 2019, 90, 404-414.	5.1	30
153	High-Quality Image Captioning With Fine-Grained and Semantic-Guided Visual Attention. IEEE Transactions on Multimedia, 2019, 21, 1681-1693.	5.2	50
154	Attention Residual Learning for Skin Lesion Classification. IEEE Transactions on Medical Imaging, 2019, 38, 2092-2103.	5.4	362
155	A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys, 2019, 51, 1-36.	16.1	440
156	Interpretable Visual Question Answering by Visual Grounding From Attention Supervision Mining. , 2019, , .		36
157	Relationship Detection Based on Object Semantic Inference and Attention Mechanisms. , 2019, , .		4
158	Eevee. , 2019, , .		1
159	Towards Safe Weakly Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43, 1-1.	9.7	52
160	Automatic Scene Recognition Based on Constructed Knowledge Space Learning. IEEE Access, 2019, 7, 102902-102910.	2.6	3
161	Creating Training Data for Scientific Named Entity Recognition with Minimal Human Effort. Lecture Notes in Computer Science, 2019, , 398-411.	1.0	10
163	Exploiting hierarchical visual features for visual question answering. Neurocomputing, 2019, 351, 187-195.	3.5	8

#	Article	IF	Citations
164	A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology, 2019, 291, 781-791.	3.6	241
165	Artificial Visual Cortex and Random Search for Object Categorization. IEEE Access, 2019, 7, 54054-54072.	2.6	13
166	A Hybrid Model for Forecasting Traffic Flow: Using Layerwise Structure and Markov Transition Matrix. IEEE Access, 2019, 7, 26002-26012.	2.6	18
167	Multi-Modal Detection Fusion on a Mobile UGV for Wide-Area, Long-Range Surveillance. , 2019, , .		1
168	Soft Transfer Learning via Gradient Diagnosis for Visual Relationship Detection. , 2019, , .		10
169	Beyond Majority Voting: A Coarse-to-Fine Label Filtration for Heavily Noisy Labels. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30, 3774-3787.	7.2	8
170	MoQA â€“ A Multi-modal Question Answering Architecture. Lecture Notes in Computer Science, 2019, , 106-113.	1.0	3
171	Application of artificial intelligence algorithms in image processing. Journal of Visual Communication and Image Representation, 2019, 61, 42-49.	1.7	37
172	A hierarchical recurrent approach to predict scene graphs from a visualâ€attentionâ€oriented perspective. Computational Intelligence, 2019, 35, 496-516.	2.1	17
173	Information fusion in visual question answering: A Survey. Information Fusion, 2019, 52, 268-280.	11.7	52
174	Social media data for conservation science: A methodological overview. Biological Conservation, 2019, 233, 298-315.	1.9	269
175	Co-Attention Network With Question Type for Visual Question Answering. IEEE Access, 2019, 7, 40771-40781.	2.6	36
176	Geospatial relation captioning for high-spatial-resolution images by using an attention-based neural network. International Journal of Remote Sensing, 2019, 40, 6482-6498.	1.3	13
177	Beyond Supervised Learning: A Computer Vision Perspective. Journal of the Indian Institute of Science, 2019, 99, 177-199.	0.9	12
178	Know More Say Less: Image Captioning Based on Scene Graphs. IEEE Transactions on Multimedia, 2019, 21, 2117-2130.	5.2	116
179	Hierarchical LSTMs with Adaptive Attention for Visual Captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42, 1-1.	9.7	143
180	Natural Language Guided Visual Relationship Detection. , 2019, , .		32
181	Crowdsourcing for search engines: perspectives and challenges. International Journal of Crowd Science, 2019, 3, 49-62.	1.1	2

#	Article	IF	Citations
182	Learning Object Context for Dense Captioning. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, 8650-8657.	3.6	23
183	MUREL: Multimodal Relational Reasoning for Visual Question Answering. , 2019, , .		167
184	Hierarchical Attention Network for Image Captioning. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, 8957-8964.	3.6	62
185	Semantic Correlations Loss: Improving Model Interpretability for Multi-class Classification. , 2019, , .		1
186	Exploring Semantic Relationships for Image Captioning without Parallel Data. , 2019, , .		13
187	Thai Scene Graph Generation from Images and Applications. , 2019, , .		0
188	Object-based Image Discrimination Relationship Recognition. , 2019, , .		0
189	Learning to Generate Unambiguous Spatial Referring Expressions for Real-World Environments. , 2019, , .		7
190	3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera. , 2019, , .		115
191	Situational Fusion of Visual Representation for Visual Navigation. , 2019, , .		36
192	Visual Question Answering over Scene Graph. , 2019, , .		14
193	BRIDGE: Building Plan Repository for Image Description Generation, and Evaluation. , 2019, , .		9
194	Gaining Extra Supervision via Multi-task learning for Multi-Modal Video Question Answering. , 2019, , .		12
195	Specifying Object Attributes and Relations in Interactive Scene Generation. , 2019, , .		100
196	Zero-Shot Object Detection with Textual Descriptions. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, 8690-8697.	3.6	49
197	Visual Relationships as Functions:Enabling Few-Shot Scene Graph Prediction. , 2019, , .		13
198	Language-Agnostic Visual-Semantic Embeddings. , 2019, , .		27
199	Residual Self-Attention for Visual Question Answering. , 2019, , .		1

#	Article	IF	Citations
200	Zero-Shot Learning on Human-Object Interaction Recognition in Video. , 2019, , .		0
201	Image Generation From Layout. , 2019, , .		93
202	Visual Narrative Technology of Paintings Based on Image Objects. , 2019, , .		1
203	Grounded Video Description. , 2019, , .		105
204	Towards VQA Models That Can Read. , 2019, , .		196
205	Image Classification with Visual Relationship. , 2019, , .		0
206	GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. , 2019, , .		400
207	Layout and Context Understanding for Image Synthesis with Scene Graphs. , 2019, , .		7
208	Transfer Learning via Unsupervised Task Discovery for Visual Question Answering. , 2019, , .		7
209	SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition. , 2019, , .		11
210	Align2Ground: Weakly Supervised Phrase Grounding Guided by Image-Caption Alignment. , 2019, , .		52
211	Hierarchy Parsing for Image Captioning. , 2019, , .		131
212	Generating Diverse and Descriptive Image Captions Using Visual Paraphrases. , 2019, , .		24
213	ACMM: Aligned Cross-Modal Memory for Few-Shot Image and Sentence Matching. , 2019, , .		46
214	Towards Unsupervised Image Captioning With Shared Multimodal Embeddings. , 2019, , .		47
215	Seq-SG2SL: Inferring Semantic Layout From Scene Graph Through Sequence to Sequence Learning. , 2019, , .		6
216	Unpaired Image-to-Speech Synthesis With Multimodal Information Bottleneck. , 2019, , .		16
217	Entangled Transformer for Image Captioning. , 2019, , .		182

#	Article	IF	Citations
218	ShapeMask: Learning to Segment Novel Objects by Refining Shape Priors. , 2019, , .		75
219	Pose-Aware Multi-Level Feature Network for Human Object Interaction Detection. , 2019, , .		135
220	Language-Conditioned Graph Networks for Relational Reasoning. , 2019, , .		83
221	Unpaired Image Captioning via Scene Graph Alignments. , 2019, , .		86
222	SynthRel0: Towards a Diagnostic Dataset for Relational Representation Learning. , 2019, , .		0
223	Scene Graph Prediction with Limited Labels. , 2019, , .		43
224	Scaling Object Detection by Transferring Classification Weights. , 2019, , .		12
225	No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. , 2019, , .		62
226	Relation-Aware Graph Attention Network for Visual Question Answering. , 2019, , .		193
227	Scene Text Visual Question Answering. , 2019, , .		128
228	Counting Attention Based on Classification Confidence for Visual Question Answering. , 2019, , .		1
229	Making History Matter: History-Advantage Sequence Training for Visual Dialog. , 2019, , .		40
230	Attention on Attention for Image Captioning. , 2019, , .		502
231	Zero-Shot Grounding of Objects From Natural Language Queries. , 2019, , .		77
232	Reflective Decoding Network for Image Captioning. , 2019, , .		60
233	Deliberate Attention Networks for Image Captioning. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33, 8320-8327.	3.6	35
234	Human Uncertainty Makes Classification More Robust. , 2019, , .		63
235	Counterfactual Critic Multi-Agent Training for Scene Graph Generation. , 2019, , .		91

#	Article	IF	Citations
236	Scene Graph Generation With External Knowledge and Image Reconstruction. , 2019, , .		168
237	VEKG: Video Event Knowledge Graph to Represent Video Streams for Complex Event Pattern Matching. , 2019, , .		11
238	Learning to Collocate Neural Modules for Image Captioning. , 2019, , .		52
239	Scene Graph Prediction With Limited Labels. , 2019, 2019, 2580-2590.		12
240	ICDAR 2019 Competition on Scene Text Visual Question Answering. , 2019, , .		13
241	Compensating Supervision Incompleteness with Prior Knowledge in Semantic Image Interpretation. , 2019, , .		10
242	From Recognition to Cognition: Visual Commonsense Reasoning. , 2019, , .		316
243	Deep Metric Learning Beyond Binary Supervision. , 2019, , .		60
244	OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. , 2019, , .		160
245	Transferable Interactiveness Knowledge for Human-Object Interaction Detection. , 2019, , .		151
246	Attentive Relational Networks for Mapping Images to Scene Graphs. , 2019, , .		91
247	Cross-Modal Relationship Inference for Grounding Referring Expressions. , 2019, , .		69
248	Neural Sequential Phrase Grounding (SeqGROUND). , 2019, , .		23
249	Knowledge-Embedded Routing Network for Scene Graph Generation. , 2019, , .		243
250	Visual Question Answering as Reading Comprehension. , 2019, , .		23
251	On Exploring Undetermined Relationships for Visual Relationship Detection. , 2019, , .		57
252	Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-To-Image Translation. , 2019, , .		48
253	Context and Attribute Grounded Dense Captioning. , 2019, , .		30

#	Article	IF	Citations
254	Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Captioning. , 2019, , .		50
255	Adversarial Inference for Multi-Sentence Video Description. , 2019, , .		50
256	Learning to Compose Dynamic Tree Structures for Visual Contexts. , 2019, , .		251
257	Cycle-Consistency for Robust Visual Question Answering. , 2019, , .		77
258	Exploring Context and Visual Pattern of Relationship for Scene Graph Generation. , 2019, , .		60
259	Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. , 2019, , .		112
260	Visual Query Answering by Entity-Attribute Graph Matching and Reasoning. , 2019, , .		12
261	Look Back and Predict Forward in Image Captioning. , 2019, , .		91
262	Intention Oriented Image Captions With Guiding Objects. , 2019, , .		36
263	Long-Term Feature Banks for Detailed Video Understanding. , 2019, , .		245
264	Information Maximizing Visual Question Generation. , 2019, , .		52
265	Not All Frames Are Equal: Weakly-Supervised Video Grounding With Contextual Similarity and Visual Clustering Losses. , 2019, , .		26
266	Auto-Encoding Scene Graphs for Image Captioning. , 2019, , .		459
267	Region Based Anomaly Detection with Real-Time Training and Analysis. , 2019, , .		1
268	An End-To-End Network for Generating Social Relationship Graphs. , 2019, , .		24
269	Graphical Contrastive Losses for Scene Graph Parsing. , 2019, , .		125
270	Multi-Level Multimodal Common Semantic Space for Image-Phrase Grounding. , 2019, , .		28
271	Answer Them All! Toward Universal Visual Question Answering Models. , 2019, , .		56

#	Article	IF	Citations
272	Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. , 2019, , .		182
273	Visual Semantic Reasoning for Image-Text Matching. , 2019, , .		292
274	Phrase Localization Without Paired Training Examples. , 2019, , .		19
275	Image Synthesis From Reconfigurable Layout and Style. , 2019, , .		61
276	A Denoising Framework for Image Caption. , 2019, , .		1
277	On Class Imbalance and Background Filtering in Visual Relationship Detection. , 2019, , .		2
278	Cross-Modal Image-Text Retrieval with Semantic Consistency. , 2019, , .		46
279	Bootstrapping Knowledge Graphs From Images and Text. Frontiers in Neurorobotics, 2019, 13, 93.	1.6	4
280	Movienet: a movie multilayer network model using visual and textual semantic cues. Applied Network Science, 2019, 4, .	0.8	7
281	Recursive Visual Attention in Visual Dialog. , 2019, , .		71
282	Scene graph captioner: Image captioning based on structural visual representation. Journal of Visual Communication and Image Representation, 2019, 58, 477-485.	1.7	59
283	A multimodal fusion approach for image captioning. Neurocomputing, 2019, 329, 476-485.	3.5	38
284	From BoW to CNN: Two Decades of Texture Representation for Texture Classification. International Journal of Computer Vision, 2019, 127, 74-109.	10.9	247
285	Focal Visual-Text Attention for Memex Question Answering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41, 1893-1908.	9.7	49
287	COSMO: Contextualized scene modeling with Boltzmann Machines. Robotics and Autonomous Systems, 2019, 113, 132-148.	3.0	10
288	Deep Learning From Noisy Image Labels With Quality Embedding. IEEE Transactions on Image Processing, 2019, 28, 1909-1922.	6.0	58
289	Word-to-region attention network for visual question answering. Multimedia Tools and Applications, 2019, 78, 3843-3858.	2.6	24
290	Team NimbRo at MBZIRC 2017: Autonomous valve stem turning using a wrench. Journal of Field Robotics, 2019, 36, 170-182.	3.2	10

#	Article	IF	Citations
291	Hierarchical Scene Parsing by Weakly Supervised Learning with Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41, 596-610.	9.7	16
292	Biometric surveillance using visual question answering. Pattern Recognition Letters, 2019, 126, 111-118.	2.6	15
293	Image and Sentence Matching via Semantic Concepts and Order Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42, 636-650.	9.7	24
294	Hierarchical Deep Neural Network for Image Captioning. Neural Processing Letters, 2020, 52, 1057-1067.	2.0	14
295	Visual relationship detection based on bidirectional recurrent neural network. Multimedia Tools and Applications, 2020, 79, 35297-35313.	2.6	8
296	Context-Aware Visual Policy Network for Fine-Grained Image Captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 710-722.	9.7	72
297	Learning visual features for relational CBIR. International Journal of Multimedia Information Retrieval, 2020, 9, 113-124.	3.6	9
298	Learning visual relationship and context-aware attention for image captioning. Pattern Recognition, 2020, 98, 107075.	5.1	98
299	Multimodal feature fusion by relational reasoning and attention for visual question answering. Information Fusion, 2020, 55, 116-126.	11.7	49
300	Recall What You See Continually Using GridLSTM in Image Captioning. IEEE Transactions on Multimedia, 2020, 22, 808-818.	5.2	31
301	3-D Scene Graph: A Sparse and Semantic Representation of Physical Environments for Intelligent Agents. IEEE Transactions on Cybernetics, 2020, 50, 4921-4933.	6.2	37
302	Text-based indoor place recognition with deep neural network. Neurocomputing, 2020, 390, 239-247.	3.5	5
303	Compact Wideband Wide-Angle Polarization-Free Metasurface Lens Antenna Array for Multibeam Base Stations. IEEE Transactions on Antennas and Propagation, 2020, 68, 1378-1388.	3.1	44
304	Multimodal Transformer With Multi-View Visual Representation for Image Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30, 4467-4480.	5.6	258
305	Text mining and deep learning for disease classification. , 2020, , 109-135.		1
306	INGRESS: Interactive visual grounding of referring expressions. International Journal of Robotics Research, 2020, 39, 217-232.	5.8	46
307	The Focusâ€“Aspectâ€“Value model for predicting subjective visual attributes. International Journal of Multimedia Information Retrieval, 2020, 9, 47-60.	3.6	1
308	Cross-scale fusion detection with global attribute for dense captioning. Neurocomputing, 2020, 373, 98-108.	3.5	4

#	Article	IF	Citations
309	Knowledge Fusion via Joint Tensor and Matrix Factorization. Cognitive Computation, 2020, 12, 642-653.	3.6	2
310	Multi-Modal fusion with multi-level attention for Visual Dialog. Information Processing and Management, 2020, 57, 102152.	5.4	11
311	SHOP-VRB: A Visual Reasoning Benchmark for Object Perception. , 2020, , .		8
312	Relation R-CNN: A Graph Based Relation-Aware Network for Object Detection. IEEE Signal Processing Letters, 2020, 27, 1680-1684.	2.1	47
313	Symbiotic Attention for Egocentric Action Recognition With Object-Centric Alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 6605-6617.	9.7	50
314	Visual Relational Reasoning for Image Caption. , 2020, , .		3
315	Richpedia: A Large-Scale, Comprehensive Multi-Modal Knowledge Graph. Big Data Research, 2020, 22, 100159.	2.6	42
316	Visual manipulation relationship recognition in object-stacking scenes. Pattern Recognition Letters, 2020, 140, 34-42.	2.6	12
317	Dual Semantic Relationship Attention Network for Image-Text Matching. , 2020, , .		1
318	Learning Multimodal Representations by Symmetrically Transferring Local Structures. Symmetry, 2020, 12, 1504.	1.1	0
319	Visual Question Answering on 360Â° Images. , 2020, , .		8
320	Joint Commonsense and Relation Reasoning for Image and Video Captioning. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34, 10973-10980.	3.6	27
321	Classifying All Interacting Pairs in a Single Shot. , 2020, , .		2
322	Graphical Semantic Authentication. , 2020, , .		1
323	Self-Supervised Feature Augmentation for Large Image Object Detection. IEEE Transactions on Image Processing, 2020, 29, 6745-6758.	6.0	25
324	PFAN++: Bi-Directional Image-Text Retrieval With Position Focused Attention Network. IEEE Transactions on Multimedia, 2021, 23, 3362-3376.	5.2	15
325	Multi-Modality Global Fusion Attention Network for Visual Question Answering. Electronics (Switzerland), 2020, 9, 1882.	1.8	4
326	ViSeR: Visual Self-Regularization. , 2020, , .		0

#	Article	IF	Citations
327	CPARR: Category-based Proposal Analysis for Referring Relationships. , 2020, , .		3
328	Quality and Relevance Metrics for Selection of Multimodal Pretraining Data. , 2020, , .		1
329	Exploring Phrase Grounding without Training: Contextualisation and Extension to Text-Based Image Retrieval. , 2020, , .		4
330	Semantics-Preserving Graph Propagation for Zero-Shot Object Detection. IEEE Transactions on Image Processing, 2020, 29, 8163-8176.	6.0	45
331	PPDM: Parallel Point Detection and Matching for Real-Time Human-Object Interaction Detection. , 2020, , .		141
332	Graph Structured Network for Image-Text Matching. , 2020, , .		132
333	On Diversity in Image Captioning: Metrics and Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 1035-1049.	9.7	15
334	Zero-Shot Object Detection: Joint Recognition and Localization of Novel Concepts. International Journal of Computer Vision, 2020, 128, 2979-2999.	10.9	34
335	Deep Bayesian Network for Visual Question Generation. , 2020, , .		6
336	Investigating Class-Level Difficulty Factors In Multi-Label Classification Problems. , 2020, , .		1
337	Spatial-Content Image Search in Complex Scenes. , 2020, , .		10
338	AACR: Feature Fusion Effects of Algebraic Amalgamation Composed Representation on (De)Compositional Network for Caption Generation for Images. SN Computer Science, 2020, 1, 1.	2.3	6
339	Explaining VQA predictions using visual grounding and a knowledge base. Image and Vision Computing, 2020, 101, 103968.	2.7	12
340	Show, Recall, and Tell: Image Captioning with Recall Mechanism. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34, 12176-12183.	3.6	31
341	Unified Vision-Language Pre-Training for Image Captioning and VQA. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34, 13041-13049.	3.6	361
342	Multi-task Compositional Network for Visual Relationship Detection. International Journal of Computer Vision, 2020, 128, 2146-2165.	10.9	22
343	Adaptive Cross-Modal Embeddings for Image-Text Alignment. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34, 12313-12320.	3.6	39
344	X-Linear Attention Networks for Image Captioning. , 2020, , .		314

#	Article	IF	Citations
345	Machine Learning Paradigms. Learning and Analytics in Intelligent Systems, 2020, , .	0.5	8
346	Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval. , 2020, , .		123
347	Answering Questions about Data Visualizations using Efficient Bimodal Fusion. , 2020, , .		21
348	Cross-modal knowledge reasoning for knowledge-based visual question answering. Pattern Recognition, 2020, 108, 107563.	5.1	54
349	Personalizing Fast-Forward Videos Based on Visual and Textual Features from Social Network. , 2020, , .		2
350	Unbiased Scene Graph Generation From Biased Training. , 2020, , .		322
351	GPS-Net: Graph Property Sensing Network for Scene Graph Generation. , 2020, , .		132
352	Learning 3D Semantic Scene Graphs From 3D Indoor Reconstructions. , 2020, , .		73
353	Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector. , 2020, , .		275
354	More Grounded Image Captioning by Distilling Image-Text Matching Model. , 2020, , .		73
355	Image Hallucination From Attribute Pairs. IEEE Transactions on Cybernetics, 2022, 52, 568-581.	6.2	6
356	Robotic Understanding of Object Semantics by Referringto a Dictionary. International Journal of Social Robotics, 2020, 12, 1251-1263.	3.1	5
357	An end-to-end graph convolutional kernel support vector machine. Applied Network Science, 2020, 5, .	0.8	1
358	Context-Aware Group Captioning via Self-Attention and Contrastive Features. , 2020, , .		22
359	Context-Aware Attention Network for Image-Text Retrieval. , 2020, , .		129
360	Weakly Supervised Visual Semantic Parsing. , 2020, , .		35
361	Temporal Reasoning via Audio Question Answering. IEEE/ACM Transactions on Audio Speech and Language Processing, 2020, 28, 2283-2294.	4.0	10
362	Modality Shifting Attention Network for Multi-Modal Video Question Answering. , 2020, , .		46

#	Article	IF	Citations
363	Visual Question Answering With Dense Inter- and Intra-Modality Interactions. IEEE Transactions on Multimedia, 2021, 23, 3518-3529.	5.2	20
364	Component Analysis for Visual Question Answering Architectures. , 2020, , .		1
365	Improving Deep Learning Approaches for Human Activity Recognition based on Natural Language Processing of Action Labels. , 2020, , .		7
366	One-Shot Texture Retrieval Using Global Grouping Metric. IEEE Transactions on Multimedia, 2020, , 1-1.	5.2	4
367	Learn and Tell: Learning Priors for Image Caption Generation. Applied Sciences (Switzerland), 2020, 10, 6942.	1.3	1
368	Ontological Approach to Image Captioning Evaluation. Pattern Recognition and Image Analysis, 2020, 30, 288-294.	0.6	1
369	Unsupervised Visualâ€“Textual Correlation Learning With Fine-Grained Semantic Alignment. IEEE Transactions on Cybernetics, 2022, 52, 3669-3683.	6.2	13
370	A Sparse Transformer-Based Approach for Image Captioning. IEEE Access, 2020, 8, 213437-213446.	2.6	2
371	TSM: Topological Scene Map for Representation in Indoor Environment Understanding. IEEE Access, 2020, 8, 185870-185884.	2.6	4
372	Hybrid Attention Distribution and Factorized Embedding Matrix in Image Captioning. IEEE Access, 2020, 8, 154453-154460.	2.6	1
373	SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability. , 2020, , .		19
374	Federated Learning for Vision-and-Language Grounding Problems. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34, 11572-11579.	3.6	33
375	DrunaliaCap: Image Captioning for Drug-Related Paraphernalia With Deep Learning. IEEE Access, 2020, 8, 161326-161336.	2.6	3
376	An Entropy Clustering Approach for Assessing Visual Question Difficulty. IEEE Access, 2020, 8, 180633-180645.	2.6	2
377	Detecting Human-Object Interactions via Functional Generalization. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34, 10460-10469.	3.6	56
378	Visual-Semantic Matching by Exploring High-Order Attention and Distraction. , 2020, , .		24
379	Hierarchical Graph Attention Network for Visual Relationship Detection. , 2020, , .		45
380	Hypergraph Attention Networks for Multimodal Learning. , 2020, , .		33

#	Article	IF	Citations
381	SPARE3D: A Dataset for SPAtial REasoning on Three-View Line Drawings. , 2020, , .		4
382	Semantic Tree-Based 3D Scene Model Recognition. , 2020, , .		1
383	Interactive Dual Generative Adversarial Networks for Image Captioning. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34, 11588-11595.	3.6	20
384	Illuminating the dark spaces of healthcare with ambient intelligence. Nature, 2020, 585, 193-202.	13.7	139
385	An Effective Dense Co-Attention Networks for Visual Question Answering. Sensors, 2020, 20, 4897.	2.1	12
386	Multimodal Machine Learning for Natural Language Processing: Disambiguating Prepositional Phrase Attachments with Images. Neural Processing Letters, 2021, 53, 3095-3121.	2.0	4
387	Domain Adaptation in Computer Vision with Deep Learning. , 2020, , .		7
388	Semantic Image Manipulation Using Scene Graphs. , 2020, , .		52
389	Deep Convolutional Networks for Construction Object Detection Under Different Visual Conditions. Frontiers in Built Environment, 2020, 6, .	1.2	34
390	Affinity Graph Supervision for Visual Recognition. , 2020, , .		4
391	ActBERT: Learning Global-Local Video-Text Representations. , 2020, , .		203
392	Graph-Structured Referring Expression Reasoning in the Wild. , 2020, , .		52
393	Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs. , 2020, , .		134
394	Iterative Answer Prediction With Pointer-Augmented Multimodal Transformers for TextVQA. , 2020, , .		106
395	TA-Student VQA: Multi-Agents Training by Self-Questioning. , 2020, , .		0
396	Cops-Ref: A New Dataset and Task on Compositional Referring Expression Comprehension. , 2020, , .		26
397	On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering. , 2020, , .		42
398	Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs. , 2020, , .		150

#	Article	IF	Citations
399	In Defense of Grid Features for Visual Question Answering. , 2020, , .		180
400	VQA With No Questions-Answers Training. , 2020, , .		8
401	Video Object Grounding Using Semantic Roles in Language Description. , 2020, , .		25
402	12-in-1: Multi-Task Vision and Language Representation Learning. , 2020, , .		207
403	Meshed-Memory Transformer for Image Captioning. , 2020, , .		496
404	Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. , 2020, , .		43
405	Spatial Cognition XII. Lecture Notes in Computer Science, 2020, , .	1.0	2
406	Two Causal Principles for Improving Visual Dialog. , 2020, , .		64
407	Spatio-Temporal Graph for Video Captioning With Knowledge Distillation. , 2020, , .		139
408	A Real-Time Cross-Modality Correlation Filtering Method for Referring Expression Comprehension. , 2020, , .		99
409	Violin: A Large-Scale Dataset for Video-and-Language Inference. , 2020, , .		27
410	Multi-Modality Cross Attention Network for Image and Sentence Matching. , 2020, , .		164
411	Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap. IEEE Access, 2020, 8, 218386-218400.	2.6	46
412	Multi-Modal Explicit Sparse Attention Networks for Visual Question Answering. Sensors, 2020, 20, 6758.	2.1	10
413	Scene Recognition Based on Recurrent Memorized Attention Network. Electronics (Switzerland), 2020, 9, 2038.	1.8	5
414	Skin Lesion Classification Using Densely Connected Convolutional Networks with Attention Residual Learning. Sensors, 2020, 20, 7080.	2.1	20
415	Multi-Label Remote Sensing Image Scene Classification by Combining a Convolutional Neural Network and a Graph Neural Network. Remote Sensing, 2020, 12, 4003.	1.8	48
416	On machine vision and photographic imagination. AI and Society, 2021, 36, 1153-1165.	3.1	4

#	Article	IF	Citations
417	Attentional Colorization Networks with Adaptive Group-Instance Normalization. Information (Switzerland), 2020, 11, 479.	1.7	2
418	Revisiting Image-Language Networks for Open-Ended Phrase Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 2155-2167.	9.7	9
419	Modeling coherence and diversity for image paragraph captioning. , 2020, , .		1
420	Learning to Recognize Visual Concepts for Visual Question Answering With Structural Label Space. IEEE Journal on Selected Topics in Signal Processing, 2020, 14, 494-505.	7.3	8
421	Exploring Entity-Level Spatial Relationships for Image-Text Matching. , 2020, , .		2
422	Attentive Gated Graph Neural Network for Image Scene Graph Generation. Symmetry, 2020, 12, 511.	1.1	3
423	RepeatPadding: Balancing words and sentence length for language comprehension in visual question answering. Information Sciences, 2020, 529, 166-178.	4.0	5
424	Local relation network with multilevel attention for visual question answering. Journal of Visual Communication and Image Representation, 2020, 73, 102762.	1.7	6
425	Multi-Layer Content Interaction Through Quaternion Product for Visual Question Answering. , 2020, , .		10
426	Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders. Nature Biomedical Engineering, 2020, 4, 767-777.	11.6	42
427	Generating Images from Arabic Story-Text using Scene Graph. , 2020, , .		3
428	Textual-Visual Reference-Aware Attention Network for Visual Dialog. IEEE Transactions on Image Processing, 2020, 29, 6655-6666.	6.0	16
429	Cross-modal learning with prior visual relation knowledge. Knowledge-Based Systems, 2020, 203, 106150.	4.0	7
430	A Comprehensive Pipeline for Complex Text-to-Image Synthesis. Journal of Computer Science and Technology, 2020, 35, 522-537.	0.9	10
431	The Open Images Dataset V4. International Journal of Computer Vision, 2020, 128, 1956-1981.	10.9	772
432	Visualâ€"Textual Hybrid Sequence Matching for Joint Reasoning. IEEE Transactions on Cybernetics, 2021, 51, 5692-5705.	6.2	7
433	Visual Object Search by Learning Spatial Context. IEEE Robotics and Automation Letters, 2020, 5, 1279-1286.	3.3	36
434	Task-Agnostic Object Recognition for Mobile Robots through Few-Shot Image Matching. Electronics (Switzerland), 2020, 9, 380.	1.8	7

#	Article	IF	Citations
435	Unsupervised Cross-Media Retrieval Using Domain Adaptation With Scene Graph. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30, 4368-4379.	5.6	38
436	Stacked squeeze-and-excitation recurrent residual network for visual-semantic matching. Pattern Recognition, 2020, 105, 107359.	5.1	11
437	MRA-Net: Improving VQA Via Multi-Modal Relation Attention Network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 318-329.	9.7	51
438	Interactive Natural Language Grounding via Referring Expression Comprehension and Scene Graph Parsing. Frontiers in Neurorobotics, 2020, 14, 43.	1.6	7
439	Constructing Geospatial Concept Graphs from Tagged Images for Geo-Aware Fine-Grained Image Recognition. ISPRS International Journal of Geo-Information, 2020, 9, 354.	1.4	4
440	Comparison of state-of-the-art deep learning APIs for image multi-label classification using semantic metrics. Expert Systems With Applications, 2020, 161, 113656.	4.4	13
441	Visually grounded paraphrase identification via gating and phrase localization. Neurocomputing, 2020, 404, 165-172.	3.5	3
442	Answer Again: Improving VQA with Cascaded-Answering Model. IEEE Transactions on Knowledge and Data Engineering, 2020, , 1-1.	4.0	9
443	Improving visual relationship detection using linguistic and spatial cues. ETRI Journal, 2020, 42, 399-410.	1.2	4
444	A Text-Guided Graph Structure for Image Captioning. , 2020, , .		1
445	Layout2image: Image Generation from Layout. International Journal of Computer Vision, 2020, 128, 2418-2435.	10.9	14
446	Multi-Modal Memory Enhancement Attention Network for Image-Text Matching. IEEE Access, 2020, 8, 38438-38447.	2.6	10
447	Relationship-Embedded Representation Learning for Grounding Referring Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43, 2765-2779.	9.7	33
448	Cross-Modal Attention With Semantic Consistence for Imageâ€“Text Matching. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31, 5412-5425.	7.2	148
449	Dual-CNN: A Convolutional language decoder for paragraph image captioning. Neurocomputing, 2020, 396, 92-101.	3.5	29
450	Multimodal Encoder-Decoder Attention Networks for Visual Question Answering. IEEE Access, 2020, 8, 35662-35671.	2.6	37
451	Object detection in optical remote sensing images by integrating object-to-object relationships. Remote Sensing Letters, 2020, 11, 416-425.	0.6	7
452	VATAS: An Open-Source Web Platform for Visual and Textual Analysis of Social Media. Journal of the Society for Social Work and Research, 2020, 11, 133-155.	0.9	5

#	Article	IF	Citations
453	Context-based information generation for managing UAV-acquired data using image captioning. Automation in Construction, 2020, 112, 103116.	4.8	40
455	Cascaded Revision Network for Novel Object Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30, 3413-3421.	5.6	23
456	Interactive Natural Language-Based Person Search. IEEE Robotics and Automation Letters, 2020, 5, 1851-1858.	3.3	4
457	Show, Tell, and Polish: Ruminant Decoding for Image Captioning. IEEE Transactions on Multimedia, 2020, 22, 2149-2162.	5.2	35
458	Handwritten Mathematical Expression Recognition via Paired Adversarial Learning. International Journal of Computer Vision, 2020, 128, 2386-2401.	10.9	57
459	MAVA: Multi-Level Adaptive Visual-Textual Alignment by Cross-Media Bi-Attention Mechanism. IEEE Transactions on Image Processing, 2020, 29, 2728-2741.	6.0	29
460	Utilizing the platform economy effect through EWOM: Does the platform matter?. International Journal of Production Economics, 2020, 227, 107663.	5.1	31
461	Revisiting EmbodiedQA: A Simple Baseline and Beyond. IEEE Transactions on Image Processing, 2020, 29, 3984-3992.	6.0	20
462	Panoptic Segmentation-Based Attention for Image Captioning. Applied Sciences (Switzerland), 2020, 10, 391.	1.3	3
463	Domain-Specific Image Caption Generator with Semantic Ontology. , 2020, , .		11
464	What and where: A context-based recommendation system for object insertion. Computational Visual Media, 2020, 6, 79-93.	10.8	10
465	Visual question answering: a state-of-the-art review. Artificial Intelligence Review, 2020, 53, 5705-5745.	9.7	30
466	Group Activity Recognition by Using Effective Multiple Modality Relation Representation With Temporal-Spatial Attention. IEEE Access, 2020, 8, 65689-65698.	2.6	12
467	On Visual-Textual-Knowledge Entity Linking. , 2020, , .		2
468	Selective residual learning for Visual Question Answering. Neurocomputing, 2020, 402, 366-374.	3.5	15
469	Dense-CaptionNet: a Sentence Generation Architecture for Fine-grained Description of Image Semantics. Cognitive Computation, 2021, 13, 595.	3.6	9
470	Integrating Part of Speech Guidance for Image Captioning. IEEE Transactions on Multimedia, 2021, 23, 92-104.	5.2	30
471	Multi-Level Knowledge Injecting for Visual Commonsense Reasoning. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31, 1042-1054.	5.6	14

#	Article	IF	Citations
472	IncDet: In Defense of Elastic Weight Consolidation for Incremental Object Detection. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32, 2306-2319.	7.2	21
473	Improving Visual Relationship Detection With Two-Stage Correlation Exploitation. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31, 2751-2763.	5.6	7
474	Explaining the semantics capturing capability of scene graph generation models. Pattern Recognition, 2021, 110, 107427.	5.1	11
475	Place perception from the fusion of different image representation. Pattern Recognition, 2021, 110, 107680.	5.1	5
476	Multimodal deep fusion for image question answering. Knowledge-Based Systems, 2021, 212, 106639.	4.0	18
477	Learning to transfer focus of graph neural network for scene graph parsing. Pattern Recognition, 2021, 112, 107707.	5.1	18
478	Multi-level similarity learning for image-text retrieval. Information Processing and Management, 2021, 58, 102432.	5.4	23
479	Hier R-CNN: Instance-Level Human Parts Detection and A New Benchmark. IEEE Transactions on Image Processing, 2021, 30, 39-54.	6.0	31
480	Deep Relation Embedding for Cross-Modal Retrieval. IEEE Transactions on Image Processing, 2021, 30, 617-627.	6.0	25
481	Structural fragmentation in scene graphs. Knowledge-Based Systems, 2021, 211, 106504.	4.0	1
482	MGRL: Graph neural network based inference in a Markov network with reinforcement learning for visual navigation. Neurocomputing, 2021, 421, 140-150.	3.5	17
483	CRUR: coupled-recurrent unit for unification, conceptualization and context capture for language representation - a generalization of bi directional LSTM. Multimedia Tools and Applications, 2021, 80, 9917-9959.	2.6	3
484	Unifying neural learning and symbolic reasoning for spinal medical report generation. Medical Image Analysis, 2021, 67, 101872.	7.0	22
485	Learning Dual Semantic Relations With Graph Attention for Image-Text Matching. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31, 2866-2879.	5.6	48
486	Fine-Grained Image Captioning With Global-Local Discriminative Objective. IEEE Transactions on Multimedia, 2021, 23, 2413-2427.	5.2	45
487	Modelling relations with prototypes for visual relation detection. Multimedia Tools and Applications, 2021, 80, 22465-22486.	2.6	0
488	Interpreting the Rhetoric of Visual Advertisements. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43, 1308-1323.	9.7	14
489	Interpretable Visual Question Answering by Reasoning on Dependency Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43, 887-901.	9.7	30

#	Article	IF	Citations
490	Visual Semantic Information Pursuit: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43, 1404-1422.	9.7	16
491	Visual Scanpath Prediction Using IOR-ROI Recurrent Mixture Density Network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43, 2101-2118.	9.7	24
492	Towards Open Ended and Free Form Visual Question Answering: Modeling VQA as a Factoid Question Answering Problem. Lecture Notes in Networks and Systems, 2021, , 749-759.	0.5	0
493	VSR++: Improving Visual Semantic Reasoning for Fine-Grained Image-Text Matching. , 2021, , .		5
494	Visual Relationship Detection With Visual-Linguistic Knowledge From Multimodal Representations. IEEE Access, 2021, 9, 50441-50451.	2.6	15
495	Improving Visual Relation Detection using Depth Maps. , 2021, , .		7
496	Background Learnable Cascade for Zero-Shot Object Detection. Lecture Notes in Computer Science, 2021, , 107-123.	1.0	11
497	IQ-VQA: Intelligent Visual Question Answering. Lecture Notes in Computer Science, 2021, , 357-370.	1.0	2
498	Attention in Reasoning: Dataset, Analysis, and Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 7310-7326.	9.7	3
499	Synthetic Data for Deep Learning. Springer Optimization and Its Applications, 2021, , .	0.6	98
500	CSKG: The CommonSense Knowledge Graph. Lecture Notes in Computer Science, 2021, , 680-696.	1.0	17
501	Grounding Plural Phrases: Countering Evaluation Biases by Individuation. , 2021, , .		0
502	Multi-Modal Image Captioning for the Visually Impaired. , 2021, , .		8
503	Neural Metaphor Detection with Visibility Embeddings. , 2021, , .		2
504	Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering. , 2021, , .		26
505	Person-in-Context Synthesis with Compositional Structural Space. , 2021, , .		1
506	Dual Path Multi-Modal High-Order Features for Textual Content based Visual Question Answering. , 2021, , .		0
507	Force Banner for the recognition of spatial relations. , 2021, , .		1

#	Article	IF	Citations
508	FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding. IEEE Access, 2021, 9, 89644-89654.	2.6	85
509	Context for Object Detection via Lightweight Global and Mid-level Representations. , 2021, , .		0
510	Task-Adaptive Attention for Image Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 43-51.	5.6	146
511	Probing Spatial Clues: Canonical Spatial Templates for Object Relationship Understanding. IEEE Access, 2021, 9, 134298-134318.	2.6	2
512	Hierarchical Reasoning Network for Human-Object Interaction Detection. IEEE Transactions on Image Processing, 2021, 30, 8306-8317.	6.0	9
513	Attention Guided Relation Detection Approach for Video Visual Relation Detection. IEEE Transactions on Multimedia, 2022, 24, 3896-3907.	5.2	2
514	Learning Layout and Style Reconfigurable GANs for Controllable Image Synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, PP, 1-1.	9.7	21
515	Unsupervised Vision-and-Language Pre-training Without Parallel Images and Captions. , 2021, , .		11
516	Suppressing Biased Samples for Robust VQA. IEEE Transactions on Multimedia, 2022, 24, 3405-3415.	5.2	11
517	Plot and Rework: Modeling Storylines for Visual Storytelling. , 2021, , .		7
518	Spatial-Temporal Graphs for Cross-Modal Text2Video Retrieval. IEEE Transactions on Multimedia, 2022, 24, 2914-2923.	5.2	36
519	End-to-End Video Question-Answer Generation With Generator-Pretester Network. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31, 4497-4507.	5.6	15
520	Modular Graph Attention Network for Complex Visual Relational Reasoning. Lecture Notes in Computer Science, 2021, , 137-153.	1.0	1
521	Adaptive Spatio-Temporal Graph Enhanced Vision-Language Representation for Video QA. IEEE Transactions on Image Processing, 2021, 30, 5477-5489.	6.0	12
522	Image Captioning Through Image Transformer. Lecture Notes in Computer Science, 2021, , 153-169.	1.0	22
523	Improving Visual Reasoning Through Semantic Representation. IEEE Access, 2021, 9, 91476-91486.	2.6	121
524	Using Scene Graphs for Detecting Visual Relationships. , 2021, , .		0
525	Integrating Historical States and Co-attention Mechanism for Visual Dialog. , 2021, , .		1

#	Article	IF	Citations
526	A Review on Explainability in Multimodal Deep Neural Nets. IEEE Access, 2021, 9, 59800-59821.	2.6	78
527	Adaptive Word Embedding Module for Semantic Reasoning in Large-scale Detection. , 2021, , .		1
528	Integrating Multihub Driven Attention Mechanism and Big Data Analytics for Virtual Representation of Visual Scenes. IEEE Transactions on Industrial Informatics, 2022, 18, 1435-1444.	7.2	5
529	A Hybrid Approach for Semantic Image Annotation. IEEE Access, 2021, 9, 131977-131994.	2.6	0
530	VSRN: Visual-Semantic Relation Network for Video Visual Relation Inference. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 768-777.	5.6	3
531	Human-centric Relation Segmentation: Dataset and Solution. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, PP, 1-1.	9.7	1
532	Cross-modal multi-relationship aware reasoning for image-text matching. Multimedia Tools and Applications, 2022, 81, 12005-12027.	2.6	4
533	A Systematic Survey of ML Datasets for Prime CV Research Areasâ€"Media and Metadata. Data, 2021, 6, 12.	1.2	1
534	Second Order Enhanced Multi-glimpse Attention in Visual Question Answering. Lecture Notes in Computer Science, 2021, , 87-103.	1.0	0
535	Message-Passing-Driven Triplet Representation for Geo-Object Relational Inference in HRSI. IEEE Geoscience and Remote Sensing Letters, 2022, 19, 1-5.	1.4	0
536	Knowledge-Routed Visual Question Reasoning: Challenges for Deep Representation Embedding. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33, 2758-2767.	7.2	7
537	DeCEMBERT: Learning from Noisy Instructional Videos via Dense Captions and Entropy Minimization. , 2021, , .		18
538	Cross-lingual Cross-modal Pretraining for Multimodal Retrieval. , 2021, , .		14
539	IPGN: Interactiveness Proposal Graph Network for Human-Object Interaction Detection. IEEE Transactions on Image Processing, 2021, 30, 6583-6593.	6.0	19
540	Multi-modal Contextual Graph Neural Network for Text Visual Question Answering. , 2021, , .		3
541	Answer-checking in Context: A Multi-modal Fully Attention Network for Visual Question Answering. , 2021, , .		3
542	Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers. Transactions of the Association for Computational Linguistics, 2021, 9, 570-585.	3.2	30
543	KM-BART: Knowledge Enhanced Multimodal BART for Visual Commonsense Generation. , 2021, , .		11

#	Article	IF	Citations
544	Exploring Pairwise Relationships Adaptively From Linguistic Context in Image Captioning. IEEE Transactions on Multimedia, 2022, 24, 3101-3113.	5.2	13
545	Mutual Attention Inception Network for Remote Sensing Visual Question Answering. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-14.	2.7	48
546	ACP++: Action Co-Occurrence Priors for Human-Object Interaction Detection. IEEE Transactions on Image Processing, 2021, 30, 9150-9163.	6.0	10
547	Difficulty-Controllable Visual Question Generation. Lecture Notes in Computer Science, 2021, , 332-347.	1.0	4
548	Relation-Aware Reasoning with Graph Convolutional Network. Lecture Notes in Computer Science, 2021, , 52-64.	1.0	0
549	Multiscale Conditional Relationship Graph Network for Referring Relationships in Images. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14, 752-760.	2.6	5
550	Validity-Based Sampling and Smoothing Methods for Multiple Reference Image Captioning. , 2021, , .		0
551	Probing Contextual Language Models for Common Ground with Visual Representations. , 2021, , .		3
552	OCID-Ref: A 3D Robotic Dataset With Embodied Language For Clutter Scene Grounding. , 2021, , .		3
553	Addressing Class Imbalance in Scene Graph Parsing by Learning to Contrast and Score. Lecture Notes in Computer Science, 2021, , 461-477.	1.0	0
554	Exploring and Exploiting the Hierarchical Structure of a Scene for Scene Graph Generation. , 2021, , .		0
555	Zero-Shot Video Event Detection With High-Order Semantic Concept Discovery and Matching. IEEE Transactions on Multimedia, 2022, 24, 1896-1908.	5.2	7
556	Relation Regularized Scene Graph Generation. IEEE Transactions on Cybernetics, 2022, 52, 5961-5972.	6.2	7
557	Transferable Interactiveness Knowledge for Human-Object Interaction Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, PP, 1-1.	9.7	12
558	An Integrative Review of Image Captioning Research. Journal of Physics: Conference Series, 2021, 1748, 042060.	0.3	3
559	Evaluation metrics for measuring bias in search engine results. Information Retrieval, 2021, 24, 85-113.	1.6	20
560	P â‰ˆ NP, at least in Visual Question Answering. , 2021, , .		0
561	MAGNet: Multi-Region Attention-Assisted Grounding of Natural Language Queries at Phrase Level. , 2021, , .		0

#	Article	IF	Citations
562	Human-Centric Spatio-Temporal Video Grounding With Visual Transformers. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 8238-8249.	5.6	36
563	Towards Visual Question Answering on Pathology Images. , 2021, , .		9
564	E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. , 2021, , .		31
565	Region-Aware Image Captioning via Interaction Learning. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 3685-3696.	5.6	79
566	Bilinear Graph Networks for Visual Question Answering. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34, 1023-1034.	7.2	15
567	Data-efficient Alignment of Multimodal Sequences by Aligning Gradient Updates and Internal Feature Distributions. , 2021, , .		0
568	Multi-Modal Reasoning Graph for Scene-Text Based Fine-Grained Image Classification and Retrieval. , 2021, , .		17
569	Enhancing Neural Machine Translation With Dual-Side Multimodal Awareness. IEEE Transactions on Multimedia, 2022, 24, 3013-3024.	5.2	4
570	Align R-CNN: A Pairwise Head Network for Visual Relationship Detection. IEEE Transactions on Multimedia, 2022, 24, 1266-1276.	5.2	6
571	Iterative Visual Relationship Detection via Commonsense Knowledge Graph. Big Data Research, 2021, 23, 100175.	2.6	2
572	Scene Graph Generation With Hierarchical Context. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32, 909-915.	7.2	23
573	A Novel Approach to Scene Graph Vectorization. , 2021, , .		1
574	Arrow R-CNN for handwritten diagram recognition. International Journal on Document Analysis and Recognition, 2021, 24, 3-17.	2.7	16
575	Scene graph generation via multi-relation classification and cross-modal attention coordinator. , 2021, , .		0
576	Interactive re-ranking for cross-modal retrieval based on object-wise question answering. , 2021, , .		3
577	Detection of pneumonia infection in lungs from chest X-ray images using deep convolutional neural network and content-based image retrieval techniques. Journal of Ambient Intelligence and Humanized Computing, 2021, , 1-8.	3.3	30
578	Multi-View Attention Network for Visual Dialog. Applied Sciences (Switzerland), 2021, 11, 3009.	1.3	10
579	Attribute-Image Similarity Measure for Multimodal Attention Mechanism. , 2021, , .		0

#	Article	IF	Citations
580	Scene graph generation by multi-level semantic tasks. Applied Intelligence, 2021, 51, 7781-7793.	3.3	5
581	Relationship graph learning network for visual relationship detection. , 2021, , .		1
582	Joint embedding VQA model based on dynamic word vector. PeerJ Computer Science, 2021, 7, e353.	2.7	123
583	A survey of deep learning-based visual question answering. Journal of Central South University, 2021, 28, 728-746.	1.2	4
584	Scene Graph Inference via Multi-Scale Context Modeling. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31, 1031-1041.	5.6	13
585	A Survey on Multimodal Deep Learning for Image Synthesis. , 2021, , .		3
586	Features to Text: A Comprehensive Survey of Deep Learning on Semantic Segmentation and Image Captioning. Complexity, 2021, 2021, 1-19.	0.9	10
587	Information to Wisdom: Commonsense Knowledge Extraction and Compilation. , 2021, , .		16
588	Visual relationship detection with recurrent attention and negative sampling. Neurocomputing, 2021, 434, 55-66.	3.5	10
589	3-D Relation Network for visual relation recognition in videos. Neurocomputing, 2021, 432, 91-100.	3.5	16
590	Polysemy Deciphering Network for Robust Humanâ€“Object Interaction Detection. International Journal of Computer Vision, 2021, 129, 1910-1929.	10.9	24
591	Multimodal graph inference network for scene graph generation. Applied Intelligence, 2021, 51, 8768.	3.3	1
592	Counterfactual attribute-based visual explanations for classification. International Journal of Multimedia Information Retrieval, 2021, 10, 127-140.	3.6	2
593	Exploiting structured high-level knowledge for domain-specific visual classification. Pattern Recognition, 2021, 112, 107806.	5.1	6
594	Linguistically-aware attention for reducing the semantic gap in vision-language tasks. Pattern Recognition, 2021, 112, 107812.	5.1	10
595	A Survey on Image Captioning datasets and Evaluation Metrics. IOP Conference Series: Materials Science and Engineering, 2021, 1116, 012184.	0.3	1
596	Knowledge-driven description synthesis for floor plan interpretation. International Journal on Document Analysis and Recognition, 2021, 24, 19.	2.7	5
597	Image captioning via proximal policy optimization. Image and Vision Computing, 2021, 108, 104126.	2.7	10

#	Article	IF	Citations
598	Generating Accurate Caption Units for Figure Captioning. , 2021, , .		18
599	Radial Graph Convolutional Network for Visual Question Generation. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32, 1654-1667.	7.2	32
600	Visual knowledge: an attempt to explore machine creativity. Frontiers of Information Technology and Electronic Engineering, 2021, 22, 619-624.	1.5	1
601	Scene-Graph-Guided message passing network for dense captioning. Pattern Recognition Letters, 2021, 145, 187-193.	2.6	5
602	Video action detection by learning graph-based spatio-temporal interactions. Computer Vision and Image Understanding, 2021, 206, 103187.	3.0	13
603	Ideal Words. KI - Kunstliche Intelligenz, 2021, 35, 271-290.	2.2	4
604	NLP Meets Vision for Visual Interpretation - A Retrospective Insight and Future directions. , 2021, , .		2
605	OCNet: Object Context for Semantic Segmentation. International Journal of Computer Vision, 2021, 129, 2375-2398.	10.9	121
606	SRR-LGR: Localâ€“Global Information-Reasoned Social Relation Recognition for Human-Oriented Observation. Remote Sensing, 2021, 13, 2038.	1.8	4
607	Inferring spatial relations from textual descriptions of images. Pattern Recognition, 2021, 113, 107847.	5.1	2
608	Medical Big Data Analysis with Attention and Large Margin Loss Model for Skin Lesion Application. Journal of Signal Processing Systems, 2021, 93, 827-839.	1.4	3
609	Weakly-supervised video object localization with attentive spatio-temporal correlation. Pattern Recognition Letters, 2021, 145, 232-239.	2.6	2
610	An Improved Attention for Visual Question Answering. , 2021, , .		25
611	ObjectGraphs: Using Objects and a Graph Convolutional Network for the Bottom-up Recognition and Explanation of Events in Video. , 2021, , .		6
613	Show and Speak: Directly Synthesize Spoken Description of Images. , 2021, , .		2
614	Align or attend? Toward More Efficient and Accurate Spoken Word Discovery Using Speech-to-Image Retrieval. , 2021, , .		4
615	Integrating Scene Semantic Knowledge into Image Captioning. ACM Transactions on Multimedia Computing, Communications and Applications, 2021, 17, 1-22.	3.0	28
616	Towards goal-oriented semantic signal processing: Applications and future challenges. , 2021, 119, 103134.		26

#	Article	IF	Citations
617	Interpretable visual reasoning: A survey. Image and Vision Computing, 2021, 112, 104194.	2.7	5
618	Class-Selective Mini-Batching and Multitask Learning for Visual Relationship Recognition. SAIEE Africa Research Journal, 2021, 112, 99-109.	1.1	0
619	Knowledge Reasoning for Semantic Segmentation. , 2021, , .		6
620	Practical Cross-modal Manifold Alignment for Robotic Grounded Language Learning. , 2021, , .		0
621	Continual learning in cross-modal retrieval. , 2021, , .		6
622	Scaling Human-Object Interaction Recognition in the Video through Zero-Shot Learning. Computational Intelligence and Neuroscience, 2021, 2021, 1-15.	1.1	3
623	TIED: A Cycle Consistent Encoder-Decoder Model for Text-to-Image Retrieval. , 2021, , .		4
624	Linguistic issues behind visual question answering. Language and Linguistics Compass, 2021, 15, e12417.	1.3	6
625	Integrating knowledge-based sparse representation for image detection. Neurocomputing, 2021, 442, 173-183.	3.5	0
626	Target-Tailored Source-Transformation for Scene Graph Generation. , 2021, , .		1
627	Background and foreground disentangled generative adversarial network for scene image synthesis. Computers and Graphics, 2021, 97, 54-66.	1.4	3
629	Exploring Explicit And Implicit Visual Relationships For Image Captioning. , 2021, , .		3
630	Visual relationship detection with region topology structure. Information Sciences, 2021, 564, 384-395.	4.0	8
631	Semantic Relation Model and Dataset for Remote Sensing Scene Understanding. ISPRS International Journal of Geo-Information, 2021, 10, 488.	1.4	5
632	How saccadic vision might help with the interpretability of deep networks. , 2021, , .		0
633	LPF: A Language-Prior Feedback Objective Function for De-biased Visual Question Answering. , 2021, , .		10
634	Towards Generating and Evaluating Iconographic Image Captions of Artworks. Journal of Imaging, 2021, 7, 123.	1.7	17
635	Visual question answering based on local-scene-aware referring expression generation. Neural Networks, 2021, 139, 158-167.	3.3	14

#	Article	IF	Citations
636	Relationship-Aware Primal-Dual Graph Attention Network For Scene Graph Generation. , 2021, , .		0
637	Do We Really Reduce Bias for Scene Graph Generation?. , 2021, , .		0
638	UUCT - HyMP: Towards Tracking Dispersed Crowd Groups from UAVs. , 2021, , .		2
639	Global Object Proposals for Improving Multi-Sentence Video Descriptions. , 2021, , .		1
640	Adapted Graph Reasoning and Filtration for Description-Image Retrieval. , 2021, , .		2
641	Robust Visual Relationship Detection towards Sparse Images in Internet-of-Things. Wireless Communications and Mobile Computing, 2021, 2021, 1-10.	0.8	1
642	Natural language guided object retrieval in images. Acta Informatica, 2021, 58, 243-261.	0.5	1
643	Attention LSTM for Scene Graph Generation. , 2021, , .		1
644	Select, Substitute, Search: A New Benchmark for Knowledge-Augmented Visual Question Answering. , 2021, , .		8
645	Learning Homogeneous and Heterogeneous Co-Occurrences for Unsupervised Cross-Modal Retrieval. , 2021, , .		3
646	Multiple Hub-Driven Attention Graph Network for Scene Graph Generation. , 2021, , .		1
647	Visual-Semantic Dual Channel Network for Visual Question Answering. , 2021, , .		1
648	Passage Retrieval for Outside-Knowledge Visual Question Answering. , 2021, , .		8
649	PAL-Net: Predicate-Aware Learning Network for Visual Relationship Recognition. , 2021, , .		0
650	Auxiliary Bi-Level Graph Representation for Cross-Modal Image-Text Retrieval. , 2021, , .		8
651	On the Limitations of Visual-Semantic Embedding Networks for Image-to-Text Information Retrieval. Journal of Imaging, 2021, 7, 125.	1.7	8
652	Looking Back and Forward: Enhancing Image Captioning with Global Semantic Guidance. , 2021, , .		0
653	Dense Video Captioning with Hierarchical Attention-Based Encoder-Decoder Networks. , 2021, , .		1

#	Article	IF	Citations
654	GilBERT: Generative Vision-Language Pre-Training for Image-Text Retrieval. , 2021, , .		18
655	Heterogeneous Attention Network for Effective and Efficient Cross-modal Retrieval. , 2021, , .		35
656	Few-shot Object Detection with Camouflage Animals. Journal of Physics: Conference Series, 2021, 1961, 012005.	0.3	0
657	DiMBERT: Learning Vision-Language Grounded Representations with Disentangled Multimodal-Attention. ACM Transactions on Knowledge Discovery From Data, 2022, 16, 1-19.	2.5	3
658	ReLaB: Reliable Label Bootstrapping for Semi-Supervised Learning. , 2021, , .		3
659	DMRFNet: Deep Multimodal Reasoning and Fusion for Visual Question Answering and explanation generation. Information Fusion, 2021, 72, 70-79.	11.7	27
660	Multi goals and multi scenes visual mapless navigation in indoor using meta-learning and scene priors. Neurocomputing, 2021, 449, 368-377.	3.5	5
661	Heterogeneous Excitation-and-Squeeze Network for visual dialog. Neurocomputing, 2021, 449, 399-410.	3.5	5
662	HSGMP. , 2021, , .		6
663	Language-Conditioned Region Proposal and Retrieval Network for Referring Expression Comprehension. , 2021, , .		1
664	Fine-Grained Cross-Modal Retrieval for Cultural Items with Focal Attention and Hierarchical Encodings. Computers, 2021, 10, 105.	2.1	2
665	DIFFERENT APPLICATION AREAS OF OBJECT DETECTION WITH DEEP LEARNING. AkÄ±llÄ± UlaÅŸÄ±m Sistemleri Ve UygulamalarÄ± Dergisi:, 0, , .	0.2	1
666	Neural Symbolic Representation Learning for Image Captioning. , 2021, , .		5
667	Web-Scale Generic Object Detection at Microsoft Bing. , 2021, , .		0
668	3DRM: Pair-wise relation module for 3D object detection. Computers and Graphics, 2021, 98, 58-70.	1.4	5
669	Learning to Select. , 2021, , .		5
670	Graph-LSTM with Global Attribute for Scene Graph Generation. Journal of Physics: Conference Series, 2021, 2003, 012001.	0.3	2
672	Say What You Are Looking At: An Attention-Based Interactive System for Autistic Children. Applied Sciences (Switzerland), 2021, 11, 7426.	1.3	1

#	Article	IF	Citations
673	Rethinking semantic-visual alignment in zero-shot object detection via a softplus margin focal loss. Neurocomputing, 2021, 449, 117-135.	3.5	9
674	Boosting convolutional image captioning with semantic content and visual relationship. Displays, 2021, 70, 102069.	2.0	26
675	Residual Spatiotemporal Autoencoder with Skip Connected and Memory Guided Network for Detecting Video Anomalies. Neural Processing Letters, 2021, 53, 4677-4692.	2.0	6
676	Pre-trained models: Past, present and future. AI Open, 2021, 2, 225-250.	9.1	253
677	PAM: Understanding Product Images in Cross Product Category Attribute Extraction. , 2021, , .		8
678	Improving Object Detection And Attribute Recognition By Feature Entanglement Reduction. , 2021, , .		0
679	Contextual Label Transformation For Scene Graph Generation. , 2021, , .		0
680	Reasoning like Humans: On Dynamic Attention Prior in Image Captioning. Knowledge-Based Systems, 2021, 228, 107313.	4.0	10
681	Multimodal feature-wise co-attention method for visual question answering. Information Fusion, 2021, 73, 1-10.	11.7	38
682	A survey on generative adversarial network-based text-to-image synthesis. Neurocomputing, 2021, 451, 316-336.	3.5	25
683	A detailed review of prevailing image captioning methods using deep learning techniques. Multimedia Tools and Applications, 2022, 81, 1313-1336.	2.6	5
684	Sparse And Structured Visual Attention. , 2021, , .		2
685	Dimensions of commonsense knowledge. Knowledge-Based Systems, 2021, 229, 107347.	4.0	24
686	Pick-Object-Attack: Type-specific adversarial attack for object detection. Computer Vision and Image Understanding, 2021, 211, 103257.	3.0	6
687	Revisiting image captioning via maximum discrepancy competition. Pattern Recognition, 2022, 122, 108358.	5.1	9
688	MUMC: Minimizing uncertainty of mixture of cues. Image and Vision Computing, 2021, 115, 104280.	2.7	1
689	Multi-type decision fusion network for visual Q&A. Image and Vision Computing, 2021, 115, 104281.	2.7	5
690	Metadata based need-to-know view in large-scale video surveillance systems. Computers and Security, 2021, 111, 102452.	4.0	3

#	Article	IF	Citations
691	Adversarial text-to-image synthesis: A review. Neural Networks, 2021, 144, 187-209.	3.3	78
692	Accuracy vs. complexity: A trade-off in visual question answering models. Pattern Recognition, 2021, 120, 108106.	5.1	15
693	Multimodal research in vision and language: A review of current and emerging trends. Information Fusion, 2022, 77, 149-171.	11.7	36
694	Atom correlation based graph propagation for scene graph generation. Pattern Recognition, 2022, 122, 108300.	5.1	9
695	A multimodal attention fusion network with a dynamic vocabulary for TextVQA. Pattern Recognition, 2022, 122, 108214.	5.1	12
696	Automatic Detection of Discrimination Actions from Social Images. Electronics (Switzerland), 2021, 10, 325.	1.8	0
698	Stimuli-Aware Visual Emotion Analysis. IEEE Transactions on Image Processing, 2021, 30, 7432-7445.	6.0	23
699	Learning to Select Question-Relevant Relations for Visual Question Answering. , 2021, , .		1
700	Video Question Answering with Phrases via Semantic Roles. , 2021, , .		4
701	Global-Affine and Local-Specific Generative Adversarial Network for semantic-guided image generation. Mathematical Foundations of Computing, 2021, 4, 145.	0.7	0
703	Automatically Generating Natural Language Descriptions of Images by a Deep Hierarchical Framework. IEEE Transactions on Cybernetics, 2022, 52, 7441-7452.	6.2	6
704	Multitask Learning for Visual Question Answering. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34, 1380-1394.	7.2	4
705	Optimistic Agent: Accurate Graph-Based Value Estimation for More Successful Visual Navigation. , 2021, , .		7
706	Document Domain Randomization for Deep Learning Document Layout Extraction. Lecture Notes in Computer Science, 2021, , 497-513.	1.0	1
707	Rakutenâ€™s Participation in WAT 2021: Examining the Effectiveness of Pre-trained Models for Multilingual and Multimodal Machine Translation. , 2021, , .		1
708	Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection. IEEE Transactions on Multimedia, 2022, 24, 3455-3468.	5.2	22
709	Relatable Clothing: Soft-Attention Mechanism for Detecting Worn/Unworn Objects. IEEE Access, 2021, 9, 108782-108792.	2.6	2
710	Maintaining Common Ground in Dynamic Environments. Transactions of the Association for Computational Linguistics, 2021, 9, 995-1011.	3.2	0

#	Article	IF	Citations
711	Graph Reasoning-Based Emotion Recognition Network. IEEE Access, 2021, 9, 6488-6497.	2.6	15
712	Graphhopper: Multi-hop Scene Graph Reasoning for Visual Question Answering. Lecture Notes in Computer Science, 2021, , 111-127.	1.0	7
713	Self-Distillation for Few-Shot Image Captioning. , 2021, , .		8
714	Toward Region-Aware Attention Learning for Scene Graph Generation. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33, 7655-7666.	7.2	11
715	Cross-Lingual Visual Grounding. IEEE Access, 2021, 9, 349-358.	2.6	1
716	Unpaired Image Captioning With semantic-Constrained Self-Learning. IEEE Transactions on Multimedia, 2022, 24, 904-916.	5.2	21
717	Few-Shot Image and Sentence Matching via Aligned Cross-Modal Memory. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 2968-2983.	9.7	3
718	A Novel Attention-based Aggregation Function to Combine Vision and Language. , 2021, , .		3
719	Multi-Gate Attention Network for Image Captioning. IEEE Access, 2021, 9, 69700-69709.	2.6	12
720	Iconographic Image Captioning for Artworks. Lecture Notes in Computer Science, 2021, , 502-516.	1.0	9
721	Probing Multi-modal Machine Translation with Pre-trained Language Model. , 2021, , .		3
722	Zero-Shot Scene Graph Relation Prediction Through Commonsense Knowledge Integration. Lecture Notes in Computer Science, 2021, , 466-482.	1.0	13
723	Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. Transactions of the Association for Computational Linguistics, 2021, 9, 978-994.	3.2	37
724	Re-Attention for Visual Question Answering. IEEE Transactions on Image Processing, 2021, 30, 6730-6743.	6.0	41
725	DocVQA: A Dataset for VQA on Document Images. , 2021, , .		78
726	Cross-modality co-attention networks for visual question answering. Soft Computing, 2021, 25, 5411-5421.	2.1	17
727	VisQA: X-raying Vision and Language Reasoning in Transformers. IEEE Transactions on Visualization and Computer Graphics, 2022, 28, 976-986.	2.9	10
728	Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. IEEE Transactions on Visualization and Computer Graphics, 2022, 28, 1073-1083.	2.9	29

#	Article	IF	Citations
729	Diverse and Coherent Paragraph Generation from Images. Lecture Notes in Computer Science, 2018, , 747-763.	1.0	34
730	Zoom-Net: Mining Deep Feature Interactions for Visual Relationship Recognition. Lecture Notes in Computer Science, 2018, , 330-347.	1.0	78
731	Hierarchical Relational Networks for Group Activity Recognition and Retrieval. Lecture Notes in Computer Science, 2018, , 742-758.	1.0	77
732	Stacked Cross Attention for Image-Text Matching. Lecture Notes in Computer Science, 2018, , 212-228.	1.0	528
733	Learning to Segment via Cut-and-Paste. Lecture Notes in Computer Science, 2018, , 39-54.	1.0	43
734	Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering. Lecture Notes in Computer Science, 2018, , 460-477.	1.0	51
735	NNEval: Neural Network Based Evaluation Metric for Image Captioning. Lecture Notes in Computer Science, 2018, , 39-55.	1.0	6
736	Graph R-CNN for Scene GraphÂGeneration. Lecture Notes in Computer Science, 2018, , 690-706.	1.0	351
737	Dynamic Multimodal Instance Segmentation Guided by Natural Language Queries. Lecture Notes in Computer Science, 2018, , 656-672.	1.0	75
738	Shuffle-Then-Assemble: Learning Object-Agnostic Visual Relationship Features. Lecture Notes in Computer Science, 2018, , 38-54.	1.0	40
739	Visual Question Generation for Class Acquisition of Unknown Objects. Lecture Notes in Computer Science, 2018, , 492-507.	1.0	11
740	Object Level Visual Reasoning in Videos. Lecture Notes in Computer Science, 2018, , 106-122.	1.0	65
741	Compositional Learning for Human Object Interaction. Lecture Notes in Computer Science, 2018, , 247-264.	1.0	67
742	Exploring Visual Relationship for Image Captioning. Lecture Notes in Computer Science, 2018, , 711-727.	1.0	440
743	Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. Lecture Notes in Computer Science, 2018, , 353-369.	1.0	80
744	ADVISE: Symbolism and External Knowledge for Decoding Advertisements. Lecture Notes in Computer Science, 2018, , 868-886.	1.0	23
745	Hierarchical Vision-Language Alignment for Video Captioning. Lecture Notes in Computer Science, 2019, , 42-54.	1.0	12
746	Visual Graphs from Motion (VGfM): Scene Understanding with Object Geometry Reasoning. Lecture Notes in Computer Science, 2019, , 330-346.	1.0	6

#	Article	IF	Citations
747	Artpedia: A New Visual-Semantic Dataset with Visual and Contextual Sentences in the Artistic Domain. Lecture Notes in Computer Science, 2019, , 729-740.	1.0	25
748	Image-to-Image Translation to Unfold the Reality of Artworks: An Empirical Analysis. Lecture Notes in Computer Science, 2019, , 741-752.	1.0	3
749	Deep Learning-Based Video Retrieval Using Object Relationships and Associated Audio Classes. Lecture Notes in Computer Science, 2020, , 803-808.	1.0	1
750	AQuA: ASP-Based Visual Question Answering. Lecture Notes in Computer Science, 2020, , 57-72.	1.0	8
751	Utilising Information Foraging Theory for User Interaction with Image Query Auto-Completion. Lecture Notes in Computer Science, 2020, , 666-680.	1.0	4
752	Contextual Heterogeneous Graph Network for Human-Object Interaction Detection. Lecture Notes in Computer Science, 2020, , 248-264.	1.0	33
753	Captioning Images Taken by People Who Are Blind. Lecture Notes in Computer Science, 2020, , 417-434.	1.0	67
754	Large-Scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline. Lecture Notes in Computer Science, 2020, , 336-352.	1.0	50
755	Seeing the Un-Scene: Learning Amodal Semantic Maps for Room Navigation. Lecture Notes in Computer Science, 2020, , 513-529.	1.0	18
756	Representation Learning on Visual-Symbolic Graphs for Video Understanding. Lecture Notes in Computer Science, 2020, , 71-90.	1.0	13
757	Semantic Equivalent Adversarial Data Augmentation for Visual Question Answering. Lecture Notes in Computer Science, 2020, , 437-453.	1.0	19
758	Controlling Style and Semantics in Weakly-Supervised Image Generation. Lecture Notes in Computer Science, 2020, , 482-499.	1.0	20
759	Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models. Lecture Notes in Computer Science, 2020, , 565-580.	1.0	34
760	Hallucinating Visual Instances in Total Absentia. Lecture Notes in Computer Science, 2020, , 264-282.	1.0	10
761	Connecting Vision and Language with Localized Narratives. Lecture Notes in Computer Science, 2020, , 647-664.	1.0	50
762	Comprehensive Image Captioning via Scene Graph Decomposition. Lecture Notes in Computer Science, 2020, , 211-229.	1.0	48
763	UNITER: UNiversal Image-TExt Representation Learning. Lecture Notes in Computer Science, 2020, , 104-120.	1.0	541
764	Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. Lecture Notes in Computer Science, 2020, , 121-137.	1.0	452

#	Article	IF	Citations
765	Efficient Attention Mechanism for Visual Dialog that Can Handle All the Interactions Between Multiple Inputs. Lecture Notes in Computer Science, 2020, , 223-240.	1.0	23
766	Visual Question Answering on Image Sets. Lecture Notes in Computer Science, 2020, , 51-67.	1.0	10
767	Bridging Knowledge Graphs to Generate Scene Graphs. Lecture Notes in Computer Science, 2020, , 606-623.	1.0	83
768	RetrieveGAN: Image Synthesis via Differentiable Patch Retrieval. Lecture Notes in Computer Science, 2020, , 242-257.	1.0	13
769	Visual-Relation Conscious Image Generation from Structured-Text. Lecture Notes in Computer Science, 2020, , 290-306.	1.0	7
770	Neural Networks for Detecting Irrelevant Questions During Visual Question Answering. Lecture Notes in Computer Science, 2020, , 786-797.	1.0	3
771	Modeling Context Between Objects for Referring Expression Understanding. Lecture Notes in Computer Science, 2016, , 792-807.	1.0	146
772	Spatio-Temporal Attention Models for Grounded Video Captioning. Lecture Notes in Computer Science, 2017, , 104-119.	1.0	11
773	GHio-Ca: An Android Application for Automatic Image Classification. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 2018, , 248-257.	0.2	4
774	Caption-Based Region Extraction in Images. Advances in Intelligent Systems and Computing, 2020, , 27-38.	0.5	2
776	Learning Semantic-Specific Graph Representation for Multi-Label Image Recognition. , 2019, , .		160
777	Contextual Translation Embedding for Visual Relationship Detection and Scene Graph Generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43, 3820-3832.	9.7	39
778	A Survey of Single-Scene Video Anomaly Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, PP, 1-1.	9.7	73
779	Auto-encoding and Distilling Scene Graphs for Image Captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, PP, 1-1.	9.7	21
780	Sequence-to-sequence image caption generator. , 2019, , .		5
781	RAID. ACM Transactions on Graphics, 2016, 35, 1-12.	4.9	4
782	Learning and reasoning in logic tensor networks. , 2017, , .		15
783	Commonsense Knowledge in Machine Intelligence. SIGMOD Record, 2018, 46, 49-52.	0.7	63

#	Article	IF	Citations
784	Context-Aware Visual Policy Network for Sequence-Level Image Captioning. , 2018, , .		64
785	Context-Dependent Diffusion Network for Visual Relationship Detection. , 2018, , .		31
786	Sketchforme. , 2019, , .		13
787	Aligning Linguistic Words and Visual Semantic Units for Image Captioning. , 2019, , .		66
788	Towards fairer datasets. , 2020, , .		123
789	Does learning require memorization? a short tale about a long tail. , 2020, , .		70
790	ACMNet. ACM Transactions on Multimedia Computing, Communications and Applications, 2020, 16, 1-21.	3.0	6
791	Explaining with Counter Visual Attributes and Examples. , 2020, , .		7
792	Visual Relations Augmented Cross-modal Retrieval. , 2020, , .		13
793	Adversarial Attacks on Deep-learning Models in Natural Language Processing. ACM Transactions on Intelligent Systems and Technology, 2020, 11, 1-41.	2.9	193
794	Visual Question Generation. ACM Computing Surveys, 2021, 53, 1-22.	16.1	13
795	Image Captioning in Chinese and Its Application for Children with Autism Spectrum Disorder. , 2020, , .		5
796	Constrained LSTM and Residual Attention for Image Captioning. ACM Transactions on Multimedia Computing, Communications and Applications, 2020, 16, 1-18.	3.0	24
797	DeVLBert. , 2020, , .		40
798	Generalized Zero-shot Learning with Multi-source Semantic Embeddings for Scene Recognition. , 2020, , .		6
799	Weakly-Supervised Video Object Grounding by Exploring Spatio-Temporal Contexts. , 2020, , .		36
800	Relational Graph Learning for Grounded Video Description Generation. , 2020, , .		17
801	Give Me Something to Eat: Referring Expression Comprehension with Commonsense Knowledge. , 2020, , .		12

#	Article	IF	Citations
802	Boosting Visual Question Answering with Context-aware Knowledge Aggregation. , 2020, , .		29
803	Context-Aware Multi-View Summarization Network for Image-Text Matching. , 2020, , .		58
804	Combo-Attention Network for Baidu Video Advertising. , 2020, , .		26
805	Image Captioning with a Joint Attention Mechanism by Visual Concept Samples. ACM Transactions on Multimedia Computing, Communications and Applications, 2020, 16, 1-22.	3.0	13
806	FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval. , 2020, , .		65
807	CrowdCO-OP. Proceedings of the ACM on Human-Computer Interaction, 2020, 4, 1-24.	2.5	14
808	Automatic Arabic Image Captioning using RNN-LSTM-Based Language Model and CNN. International Journal of Advanced Computer Science and Applications, 2018, 9, .	0.5	24
809	Snuba. Proceedings of the VLDB Endowment, 2018, 12, 223-236.	2.1	69
810	Explorations into Deep Learning Text Architectures for Dense Image Captioning. , 0, , .		3
811	Obtaining Faithful Interpretations from Compositional Neural Networks. , 2020, , .		14
812	Words Arenâ€™t Enough, Their Order Matters: On the Robustness of Grounding Visual Referring Expressions. , 2020, , .		13
813	Aligned Dual Channel Graph Convolutional Network for Visual Question Answering. , 2020, , .		43
814	What are the Goals of Distributional Semantics?. , 2020, , .		9
815	Improving Image Captioning with Better Use of Caption. , 2020, , .		35
816	Cross-Modality Relevance for Reasoning on Language and Vision. , 2020, , .		16
817	TVQA+: Spatio-Temporal Grounding for Video Question Answering. , 2020, , .		71
818	Interactive Key-Value Memory-augmented Attention for Image Paragraph Captioning. , 2020, , .		9
819	Re-solve it: simulating the acquisition of core semantic competences from small data. , 2020, , .		3

#	Article	IF	Citations
820	Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. , 2020, , .		23
821	Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. , 2020, , .		5
822	Fine-Grained Grounding for Multimodal Speech Recognition. , 2020, , .		4
823	Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision. , 2020, , .		4
824	A Linguistic Analysis of Visually Grounded Dialogues Based on Spatial Expressions. , 2020, , .		4
825	Obj2Text: Generating Visually Descriptive Language from Object Layouts. , 2017, , .		31
826	Training for Diversity in Image Paragraph Captioning. , 2018, , .		29
827	Object Hallucination in Image Captioning. , 2018, , .		88
828	Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. , 2019, , .		123
829	Dual Attention Networks for Visual Reference Resolution in Visual Dialog. , 2019, , .		46
830	LXMERT: Learning Cross-Modality Encoder Representations from Transformers. , 2019, , .		909
831	Sunny and Dark Outside?! Improving Answer Consistency in VQA through Entailed Question Generation. , 2019, , .		26
832	Do Nuclear Submarines Have Nuclear Captains? A Challenge Dataset for Commonsense Reasoning over Adjectives and Objects. , 2019, , .		4
833	Stacking with Auxiliary Features for Visual Question Answering. , 2018, , .		9
834	Pragmatically Informative Image Captioning with Character-Level Inference. , 2018, , .		27
835	Automatic Metric Validation for Grammatical Error Correction. , 2018, , .		14
836	The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. , 2019, , .		26
837	Generating Question Relevant Captions to Aid Visual Question Answering. , 2019, , .		26

#	Article	IF	Citations
838	Improving Visual Question Answering by Referring to Generated Paragraph Captions. , 2019, , .		12
839	Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems. , 2019, , .		43
840	Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog. , 2019, , .		62
841	Multimodal Logical Inference System for Visual-Textual Entailment. , 2019, , .		4
842	Exploring the Functional and Geometric Bias of Spatial Relations Using Neural Language Models. , 2018, , .		7
843	Anaphora Resolution for Improving Spatial Relation Extraction from Text. , 2018, , .		6
844	SpatialVOC2K: A Multilingual Dataset of Images with Annotations and Features for Spatial Relations between Objects. , 2018, , .		3
845	Talking about other people: an endless range of possibilities. , 2018, , .		7
846	Learning Question-Guided Video Representation for Multi-Turn Video Question Answering. , 2019, , .		2
847	What goes into a word: generating image descriptions with top-down spatial knowledge. , 2019, , .		8
848	Object Detection Meets Knowledge Graphs. , 2017, , .		68
849	Representation Learning for Scene Graph Completion via Jointly Structural and Visual Embedding. , 2018, , .		20
850	Feature Enhancement in Attention for Visual Question Answering. , 2018, , .		10
851	Dual Visual Attention Network for Visual Dialog. , 2019, , .		25
852	Swell-and-Shrink: Decomposing Image Captioning by Transformation and Summarization. , 2019, , .		4
853	Towards a Framework for Visual Intelligence in Service Robotics: Epistemic Requirements and Gap Analysis. , 2020, , .		3
855	A Survey on Different Deep Learning Architectures for Image Captioning. WSEAS Transactions on Systems and Control, 2020, 15, 635-646.	0.5	2
856	Multi-branch Graph Network for Learning Human-Object Interaction. Lecture Notes in Computer Science, 2021, , 425-436.	1.0	1

#	Article	IF	Citations
857	Hierarchical Group-Level Emotion Recognition. IEEE Transactions on Multimedia, 2021, 23, 3892-3906.	5.2	12
858	The Role of Syntactic Planning in Compositional Image Captioning. , 2021, , .		2
859	Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 7778-7796.	9.7	148
860	SOLVER: Scene-Object Interrelated Visual Emotion Reasoning Network. IEEE Transactions on Image Processing, 2021, 30, 8686-8701.	6.0	17
861	Synthesizing Spoken Descriptions of Images. IEEE/ACM Transactions on Audio Speech and Language Processing, 2021, 29, 3242-3254.	4.0	2
862	High-Order Interaction Learning for Image Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 4417-4430.	5.6	52
863	Automatic Generation of Descriptive Titles for Video Clips Using Deep Learning. Transactions on Computational Science and Computational Intelligence, 2021, , 17-28.	0.3	8
864	A Universal Quaternion Hypergraph Network for Multimodal Video Question Answering. IEEE Transactions on Multimedia, 2023, 25, 38-49.	5.2	6
865	Dense Relational Image Captioning via Multi-Task Triple-Stream Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 7348-7362.	9.7	19
866	Hierarchical Planning for Long-Horizon Manipulation with Geometric and Symbolic Scene Graphs. , 2021, , .		27
867	A deep dense captioning framework with joint localization and contextual reasoning. Journal of Central South University, 2021, 28, 2801-2813.	1.2	0
868	Layout Structure Assisted Indoor Image Generation. , 2021, , .		0
869	Triangle-Reward Reinforcement Learning. , 2021, , .		4
870	Distributed Attention for Grounded Image Captioning. , 2021, , .		9
871	Multi-Perspective Video Captioning. , 2021, , .		11
872	Learning to Understand Traffic Signs. , 2021, , .		4
873	Neighbor-view Enhanced Model for Vision and Language Navigation. , 2021, , .		24
874	Towards Reasoning Ability in Scene Text Visual Question Answering. , 2021, , .		9

#	Article	IF	Citations
875	Similar Scenes Arouse Similar Emotions. , 2021, , .		11
876	Database-adaptive Re-ranking for Enhancing Cross-modal Image Retrieval. , 2021, , .		6
877	Screen Parsing: Towards Reverse Engineering of UI Models from Screenshots. , 2021, , .		21
878	VQA as a factoid question answering problem: A novel approach for knowledge-aware and explainable visual question answering. Image and Vision Computing, 2021, 116, 104328.	2.7	5
879	Mask and Predict. , 2021, , .		7
880	X-GGM: Graph Generative Modeling for Out-of-distribution Generalization in Visual Question Answering. , 2021, , .		10
881	Personalized Multi-modal Video Retrieval on Mobile Devices. , 2021, , .		0
882	Recovering the Unbiased Scene Graphs from the Biased Ones. , 2021, , .		49
883	Weakly-Supervised Video Object Grounding via Stable Context Learning. , 2021, , .		6
884	A Picture is Worth a Thousand Words. , 2021, , .		3
885	ION. , 2021, , .		8
886	Exploring Logical Reasoning for Referring Expression Comprehension. , 2021, , .		4
887	Position-Augmented Transformers with Entity-Aligned Mesh for TextVQA. , 2021, , .		11
888	RDMMFET: Representation of Dense Multimodality Fusion Encoder Based on Transformer. Mobile Information Systems, 2021, 2021, 1-9.	0.4	0
890	Am I Allergic to This? Assisting Sight Impaired People in the Kitchen. , 2021, , .		1
891	ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross- and Intra-modal Knowledge Integration. , 2021, , .		22
892	Beyond OCR + VQA. , 2021, , .		25
893	Dense Contrastive Visual-Linguistic Pretraining. , 2021, , .		5

#	Article	IF	Citations
894	Dual Graph Convolutional Networks with Transformer and Curriculum Learning for Image Captioning. , 2021, , .		38
895	Unifying Multimodal Transformer for Bi-directional Image and Text Generation. , 2021, , .		20
896	Focal and Composed Vision-semantic Modeling for Visual Question Answering. , 2021, , .		4
897	Deconfounded and Explainable Interactive Vision-Language Retrieval of Complex Scenes. , 2021, , .		4
898	Fully Functional Image Manipulation Using Scene Graphs in A Bounding-Box Free Way. , 2021, , .		9
899	Hierarchical Semantic Enhanced Directional Graph Network for Visual Commonsense Reasoning. , 2021, , .		0
901	MedFuseNet:ÂAnÂattention-based multimodal deep learning model for visual question answering in the medical domain. Scientific Reports, 2021, 11, 19826.	1.6	24
902	A survey of methods, datasets and evaluation metrics for visual question answering. Image and Vision Computing, 2021, 116, 104327.	2.7	19
903	Image and Video Captioning Using Deep Architectures. , 2021, , 151-174.		0
904	ACD. , 2016, , .		5
905	CLEF 2017: Multimodal Spatial Role Labeling (mSpRL) Task Overview. Lecture Notes in Computer Science, 2017, , 367-376.	1.0	9
906	Learning Action Concept Trees and Semantic Alignment Networks from Image-Description Data. Lecture Notes in Computer Science, 2017, , 19-34.	1.0	0
907	Spatial Language Understanding with Multimodal Graphs using Declarative Learning based Programming. , 2017, , .		7
908	Review on latest approaches used in Natural Language Processing for generation of Image Captioning. International Journal of Computer Science and Engineering, 2017, 4, 41-48.	0.1	0
909	Concept Mask: Large-Scale Segmentation from Semantic Concepts. Lecture Notes in Computer Science, 2018, , 542-557.	1.0	2
910	Conditional Image-Text Embedding Networks. Lecture Notes in Computer Science, 2018, , 258-274.	1.0	50
911	Deep Neural Network Based Image Captioning. Lecture Notes in Computer Science, 2018, , 335-347.	1.0	0
912	Scene Graph Generation Based on Node-Relation Context Module. Lecture Notes in Computer Science, 2018, , 134-145.	1.0	2

#	Article	IF	Citations
913	Token-level and sequence-level loss smoothing for RNN language models. , 2018, , .		10
914	Visually Guided Spatial Relation Extraction from Text. , 2018, , .		5
915	Approximate Query Matching for Graph-Based Holistic Image Retrieval. Lecture Notes in Computer Science, 2018, , 72-84.	1.0	0
916	Dense Captioning of Natural Scenes in Spanish. Lecture Notes in Computer Science, 2018, , 145-154.	1.0	2
917	Pushing the Limits of Radiology with Joint Modeling of Visual and Textual Information. , 2018, , .		3
918	A Bottom-Up and Top-Down Approach for Image Captioning using Transformer. , 2018, , .		3
919	Geometry-aware Relational Exemplar Attention for Dense Captioning. , 2019, , .		2
921	Visually grounded generation of entailments from premises. , 2019, , .		0
922	Identifying and Explaining Discriminative Attributes. , 2019, , .		1
923	Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation. , 2019, , .		35
924	Exploiting Attention for Visual Relationship Detection. Lecture Notes in Computer Science, 2019, , 331-344.	1.0	3
925	Scene Graph Generation via Convolutional Message Passing and Class-Aware Memory Embeddings. Lecture Notes in Computer Science, 2019, , 620-633.	1.0	0
926	The Latent Semantic Power of Labels: Improving Image Classification via Natural Language Semantic. Lecture Notes in Computer Science, 2019, , 175-189.	1.0	1
927	DynGraph: Visual Question Answering via Dynamic Scene Graphs. Lecture Notes in Computer Science, 2019, , 428-441.	1.0	1
928	Learning to Relate from Captions and Bounding Boxes. , 2019, , .		1
929	YouMakeup: A Large-Scale Domain-Specific Multimodal Dataset for Fine-Grained Semantic Comprehension. , 2019, , .		11
930	Phrase Grounding by Soft-Label Chain Conditional Random Field. , 2019, , .		3
931	Knowing Where to Look? Analysis on Attention of Visual Question Answering System. Lecture Notes in Computer Science, 2019, , 145-152.	1.0	3

#	Article	IF	Citations
932	Dense Image Captioning Based on Precise Feature Extraction. Communications in Computer and Information Science, 2019, , 83-90.	0.4	1
933	Scene Recognition via Bi-enhanced Knowledge Space Learning. Communications in Computer and Information Science, 2019, , 215-223.	0.4	0
934	A Framework for Queryable Video Analysis. , 2019, , .		1
935	Visual Relationship Prediction via Label Clustering and Incorporation of Depth Information. Lecture Notes in Computer Science, 2019, , 571-581.	1.0	3
936	Distributional Semantics in the Real World: Building Word Vector Representations from a Truth-Theoretic Model. , 2019, , .		2
937	Automatic Classification and Reporting of Multiple Common Thorax Diseases Using Chest Radiographs. Advances in Computer Vision and Pattern Recognition, 2019, , 393-412.	0.9	1
938	On the Role of Scene Graphs in Image Captioning. , 2019, , .		14
939	Not Just a Matter of Semantics: The Relationship Between Visual and Semantic Similarity. Lecture Notes in Computer Science, 2019, , 414-427.	1.0	6
940	Evidence Humans Provide When Explaining Data-Labeling Decisions. Lecture Notes in Computer Science, 2019, , 390-409.	1.0	0
941	What a neural language model tells us about spatial relations. , 2019, , .		0
942	Expressing Visual Relationships via Language. , 2019, , .		18
943	Video question answering by frame attention. , 2019, , .		0
944	Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge. , 2019, , .		2
945	Finding Images by Dialoguing with Image. , 2019, , .		0
946	Explainable Video Action Reasoning via Prior Knowledge and State Transitions. , 2019, , .		30
947	MUCH. , 2019, , .		3
949	Neural Joking Machine. Journal of the Japan Society for Precision Engineering, 2019, 85, 1151-1156.	0.0	0
950	Amplifying Key Cues for Human-Object-Interaction Detection. Lecture Notes in Computer Science, 2020, , 248-265.	1.0	38

#	Article	IF	Citations
952	REXUP: I REason, I EXtract, I UPdate with Structured Compositional Reasoning for Visual Question Answering. Lecture Notes in Computer Science, 2020, , 520-532.	1.0	3
953	CLEVR Parser: A Graph Parser Library for Geometric Learning on Language Grounded Image Scenes. , 2020, , .		0
954	Length-Controllable Image Captioning. Lecture Notes in Computer Science, 2020, , 712-729.	1.0	25
955	Sketching Image Gist: Human-Mimetic Hierarchical Scene Graph Generation. Lecture Notes in Computer Science, 2020, , 222-239.	1.0	21
956	Multichannel Attention Refinement for Video Question Answering. ACM Transactions on Multimedia Computing, Communications and Applications, 2020, 16, 1-23.	3.0	14
958	Comparing human and automated approaches to visual storytelling. , 2020, , 159-196.		2
959	EViLBERT: Learning Task-Agnostic Multimodal Sense Embeddings. , 2020, , .		2
960	Multi-Level Multimodal Transformer Network for Multimodal Recipe Comprehension. , 2020, , .		1
961	Classification of Chest Diseases Using Deep Learning. Advances in Intelligent Systems and Computing, 2021, , 152-158.	0.5	1
962	Explaining transformer-based image captioning models: An empirical analysis. AI Communications, 2022, 35, 111-129.	0.8	12
963	Transformers in computational visual media: A survey. Computational Visual Media, 2022, 8, 33-62.	10.8	60
964	Deep image synthesis from intuitive user input: A review and perspectives. Computational Visual Media, 2022, 8, 3-31.	10.8	21
965	Relation Network and Causal Reasoning for Image Captioning. , 2021, , .		1
966	Assorted Attention Network for Cross-Lingual Language-to-Vision Retrieval. , 2021, , .		8
967	Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language. Lecture Notes in Computer Science, 2020, , 601-618.	1.0	22
968	Organizing Tagged Knowledge: Similarity Measures and Semantic Fluency in Structure Mining. Journal of Mechanical Design, Transactions of the ASME, 2020, 142, .	1.7	4
969	Image-Based World-perceiving Knowledge Graph (WpKG) with Imprecision. Communications in Computer and Information Science, 2020, , 415-428.	0.4	1
970	Story-driven Video Editing. IEEE Transactions on Multimedia, 2021, 23, 4027-4036.	5.2	5

#	Article	IF	Citations
971	CompGuessWhat?!: A Multi-task Evaluation Framework for Grounded Language Learning. , 2020, , .		5
972	Taking a Cue From the Human. Journal of Audiovisual Translation, 2020, 3, .	0.1	2
973	A Survey on VQA: Datasets and Approaches. , 2020, , .		1
974	Research on Text to Image Based on Generative Adversarial Network. , 2020, , .		2
975	RGB-D Based Visual Navigation Using Direction Estimation Module. Lecture Notes in Computer Science, 2021, , 252-264.	1.0	0
976	Visual question answering by pattern matching and reasoning. Neurocomputing, 2022, 467, 323-336.	3.5	12
977	A Multimodal Deep Framework for Derogatory Social Media Post Identification of a Recognized Person. ACM Transactions on Asian and Low-Resource Language Information Processing, 2022, 21, 1-19.	1.3	9
978	A visual persistence model for image captioning. Neurocomputing, 2022, 468, 48-59.	3.5	7
979	A unified deep sparse graph attention network for scene graph generation. Pattern Recognition, 2022, 123, 108367.	5.1	10
980	A Survey of Deep Active Learning. ACM Computing Surveys, 2022, 54, 1-40.	16.1	294
981	Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. Lecture Notes in Computer Science, 2020, , 1-17.	1.0	28
982	MAF: Multimodal Alignment Framework for Weakly-Supervised Phrase Grounding. , 2020, , .		15
983	Image Generation with the Enhanced Latent Code and Sub-pixel Sampling. Lecture Notes in Computer Science, 2020, , 387-398.	1.0	0
984	Visually Grounded Continual Learning of Compositional Phrases. , 2020, , .		2
985	Towards Partial Supervision for Generic Object Counting in Natural Scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 1604-1622.	9.7	2
987	Learning Object Permanence from Video. Lecture Notes in Computer Science, 2020, , 35-50.	1.0	7
988	The Devil Is in Classification: A Simple Framework for Long-Tail Instance Segmentation. Lecture Notes in Computer Science, 2020, , 728-744.	1.0	72
989	AiR: Attention with Reasoning Capability. Lecture Notes in Computer Science, 2020, , 91-107.	1.0	12

#	Article	IF	Citations
990	Fine-Grained Scene-Graph-to-Image Model Based on SAGAN. Lecture Notes in Computer Science, 2020, , 325-337.	1.0	0
991	Towards Unique and Informative Captioning of Images. Lecture Notes in Computer Science, 2020, , 629-644.	1.0	13
992	Cross-Modal Representation. , 2020, , 285-317.		1
993	Visual Relationship Detection with Contextual Information. Computers, Materials and Continua, 2020, 63, 1575-1589.	1.5	2
994	Toward General Scene Graph: Integration of Visual Semantic Knowledge with Entity Synset Alignment. , 2020, , .		1
995	AI4AR: An AI-Based Mobile Application for the Automatic Generation of AR Contents. Lecture Notes in Computer Science, 2020, , 273-288.	1.0	1
996	Vision to Language: Methods, Metrics and Datasets. Learning and Analytics in Intelligent Systems, 2020, , 9-62.	0.5	2
997	Dense Captioning Using Abstract Meaning Representation. Lecture Notes in Computer Science, 2020, , 450-465.	1.0	0
998	Visual Relation Grounding in Videos. Lecture Notes in Computer Science, 2020, , 447-464.	1.0	24
999	NODIS: Neural Ordinary Differential Scene Understanding. Lecture Notes in Computer Science, 2020, , 636-653.	1.0	6
1000	Image Generation from Layout via Pair-Wise RaGAN. Communications in Computer and Information Science, 2020, , 193-206.	0.4	1
1001	Towards Knowledge-Augmented Visual Question Answering. , 2020, , .		9
1002	Spatial Descriptions on a Functional-Geometric Spectrum: the Location of Objects. Lecture Notes in Computer Science, 2020, , 219-234.	1.0	1
1003	Focus Your Attention: A Focal Attention for Multimodal Learning. IEEE Transactions on Multimedia, 2022, 24, 103-115.	5.2	5
1004	Language-Driven Region Pointer Advancement for Controllable Image Captioning. , 2020, , .		4
1005	Bounding-Box Channels for Visual Relationship Detection. Lecture Notes in Computer Science, 2020, , 682-697.	1.0	6
1006	Generating Videos of Zero-Shot Compositions of Actions and Objects. Lecture Notes in Computer Science, 2020, , 382-401.	1.0	4
1007	Detecting Human-Object Interactions with Action Co-occurrence Priors. Lecture Notes in Computer Science, 2020, , 718-736.	1.0	45

#	Article	IF	Citations
1008	Learning to Scale Multilingual Representations for Vision-Language Tasks. Lecture Notes in Computer Science, 2020, , 197-213.	1.0	8
1009	Grounded Situation Recognition. Lecture Notes in Computer Science, 2020, , 314-332.	1.0	24
1010	Learning Visual Representations withÂCaption Annotations. Lecture Notes in Computer Science, 2020, , 153-170.	1.0	40
1011	Multiple Interaction Learning with Question-Type Prior Knowledge for Constraining Answer Search Space in Visual Question Answering. Lecture Notes in Computer Science, 2020, , 496-510.	1.0	1
1012	Decision-Making Systems Based on Semantic Image Analysis. Communications in Computer and Information Science, 2020, , 102-120.	0.4	0
1013	Unsupervised Keyword Extraction for Full-Sentence VQA. , 2020, , .		0
1014	Reducing Language Biases in Visual Question Answering with Visually-Grounded Question Encoder. Lecture Notes in Computer Science, 2020, , 18-34.	1.0	17
1015	Scene Graph Modification Based on Natural Language Commands. , 2020, , .		2
1016	Iterative Visual Relationship Detection via Commonsense Knowledge Graph. Lecture Notes in Computer Science, 2020, , 210-225.	1.0	2
1017	Jointly Linking Visual and Textual Entity Mentions with Background Knowledge. Lecture Notes in Computer Science, 2020, , 264-276.	1.0	1
1018	Image Description Generation Using Deep Learning. Lecture Notes in Networks and Systems, 2020, , 1239-1244.	0.5	0
1019	ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. Lecture Notes in Computer Science, 2020, , 422-440.	1.0	40
1020	Learning Visual Commonsense for Robust Scene Graph Generation. Lecture Notes in Computer Science, 2020, , 642-657.	1.0	27
1021	Detecting Anomalous Regions from an Image based on Deep Captioning. , 2020, , .		8
1022	Diverse and Relevant Visual Storytelling with Scene Graph Embeddings. , 2020, , .		6
1023	Multimodal Sentence Summarization via Multimodal Selective Encoding. , 2020, , .		12
1024	Multi-Source Domain Adaptation by Deep CockTail Networks. , 2020, , 213-233.		0
1025	CoNAN: A Complementary Neighboring-based Attention Network for Referring Expression Generation. , 2020, , .		5

#	Article	IF	Citations
1026	Stacked Attention Networks for Referring Expressions Comprehension. Computers, Materials and Continua, 2020, 65, 2529-2541.	1.5	1
1027	Adaptively Clustering-Driven Learning for Visual Relationship Detection. IEEE Transactions on Multimedia, 2021, 23, 4515-4525.	5.2	19
1028	Fashionpedia: Ontology, Segmentation, and an Attribute Localization Dataset. Lecture Notes in Computer Science, 2020, , 316-332.	1.0	41
1029	Spatially Aware Multimodal Transformers for TextVQA. Lecture Notes in Computer Science, 2020, , 715-732.	1.0	45
1030	Knowledge-Guided Multi-Label Few-Shot Learning for General Image Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 1371-1384.	9.7	73
1031	Integrating Domain Knowledge: Using Hierarchies to Improve Deep Classifiers. Lecture Notes in Computer Science, 2020, , 3-16.	1.0	7
1032	Scene Understanding Using Deep Neural Networksâ€"Objects, Actions, and Events: A Review. Advances in Intelligent Systems and Computing, 2020, , 223-231.	0.5	3
1033	Multi-view Visual Question Answering Dataset for Real Environment Applications. Lecture Notes in Computer Science, 2020, , 384-395.	1.0	0
1034	Injecting Prior Knowledge into Image Caption Generation. Lecture Notes in Computer Science, 2020, , 369-385.	1.0	3
1035	AABO: Adaptive Anchor Box Optimization for Object Detection via Bayesian Sub-sampling. Lecture Notes in Computer Science, 2020, , 560-575.	1.0	9
1036	Drawing Dreams. Lecture Notes in Computer Science, 2020, , 290-300.	1.0	0
1037	STL-CQA: Structure-based Transformers with Localization and Encoding for Chart Question Answering. , 2020, , .		17
1038	Beyond Language: Learning Commonsense from Images for Reasoning. , 2020, , .		0
1039	Guessing the Age of Acquisition of Italian Lemmas through Linear Regression. , 2020, , .		1
1040	Visual Social Relationship Recognition. International Journal of Computer Vision, 2020, 128, 1750-1764.	10.9	16
1041	APPROACH TO IMAGE ANALYSIS FOR COMPUTER VISION SYSTEMS. Doklady BGUIR, 2020, 18, 62-70.	0.1	0
1042	Memorize, Associate and Match: Embedding Enhancement via Fine-Grained Alignment for Image-Text Retrieval. IEEE Transactions on Image Processing, 2021, 30, 9193-9207.	6.0	21
1043	Zero-Shot Instance Segmentation. , 2021, , .		21

#	Article	IF	Citations
1044	M³P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training. , 2021, , .		30
1045	LayoutTransformer: Scene Layout Generation with Conceptual and Spatial Diversity. , 2021, , .		16
1046	VirTex: Learning Visual Representations from Textual Annotations. , 2021, , .		136
1047	Towards Accurate Text-based Image Captioning with Content Diversity Exploration. , 2021, , .		33
1048	Explicit Knowledge Incorporation for Visual Reasoning. , 2021, , .		12
1049	Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. , 2021, , .		92
1050	MR Image Super-Resolution with Squeeze and Excitation Reasoning Attention Network. , 2021, , .		47
1051	VLNâ†»BERT: A Recurrent Vision-and-Language BERT for Navigation. , 2021, , .		60
1052	Exemplar-Based Open-Set Panoptic Segmentation Network. , 2021, , .		22
1053	Learning Better Visual Dialog Agents with Pretrained Visual-Linguistic Representation. , 2021, , .		8
1054	Energy-Based Learning for Scene Graph Generation. , 2021, , .		76
1055	Learning Asynchronous and Sparse Human-Object Interaction in Videos. , 2021, , .		18
1056	Benchmarking Representation Learning for Natural World Image Collections. , 2021, , .		49
1057	Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling. , 2021, , .		238
1058	How Transferable are Reasoning Patterns in VQA?. , 2021, , .		14
1059	AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. , 2021, , .		31
1060	Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. , 2021, , .		159
1061	Siamese Natural Language Tracker: Tracking by Natural Language Descriptions with Siamese Trackers. , 2021, , .		18

#	Article	IF	Citations
1062	Domain-robust VQA with diverse datasets and methods but no target labels. , 2021, , .		10
1063	Hierarchical and Partially Observable Goal-driven Policy Learning with Goals Relational Graph. , 2021, , .		8
1064	VinVL: Revisiting Visual Representations in Vision-Language Models. , 2021, , .		317
1065	Kaleido-BERT: Vision-Language Pre-training on Fashion Domain. , 2021, , .		57
1066	Fully Convolutional Scene Graph Generation. , 2021, , .		46
1067	Human-like Controllable Image Captioning with Verb-specific Semantic Roles. , 2021, , .		33
1068	Bipartite Graph Network with Adaptive Message Passing for Unbiased Scene Graph Generation. , 2021, , .		85
1069	Learning to Predict Visual Attributes in the Wild. , 2021, , .		29
1070	Open-Vocabulary Object Detection Using Captions. , 2021, , .		96
1071	SceneGraphFusion: Incremental 3D Scene Graph Prediction from RGB-D Sequences. , 2021, , .		36
1072	TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. , 2021, , .		42
1073	Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation. , 2021, , .		34
1074	Context-Aware Layout to Image Generation with Enhanced Object Appearance. , 2021, , .		20
1075	Causal Attention for Vision-Language Tasks. , 2021, , .		57
1076	Linguistic Structures as Weak Supervision for Visual Scene Graph Generation. , 2021, , .		25
1077	Transitional Adaptation of Pretrained Models for Visual Storytelling. , 2021, , .		12
1078	Spoken Moments: Learning Joint Audio-Visual Representations from Video Descriptions. , 2021, , .		19
1079	Exploiting Edge-Oriented Reasoning for 3D Point-based Scene Graph Analysis. , 2021, , .		20

#	Article	IF	Citations
1080	Boosting Video Representation Learning with Multi-Faceted Integration. , 2021, , .		7
1081	UniT: Unified Knowledge Transfer for Any-shot Object Detection and Segmentation. , 2021, , .		13
1082	Room-and-Object Aware Knowledge Reasoning for Remote Embodied Referring Expression. , 2021, , .		29
1083	Relation-aware Instance Refinement for Weakly Supervised Visual Grounding. , 2021, , .		24
1084	ArtEmis: Affective Language for Visual Art. , 2021, , .		43
1085	Learning the Best Pooling Strategy for Visual Semantic Embedding. , 2021, , .		82
1086	Probabilistic Modeling of Semantic Ambiguity for Scene Graph Generation. , 2021, , .		35
1087	KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. , 2021, , .		59
1088	TAP: Text-Aware Pre-training for Text-VQA and Text-Caption. , 2021, , .		55
1089	General Multi-label Image Classification with Transformers. , 2021, , .		140
1090	Holistic 3D Scene Understanding from a Single Image with Implicit Representation. , 2021, , .		47
1091	Similarity-based calibration method for zero-shot recognition in multi-object scenes. , 2020, , .		0
1092	VTKEL. , 2020, , .		4
1093	Loss Optimised Video Captioning using Deep-Lstm,Attention Mechanism and Weighted Loss Metrices. , 2021, , .		1
1094	RSG-Net: Towards Rich Sematic Relationship Prediction for Intelligent Vehicle in Complex Environments. , 2021, , .		4
1095	A novel SPLIT-SIM approach for efficient image retrieval. Multimedia Systems, 0, , 1.	3.0	1
1096	A framework for visual question answering with the integration of scene-text using PHOCs and fisher vectors. Expert Systems With Applications, 2022, 190, 116159.	4.4	11
1097	CCMR: A Classic-enriched Connotation-aware Music Retrieval System on Social Media with Visual Inputs. Social Network Analysis and Mining, 2021, 11, 1.	1.9	3

#	Article	IF	Citations
1098	Regional Relation Modeling for Visual Place Recognition. , 2020, , .		3
1099	Lifelogging caption generation via fourth-person vision in a humanâ€“robot symbiotic environment. ROBOMECH Journal, 2020, 7, .	0.9	2
1100	Subgraph and object contextâ€masked network for scene graph generation. IET Computer Vision, 2020, 14, 546-553.	1.3	0
1101	Crosspower. , 2020, , .		16
1102	ROSMI: A Multimodal Corpus for Map-based Instruction-Giving. , 2020, , .		1
1103	Human-Object Interaction Detection. , 2020, , .		7
1104	"I Hope This Is Helpful". Proceedings of the ACM on Human-Computer Interaction, 2020, 4, 1-26.	2.5	12
1105	Error Analysis for Visual Question Answering. Studies in Computational Intelligence, 2021, , 283-292.	0.7	1
1106	Part-Aware Interactive Learning for Scene Graph Generation. , 2020, , .		9
1107	Visual Relation of Interest Detection. , 2020, , .		7
1108	Expressional Region Retrieval. , 2020, , .		1
1109	HOSE-Net: Higher Order Structure Embedded Network for Scene Graph Generation. , 2020, , .		9
1110	Look, Read and Feel: Benchmarking Ads Understanding with Multimodal Multitask Learning. , 2020, , .		9
1111	Activity-driven Weakly-Supervised Spatio-Temporal Grounding from Untrimmed Videos. , 2020, , .		10
1112	Transferrable Referring Expression Grounding with Concept Transfer and Context Inheritance. , 2020, , .		2
1113	Object-level Attention for Aesthetic Rating Distribution Prediction. , 2020, , .		17
1114	PCPL: Predicate-Correlation Perception Learning for Unbiased Scene Graph Generation. , 2020, , .		68
1115	Video Relation Detection via Multiple Hypothesis Association. , 2020, , .		21

#	Article	IF	Citations
1116	Hierarchical Scene Graph Encoder-Decoder for Image Paragraph Captioning. , 2020, , .		13
1117	Structural Semantic Adversarial Active Learning for Image Captioning. , 2020, , .		9
1118	Scene-Aware Context Reasoning for Unsupervised Abnormal Event Detection in Videos. , 2020, , .		38
1119	Visual-Semantic Graph Matching for Visual Grounding. , 2020, , .		18
1120	Webly Supervised Image Classification with Metadata: Automatic Noisy Label Correction via Visual-Semantic Graph. , 2020, , .		6
1121	Deep Multimodal Neural Architecture Search. , 2020, , .		42
1122	Bridging the Gap between Vision and Language Domains for Improved Image Captioning. , 2020, , .		7
1123	One-shot Scene Graph Generation. , 2020, , .		16
1124	AdaHGNN. , 2020, , .		24
1125	Video Relation Detection with Trajectory-aware Multi-modal Features. , 2020, , .		12
1126	Kinetics and Scene Features for Intent Detection. , 2020, , .		1
1127	A Scientometric Visualization Analysis of Image Captioning Research From 2010 to 2020. IEEE Access, 2021, 9, 156799-156817.	2.6	1
1128	Detecting and Tracking Small and Dense Moving Objects in Satellite Videos: A Benchmark. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-18.	2.7	27
1129	AI-Based Detection, Classification and Prediction/Prognosis in Medical Imaging. PET Clinics, 2022, 17, 183-212.	1.5	31
1130	Visual Navigation via Reinforcement Learning and Relational Reasoning. , 2021, , .		1
1131	BERTHop: An Effective Vision-and-Language Model for Chest X-ray Disease Diagnosis. , 2021, , .		5
1132	Multimodal Continuous Visual Attention Mechanisms. , 2021, , .		3
1133	Fine-Grained Visual Textual Alignment for Cross-Modal Retrieval Using Transformer Encoders. ACM Transactions on Multimedia Computing, Communications and Applications, 2021, 17, 1-23.	3.0	65

#	Article	IF	Citations
1134	A thorough review of models, evaluation metrics, and datasets on image captioning. IET Image Processing, 2022, 16, 311-332.	1.4	8
1135	Bi-Directional Co-Attention Network for Image Captioning. ACM Transactions on Multimedia Computing, Communications and Applications, 2021, 17, 1-20.	3.0	18
1136	Generating visual explanations with natural language. Applied AI Letters, 2021, 2, .	1.4	3
1137	Attention-Guided Image Captioning through Word Information. Sensors, 2021, 21, 7982.	2.1	1
1138	Cross-Modal Hybrid Feature Fusion for Image-Sentence Matching. ACM Transactions on Multimedia Computing, Communications and Applications, 2021, 17, 1-23.	3.0	25
1139	Multi-level, multi-modal interactions for visual question answering over text in images. World Wide Web, 2022, 25, 1607-1623.	2.7	1
1140	What Does a Language-And-Vision Transformer See: The Impact of Semantic Information on Visual Representations. Frontiers in Artificial Intelligence, 2021, 4, 767971.	2.0	0
1142	RUArt: A Novel Text-Centered Solution for Text-Based Visual Question Answering. IEEE Transactions on Multimedia, 2023, 25, 1-12.	5.2	10
1143	Visual Intelligence through Human Interaction. Human-computer Interaction Series, 2021, , 257-314.	0.4	3
1144	Imageability- and Length-Controllable Image Captioning. IEEE Access, 2021, 9, 162951-162961.	2.6	7
1145	VisualSem: a high-quality knowledge graph for vision and language. , 2021, , .		11
1146	Visually Grounded Reasoning across Languages and Cultures. , 2021, , .		12
1147	A Survey on Computer Vision Architectures for Large Scale Image Classification using Deep Learning. International Journal of Advanced Computer Science and Applications, 2021, 12, .	0.5	1
1148	VQA-MHUG: A Gaze Dataset to Study Multimodal Neural Attention in Visual Question Answering. , 2021, , .		5
1149	MIRTT: Learning Multimodal Interaction Representations from Trilinear Transformers for Visual Question Answering. , 2021, , .		3
1150	Cross-Modal Retrieval Augmentation for Multi-Modal Classification. , 2021, , .		6
1152	Multi-Scale Fine-Grained Alignments for Image and Sentence Matching. IEEE Transactions on Multimedia, 2023, 25, 543-556.	5.2	28
1153	Deconfounded Image Captioning: A Causal Retrospect. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, PP, 1-1.	9.7	18

#	Article	IF	Citations
1154	Structured Multimodal Attentions for TextVQA. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44, 9603-9614.	9.7	20
1155	Learning Language to Symbol and Language to Vision Mapping for Visual Grounding. SSRN Electronic Journal, 0, , .	0.4	0
1156	Enhancing Visual Dialog Questioner with Entity-based Strategy Learning and Augmented Guesser. , 2021, , .		5
1157	Data Efficient Masked Language Modeling for Vision and Language. , 2021, , .		9
1158	Multi-Stream Feature Refinement Network for Human Object Interaction Detection. SSRN Electronic Journal, 0, , .	0.4	0
1159	Object Goal Visual Navigation Using Semantic Spatial Relationships. Lecture Notes in Computer Science, 2021, , 77-88.	1.0	1
1160	Improving Pre-trained Vision-and-Language Embeddings for Phrase Grounding. , 2021, , .		3
1161	Semantically Similarity-Wise Dual-Branch Network for Scene Graph Generation. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 4573-4583.	5.6	8
1162	Rethinking Denoised Auto-Encoding in Language Pre-Training. , 2021, , .		2
1163	Semantic Novelty Detection in Natural Language Descriptions. , 2021, , .		4
1164	Transformers in Vision: A Survey. ACM Computing Surveys, 2022, 54, 1-41.	16.1	754
1165	Depth and Video Segmentation Based Visual Attention for Embodied Question Answering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 6807-6819.	9.7	2
1166	Deep Learning Approaches for Fashion Knowledge Extraction From Social Media: A Review. IEEE Access, 2022, 10, 1545-1576.	2.6	12
1167	Gated attention fusion network for multimodal sentiment classification. Knowledge-Based Systems, 2022, 240, 108107.	4.0	33
1168	Word-Region Alignment-Guided Multimodal Neural Machine Translation. IEEE/ACM Transactions on Audio Speech and Language Processing, 2022, 30, 244-259.	4.0	7
1169	Sequential Transformer via an Outside-In Attention for image captioning. Engineering Applications of Artificial Intelligence, 2022, 108, 104574.	4.3	14
1170	Region-attentive multimodal neural machine translation. Neurocomputing, 2022, 476, 1-13.	3.5	9
1171	6. Datasets for Vision and Language Research. Kyokai Joho Imeji Zasshi/Journal of the Institute of Image Information and Television Engineers, 2018, 72, 672-675.	0.0	0

#	Article	IF	Citations
1172	From Machine Translated NLI Corpus to Universal Sentence Representations in Czech. , 0, , .		0
1173	Learning Object Attributes with Category-Free Grounded Language from Deep Featurization. , 2020, , .		1
1174	A Bottom-up Framework for Construction of Structured Semantic 3D Scene Graph. , 2020, , .		4
1175	A Bottom-up Paradigm for Traffic Scene Graph Representation. , 2020, , .		3
1176	Multimodal Aggregation Approach for Memory Vision-Voice Indoor Navigation with Meta-Learning. , 2020, , .		6
1177	Retargetable AR: Context-aware Augmented Reality in Indoor Scenes based on 3D Scene Graph. , 2020, , .		23
1178	Graph Self-Attention Network for Image Captioning. , 2020, , .		1
1180	Multimodal Query-Guided Object Localization. SSRN Electronic Journal, 0, , .	0.4	0
1182	AttrLostGAN: Attribute Controlled Image Synthesis fromÂReconfigurable Layout andÂStyle. Lecture Notes in Computer Science, 2021, , 361-375.	1.0	9
1183	Modular Attention Network based on Language Model for Referring Expression. , 2021, , .		0
1184	A Lightweight Visual Question Answering Model based on Semantic Similarity. , 2021, , .		1
1185	Unsupervised Traffic Scene Generation with Synthetic 3D Scene Graphs. , 2021, , .		5
1186	Text Pared into Scene Graph for Diverse Image Generation. , 2021, , .		0
1187	Super Visual Semantic Embedding for Cross-Modal Image-Text Retrieval. , 2021, , .		0
1188	Learning Depth Information in Layout for Sketch Generation from Scene Graph. , 2021, , .		0
1189	Softmax Pooling for Super Visual Semantic Embedding. , 2021, , .		2
1190	Lightweight Visual Question Answering using Scene Graphs. , 2021, , .		5
1191	Multi-Agent Embodied Visual Semantic Navigation With Scene Prior Knowledge. IEEE Robotics and Automation Letters, 2022, 7, 3154-3161.	3.3	10

#	Article	IF	Citations
1192	Deep Modular Bilinear Attention Network for Visual Question Answering. Sensors, 2022, 22, 1045.	2.1	7
1193	Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training. Neurocomputing, 2022, 490, 163-180.	3.5	18
1194	#PraCegoVer: A Large Dataset for Image Captioning in Portuguese. Data, 2022, 7, 13.	1.2	2
1195	A survey on visual transfer learning using knowledge graphs. Semantic Web, 2022, 13, 477-510.	1.1	9
1196	Human-Centric Image Captioning. Pattern Recognition, 2022, 126, 108545.	5.1	15
1198	Efficient Local Image Descriptors Learned With Autoencoders. IEEE Access, 2022, 10, 221-235.	2.6	1
1199	Analysis of Image generation using Scene graphs. , 2022, , .		0
1200	Learning Scene-Aware Spatio-Temporal GNNs for Few-Shot Early Action Prediction. IEEE Transactions on Multimedia, 2023, 25, 2061-2073.	5.2	3
1201	Deep relational self-Attention networks for scene graph generation. Pattern Recognition Letters, 2022, 153, 200-206.	2.6	4
1202	A Large Visual Question Answering Dataset forÂCultural Heritage. Lecture Notes in Computer Science, 2022, , 193-197.	1.0	2
1203	A multi-level semantic web for hard-to-specify domain concept, Pedestrian, in ML-based software. Requirements Engineering, 2022, 27, 161-182.	2.1	5
1204	Without detection: Twoâ€step clustering features with localâ€“global attention for image captioning. IET Computer Vision, 2022, 16, 280-294.	1.3	2
1205	Perspectives and Prospects on Transformer Architecture for Cross-Modal Tasks with Language and Vision. International Journal of Computer Vision, 2022, 130, 435-454.	10.9	17
1206	Repurposing existing deep networks for caption and aesthetic-guided image cropping. Pattern Recognition, 2022, 126, 108485.	5.1	7
1207	A comprehensive review of the video-to-text problem. Artificial Intelligence Review, 2022, 55, 4165-4239.	9.7	5
1208	Object recognition datasets and challenges: A review. Neurocomputing, 2022, 495, 129-152.	3.5	18
1209	Scene Graph Generation Using Depth, Spatial, and Visual Cues in 2D Images. IEEE Access, 2022, 10, 1968-1978.	2.6	1
1210	Unified Adaptive Relevance Distinguishable Attention Network for Image-Text Matching. IEEE Transactions on Multimedia, 2023, 25, 1320-1332.	5.2	15

#	Article	IF	Citations
1211	A study on video semantics; overview, challenges, and applications. Multimedia Tools and Applications, 2022, 81, 6849-6897.	2.6	7
1212	Relational attention-based Markov logic network for visual navigation. Journal of Supercomputing, 0, , 1.	2.4	0
1213	REGRAD: A Large-Scale Relational Grasp Dataset for Safe and Object-Specific Robotic Grasping in Clutter. IEEE Robotics and Automation Letters, 2022, 7, 2929-2936.	3.3	16
1214	Artificial Intelligence Models Do Not Ground Negation, Humans Do. GuessWhat?! Dialogues as a Case Study. Frontiers in Big Data, 2021, 4, 736709.	1.8	2
1215	Survey: Transformer based video-language pre-training. AI Open, 2022, 3, 1-13.	9.1	17
1216	A Generative Answer Aggregation Model for Sentence-Level Crowdsourcing Tasks. IEEE Transactions on Knowledge and Data Engineering, 2023, 35, 3299-3312.	4.0	1
1217	Which Apple Keeps Which Doctor Away? Colorful Word Representations With Visual Oracles. IEEE/ACM Transactions on Audio Speech and Language Processing, 2022, 30, 49-59.	4.0	0
1218	Towards Open-Set Scene Graph Generation With Unknown Objects. IEEE Access, 2022, 10, 11574-11583.	2.6	1
1219	The semantic typology of visually grounded paraphrases. Computer Vision and Image Understanding, 2022, 215, 103333.	3.0	2
1220	Visual-Text Reference Pretraining Model for Image Captioning. Computational Intelligence and Neuroscience, 2022, 2022, 1-10.	1.1	1
1221	Learning 3D Semantic Scene Graphs with Instance Embeddings. International Journal of Computer Vision, 2022, 130, 630-651.	10.9	4
1222	A Comprehensive Survey of Scene Graphs: Generation and Application. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 1-26.	9.7	75
1223	Visual Relationship Detection: A Survey. IEEE Transactions on Cybernetics, 2022, 52, 8453-8466.	6.2	10
1224	Discriminative Style Learning for Cross-Domain Image Captioning. IEEE Transactions on Image Processing, 2022, 31, 1723-1736.	6.0	9
1225	Diverse and styled image captioning using singular value decompositionâ€based mixture of recurrent experts. Concurrency Computation Practice and Experience, 0, , .	1.4	0
1226	Aligning and linking entity mentions in image, text, and knowledge base. Data and Knowledge Engineering, 2022, 138, 101975.	2.1	5
1227	Examining the influence of linguistic characteristics of online managerial response on return customersâ€™ change in satisfaction with hotels. International Journal of Hospitality Management, 2022, 102, 103146.	5.3	11
1228	Global-Guided Asymmetric Attention Network for Image-Text Matching. Neurocomputing, 2022, 481, 77-90.	3.5	3

#	Article	IF	Citations
1229	Dual Global Enhanced Transformer for image captioning. Neural Networks, 2022, 148, 129-141.	3.3	41
1230	Image-Text Embedding Learning via Visual and Textual Semantic Reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 641-656.	9.7	33
1231	From Show to Tell: A Survey on Deep Learning-Based Image Captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 539-559.	9.7	81
1232	Instance-level Object relation module for one-stage Object Detection. Multimedia Tools and Applications, 2022, 81, 8617-8632.	2.6	0
1233	Multiple Context Learning Networks for Visual Question Answering. Scientific Programming, 2022, 2022, 1-11.	0.5	1
1234	Pre-training Model Based on Parallel Cross-Modality Fusion Layer. PLoS ONE, 2022, 17, e0260784.	1.1	0
1235	You should know more: Learning external knowledge for visual dialog. Neurocomputing, 2022, 488, 54-65.	3.5	2
1236	OpenPubSub: Supporting Large Semantic Content Spaces in Peer-to-Peer Publish/Subscribe Systems for the Internet of Multimedia Things. IEEE Internet of Things Journal, 2022, 9, 17640-17659.	5.5	4
1237	Image-Text Matching with Intra-Modal Knowledge. SSRN Electronic Journal, 0, , .	0.4	0
1239	Multimodal Sentiment Analysis With Image-Text Interaction Network. IEEE Transactions on Multimedia, 2023, 25, 3375-3385.	5.2	21
1240	Improving Conversation Task with Visual Scene Dataset. Journal of Natural Language Processing, 2022, 29, 166-186.	0.1	0
1241	Tackling the Challenges in Scene Graph Generation With Local-to-Global Interactions. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34, 9713-9726.	7.2	3
1242	A Text-Guided Generation and Refinement Model for Image Captioning. IEEE Transactions on Multimedia, 2023, 25, 2966-2977.	5.2	4
1243	Towards a solver-aware systems architecting framework: leveraging experts, specialists and the crowd to design innovative complex systems. Design Science, 2022, 8, .	1.1	1
1244	Exploring Implicit and Explicit Relations with the Dual Relation-Aware Network for Image Captioning. Lecture Notes in Computer Science, 2022, , 97-108.	1.0	0
1245	Inductive Biases for Low Data VQA: A Data Augmentation Approach. , 2022, , .		2
1246	Adaptive Path Selection for Dynamic Image Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 5762-5775.	5.6	21
1248	Improve Image Captioning by Estimating the Gazing Patterns from the Caption. , 2022, , .		3

#	Article	IF	Citations
1249	Temporal Moment Localization via Natural Language by Utilizing Video Question Answers as a Special Variant and Bypassing NLP for Corpora. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 6174-6185.	5.6	8
1250	Zero-Shot Predicate Prediction for Scene Graph Parsing. IEEE Transactions on Multimedia, 2023, 25, 3140-3153.	5.2	1
1251	A core region captioning framework for automatic video understanding in story video contents. International Journal of Engineering Business Management, 2022, 14, 184797902210781.	2.1	2
1252	Hierarchical Feature Aggregation Based on Transformer for Image-Text Matching. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 6437-6447.	5.6	20
1253	Video Visual Relation Detection via 3D Convolutional Neural Network. IEEE Access, 2022, 10, 23748-23756.	2.6	3
1254	Divide and Conquer: Subset Matching for Scene Graph Generation in Complex Scenes. IEEE Access, 2022, 10, 39069-39079.	2.6	0
1255	SST: Spatial and Semantic Transformers for Multi-Label Image Recognition. IEEE Transactions on Image Processing, 2022, 31, 2570-2583.	6.0	31
1256	InfographicVQA. , 2022, , .		19
1257	Topic scene graphs for image captioning. IET Computer Vision, 2022, 16, 364-375.	1.3	3
1258	Improving visual question answering by combining scene-text information. Multimedia Tools and Applications, 2022, 81, 12177-12208.	2.6	9
1259	Fixed-Size Objects Encoding for Visual Relationship Detection. Neural Processing Letters, 0, , 1.	2.0	0
1260	MAA-PTG: multimodal aspect-aware product title generation. Journal of Intelligent Information Systems, 0, , 1.	2.8	1
1261	A Deep Reinforcement Learning Approach with Visual Semantic Navigation with Memory for Mobile Robots in Indoor Home Context. Journal of Intelligent and Robotic Systems: Theory and Applications, 2022, 104, .	2.0	5
1262	Different computational relations in language are captured by distinct brain systems. Cerebral Cortex, 2023, 33, 997-1013.	1.6	8
1263	Semantic association enhancement transformer with relative position for image captioning. Multimedia Tools and Applications, 2022, 81, 21349-21367.	2.6	4
1264	Caption TLSTMs: combining transformer with LSTMs for image captioning. International Journal of Multimedia Information Retrieval, 2022, 11, 111-121.	3.6	5
1265	Benchmarking the Robustness of Object Detection Based on Near-Real Military Scenes. Wireless Communications and Mobile Computing, 2022, 2022, 1-12.	0.8	1
1266	Computer Science Diagram Understanding with Topology Parsing. ACM Transactions on Knowledge Discovery From Data, 2022, 16, 1-20.	2.5	2

#	Article	IF	Citations
1267	Semanticâ€meshed and contentâ€guided transformer for image captioning. IET Computer Vision, 2022, 16, 431-444.	1.3	4
1268	A novel deep translated attention hashing for cross-modal retrieval. Multimedia Tools and Applications, 2022, 81, 26443-26461.	2.6	2
1269	Removing Partial Mismatches in Unsupervised Image Captioning. Transactions of the Japanese Society for Artificial Intelligence, 2022, 37, H-L82_1-12.	0.1	0
1270	ARM3D: Attention-based relation module for indoor 3D object detection. Computational Visual Media, 2022, 8, 395-414.	10.8	5
1271	COIN: Counterfactual Image Generation for Visual Question Answering Interpretation. Sensors, 2022, 22, 2245.	2.1	3
1272	Improving Target-driven Visual Navigation with Attention on 3D Spatial Relationships. Neural Processing Letters, 2022, 54, 3979-3998.	2.0	9
1273	AutoPoD-Mobileâ€"Semi-Automated Data Population Using Case-like Scenarios for Training and Validation in Mobile Forensics. Forensic Sciences, 2022, 2, 302-320.	0.8	0
1274	Boosting Scene Graph Generation with Visual Relation Saliency. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-17.	3.0	6
1275	Instance-sequence reasoning for video question answering. Frontiers of Computer Science, 2022, 16, 1.	1.6	9
1276	Adversarial attack and defense technologies in natural language processing: A survey. Neurocomputing, 2022, 492, 278-307.	3.5	21
1277	Reasoning With Scene Graphs for Robot Planning Under Partial Observability. IEEE Robotics and Automation Letters, 2022, 7, 5560-5567.	3.3	13
1278	Improving indoor visual navigation generalization with scene priors and Markov relational reasoning. Applied Intelligence, 2022, 52, 17600-17613.	3.3	1
1279	An analysis of graph convolutional networks and recent datasets for visual question answering. Artificial Intelligence Review, 2022, 55, 6277-6300.	9.7	14
1280	Geometry Attention Transformer with position-aware LSTMs for image captioning. Expert Systems With Applications, 2022, 201, 117174.	4.4	20
1281	Uni-EDEN: Universal Encoder-Decoder Network by Multi-Granular Vision-Language Pre-training. ACM Transactions on Multimedia Computing, Communications and Applications, 2022, 18, 1-16.	3.0	3
1282	One-shot Video Graph Generation for Explainable Action Reasoning. Neurocomputing, 2022, 488, 212-225.	3.5	6
1283	SemanticPeer: A distributional semantic peer-to-peer lookup protocol for large content spaces at internet-scale. Future Generation Computer Systems, 2022, 132, 239-253.	4.9	3
1284	Multimodal high-order relational network for vision-and-language tasks. Neurocomputing, 2022, 492, 62-75.	3.5	2

#	Article	IF	Citations
1285	Interactive Re-ranking via Object Entropy-Guided Question Answering for Cross-Modal Image Retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 2022, 18, 1-17.	3.0	5
1286	Progress of image caption: modelling, datasets, and evaluation. , 2021, , .		0
1287	Procrustean Training for Imbalanced Deep Learning. , 2021, , .		8
1288	Segmentation-grounded Scene Graph Generation. , 2021, , .		20
1289	Unconditional Scene Graph Generation. , 2021, , .		6
1290	Generative Compositional Augmentations for Scene Graph Prediction. , 2021, , .		11
1291	Exploiting Scene Graphs for Human-Object Interaction Detection. , 2021, , .		9
1292	Weakly Supervised Relative Spatial Reasoning for Visual Question Answering. , 2021, , .		6
1293	Telling the What while Pointing to the Where: Multimodal Queries for Image Retrieval. , 2021, , .		11
1294	Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. , 2021, , .		227
1295	Calibrating Concepts and Operations: Towards Symbolic Reasoning on Real Images. , 2021, , .		6
1296	Compressing Visual-linguistic Model via Knowledge Distillation. , 2021, , .		16
1297	Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation. , 2021, , .		15
1298	YouRefIt: Embodied Reference Understanding with Language and Gesture. , 2021, , .		11
1299	Graph-to-3D: End-to-End Generation and Manipulation of 3D Scenes Using Scene Graphs. , 2021, , .		12
1300	Visual Scene Graphs for Audio Source Separation. , 2021, , .		9
1301	Context-aware Scene Graph Generation with Seq2Seq Transformers. , 2021, , .		35
1302	Panoptic Narrative Grounding. , 2021, , .		7

#	Article	IF	Citations
1303	Unified Graph Structured Models for Video Understanding. , 2021, , .		17
1304	Exploring Visual Engagement Signals for Representation Learning. , 2021, , .		4
1305	Few-Shot Visual Relationship Co-Localization. , 2021, , .		1
1306	Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models. , 2021, , .		15
1307	Target Adaptive Context Aggregation for Video Scene Graph Generation. , 2021, , .		30
1308	Joint Visual Semantic Reasoning: Multi-Stage Decoder for Text Recognition. , 2021, , .		36
1309	Topic Scene Graph Generation by Attention Distillation from Caption. , 2021, , .		11
1310	Learning Generative Models of Textured 3D Meshes from Real-World Images. , 2021, , .		15
1311	SAT: 2D Semantics Assisted Training for 3D Visual Grounding. , 2021, , .		19
1312	Detecting Persuasive Atypicality by Modeling Contextual Compatibility. , 2021, , .		0
1313	Hierarchical Object-to-Zone Graph for Object Navigation. , 2021, , .		24
1314	Context Reasoning Attention Network for Image Super-Resolution. , 2021, , .		36
1315	Interpretable Visual Reasoning via Induced Symbolic Space. , 2021, , .		5
1316	Airbert: In-domain Pretraining for Vision-and-Language Navigation. , 2021, , .		44
1317	The Road to Know-Where: An Object-and-Room Informed Sequential BERT for Indoor Vision-Language Navigation. , 2021, , .		30
1318	HAIR: Hierarchical Visual-Semantic Relational Reasoning for Video Question Answering. , 2021, , .		31
1319	Ask&Confirm: Active Detail Enriching for Cross-Modal Retrieval with Partial Query. , 2021, , .		9
1320	Hierarchical Graph Attention Network for Few-shot Visual-Semantic Learning. , 2021, , .		3

#	Article	IF	Citations
1321	Zero-shot Natural Language Video Localization. , 2021, , .		13
1322	Whoâ€™s Waldo? Linking People Across Text and Images. , 2021, , .		2
1323	A Simple Baseline for Weakly-Supervised Scene Graph Generation. , 2021, , .		18
1324	MosaicOS: A Simple and Effective Use of Object-Centric Images for Long-Tailed Object Detection. , 2021, , .		22
1325	Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining. , 2021, , .		28
1326	Self-Supervised Real-to-Sim Scene Generation. , 2021, , .		13
1327	Visual Relationship Detection Using Part-and-Sum Transformers with Composite Queries. , 2021, , .		15
1328	Spatial-Temporal Transformer for Dynamic Scene Graph Generation. , 2021, , .		54
1329	Visual Distant Supervision for Scene Graph Generation. , 2021, , .		18
1330	Image Synthesis from Layout with Locality-Aware Mask Adaption. , 2021, , .		11
1331	Wasserstein Coupled Graph Learning for Cross-Modal Retrieval. , 2021, , .		10
1332	Detecting Human-Object Relationships in Videos. , 2021, , .		13
1333	Exploring Long Tail Visual Relationship Recognition with Large Vocabulary. , 2021, , .		12
1334	From General to Specific: Informative Scene Graph Generation via Balance Adjustment. , 2021, , .		49
1335	Detector-Free Weakly Supervised Grounding by Separation. , 2021, , .		8
1336	UniT: Multimodal Multitask Learning with a Unified Transformer. , 2021, , .		112
1337	Learning to Generate Scene Graph from Natural Language Supervision. , 2021, , .		32
1338	Visual-Textual Attentive Semantic Consistency for Medical Report Generation. , 2021, , .		4

#	Article	IF	Citations
1339	Auto-Parsing Network for Image Captioning and Visual Question Answering. , 2021, , .		16
1340	Edge-cloud Collaborative Architecture for Scene Graph Application System. , 2021, , .		0
1341	Cross-Modal Visual Question Answering for Remote Sensing Data: The International Conference on Digital Image Computing: Techniques and Applications (DICTA 2021). , 2021, , .		3
1342	GNES: Learning to Explain Graph Neural Networks. , 2021, , .		13
1343	Conceptual Intelligent Model for Visual Question Answering Using Attention Mechanism and Relational Reasoning. , 2021, , .		0
1344	Medical Scene Graphs and Reasoning. , 2021, , .		0
1345	Detecting Human-to-Human-or-Object (H²O) Interactions with DIABOLO. , 2021, , .		1
1346	Survey of Visual-Semantic Embedding Methods for Zero-Shot Image Retrieval. , 2021, , .		1
1347	Keyword-aware Multi-modal Enhancement Attention for Video Question Answering. , 2021, , .		0
1348	Graph Structural Attention and Increased Global Attention for Image Captioning. , 2021, , .		0
1349	Description Region Expansion-based relationship-oriented dense image captioning model. Journal of Advanced Marine Engineering and Technology, 2021, 45, 434-441.	0.1	1
1351	Optimal Deep Neural Network-Based Model for Answering Visual Medical Question. Cybernetics and Systems, 2022, 53, 403-424.	1.6	14
1352	Handwritten Mathematical Expression Recognition with Self-Attention. , 2021, , .		1
1353	Language Based Image Quality Assessment. , 2021, , .		2
1354	Deformable Geometry based Semantic Reconstruction from Scene Graphs. , 2021, , .		2
1355	DRSGN: Dual Revised Semantic Graph Structured Network for Image-Text Matching. , 2021, , .		0
1357	Explainable AI Methods - A Brief Overview. Lecture Notes in Computer Science, 2022, , 13-38.	1.0	77
1358	Learning language to symbol and language to vision mapping for visual grounding. Image and Vision Computing, 2022, , 104451.	2.7	1

#	Article	IF	Citations
1361	Aligned visual semantic scene graph for image captioning. Displays, 2022, 74, 102210.	2.0	10
1362	Exploring correlation of relationship reasoning for scene graph generation. International Journal of Machine Learning and Cybernetics, 0, , .	2.3	1
1363	Sparse co-attention visual question answering networks based on thresholds. Applied Intelligence, 2023, 53, 586-600.	3.3	11
1367	Towards Lightweight Transformer Via Group-Wise Transformation for Vision-and-Language Tasks. IEEE Transactions on Image Processing, 2022, 31, 3386-3398.	6.0	20
1370	Reinforced Causal Explainer for Graph Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 2297-2309.	9.7	8
1371	Deep Residual Weight-Sharing Attention Network With Low-Rank Attention for Visual Question Answering. IEEE Transactions on Multimedia, 2023, 25, 4282-4295.	5.2	5
1373	Visualizable or Non-visualizable? Exploring the Visualizability of Concepts in Multi-modal Knowledge Graph. Lecture Notes in Computer Science, 2022, , 180-187.	1.0	1
1375	Multi-Branch Distance-Sensitive Self-Attention Network for Image Captioning. IEEE Transactions on Multimedia, 2023, 25, 3962-3974.	5.2	4
1377	Scene Graph Refinement Network for Visual Question Answering. IEEE Transactions on Multimedia, 2023, 25, 3950-3961.	5.2	10
1378	Effective Pre-Training Method and Its Compositional Intelligence for Image Captioning. Sensors, 2022, 22, 3433.	2.1	2
1379	BPI-MVQA: a bi-branch model for medical visual question answering. BMC Medical Imaging, 2022, 22, 79.	1.4	9
1380	Clustering-Based Semi-Supervised Cross-Modal Retrieval Using Scene Graph. Journal of Circuits, Systems and Computers, 0, , .	1.0	0
1381	ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. , 2022, , .		11
1382	Visual context embeddings for zero-shot recognition. , 2022, , .		0
1383	Improving Cross-Modal Understanding in Visual Dialog Via Contrastive Learning. , 2022, , .		8
1384	Multi-View Visual Relationship Detection with Estimated Depth Map. Applied Sciences (Switzerland), 2022, 12, 4674.	1.3	0
1385	Informative Attention Supervision for Grounded Video Description. , 2022, , .		2
1386	PMP-NET: Rethinking Visual Context for Scene Graph Generation. , 2022, , .		1

#	Article	IF	Citations
1387	Multi-stream feature refinement network for human object interaction detection. Journal of Visual Communication and Image Representation, 2022, 86, 103529.	1.7	8
1388	Mixed Knowledge Relation Transformer for Image Captioning. , 2022, , .		1
1389	Position-aware image captioning with spatial relation. Neurocomputing, 2022, 497, 28-38.	3.5	5
1390	A survey on multimodal-guided visual content synthesis. Neurocomputing, 2022, 497, 110-128.	3.5	4
1391	Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing, 2023, 14, 1966-1978.	5.7	19
1392	System of Robot Learning from Multi-Modal Demonstration and Natural Language Instruction. Procedia CIRP, 2022, 107, 914-919.	1.0	3
1393	Image-Caption Pair Replacement Algorithm towards Semi-supervised Novel Object Captioning. , 2022, , .		0
1394	A Survey on Long-Tailed Visual Recognition. International Journal of Computer Vision, 2022, 130, 1837-1872.	10.9	38
1395	Enabling Improved Learning Capability of Industrial Robots with Knowledge Graph Towards Intelligent Digital Twins. , 2022, , .		1
1396	Research on the Image Description Algorithm of Double-Layer LSTM Based on Adaptive Attention Mechanism. Mathematical Problems in Engineering, 2022, 2022, 1-9.	0.6	0
1397	Visual context learning based on textual knowledge for imageâ€“text retrieval. Neural Networks, 2022, 152, 434-449.	3.3	5
1398	Visual Cluster Grounding for Image Captioning. IEEE Transactions on Image Processing, 2022, 31, 3920-3934.	6.0	11
1399	xGQA: Cross-Lingual Visual Question Answering. , 2022, , .		5
1400	Image Captioning with X-Gelu Activated Xgl Transformer. SSRN Electronic Journal, 0, , .	0.4	0
1401	Keyword Localisation in Untranscribed Speech Using Visually Grounded Speech Models. IEEE Journal on Selected Topics in Signal Processing, 2022, 16, 1454-1466.	7.3	2
1402	Regularizing Visual Semantic Embedding With Contrastive Learning for Image-Text Matching. IEEE Signal Processing Letters, 2022, 29, 1332-1336.	2.1	8
1403	Efficient Image and Sentence Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, , 1-1.	9.7	0
1404	Expressive Scene Graph Generation Using Commonsense Knowledge Infusion forÂVisual Understanding andÂReasoning. Lecture Notes in Computer Science, 2022, , 93-112.	1.0	4

#	Article	IF	Citations
1405	Weakly-Supervised Video Object Grounding via Causal Intervention. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, , 1-1.	9.7	3
1406	Matching Visual Features to Hierarchical Semantic Topics for Image Paragraph Captioning. International Journal of Computer Vision, 2022, 130, 1920-1937.	10.9	5
1407	Multi-Modal Alignment of Visual Question Answering Based on Multi-Hop Attention Mechanism. Electronics (Switzerland), 2022, 11, 1778.	1.8	4
1408	Recognition and statistical analysis of coastal marine aquacultural cages based on R3Det single-stage detector: A case study of Fujian Province, China. Ocean and Coastal Management, 2022, 225, 106244.	2.0	7
1409	Bilinear pooling in video-QA: empirical challenges and motivational drift from neurological parallels. PeerJ Computer Science, 0, 8, e974.	2.7	0
1410	Towards an effective model for lung disease classification. Applied Soft Computing Journal, 2022, 124, 109077.	4.1	4
1413	Hateful Meme Prediction Model Using Multimodal Deep Learning. , 2021, , .		2
1414	LiVLR: A Lightweight Visual-Linguistic Reasoning Framework for Video Question Answering. IEEE Transactions on Multimedia, 2023, 25, 5002-5013.	5.2	4
1415	Bypass Network for Semantics Driven Image Paragraph Captioning. SSRN Electronic Journal, 0, , .	0.4	0
1416	SRSG and S2SG: A Model and a Dataset for Scene Graph Generation of Remote Sensing Images From Segmentation Results. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-11.	2.7	0
1417	Predicate Correlation Learning for Scene Graph Generation. IEEE Transactions on Image Processing, 2022, 31, 4173-4185.	6.0	8
1418	Word2Pix: Word to Pixel Cross-Attention Transformer in Visual Grounding. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35, 1523-1533.	7.2	4
1419	SentiStory: A Multi-Layered Sentiment-Aware Generative Model for Visual Storytelling. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 8051-8064.	5.6	3
1420	KTN: Knowledge Transfer Network for Learning Multiperson 2D-3D Correspondences. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 7732-7745.	5.6	5
1421	Video Pivoting Unsupervised Multi-Modal Machine Translation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, , 1-15.	9.7	21
1422	Double-Stream Position Learning Transformer Network for Image Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32, 7706-7718.	5.6	15
1423	Knowing What to Learn: A Metric-Oriented Focal Mechanism for Image Captioning. IEEE Transactions on Image Processing, 2022, 31, 4321-4335.	6.0	12
1424	LANBIQUE: LANguage-based Blind Image QUality Evaluation. ACM Transactions on Multimedia Computing, Communications and Applications, 2022, 18, 1-19.	3.0	0

#	Article	IF	Citations
1425	Reciprocal question representation learning network for visual dialog. Applied Intelligence, 0, , .	3.3	1
1426	A cooperative approach based on self-attention with interactive attribute for image caption. Multimedia Tools and Applications, 2023, 82, 1223-1236.	2.6	3
1427	Note: Towards Devising an Efficient VQA in the Bengali Language. , 2022, , .		1
1428	Piecewise linear neural networks and deep learning. Nature Reviews Methods Primers, 2022, 2, .	11.8	11
1429	Computer Vision-Based Hazard Identification of Construction Site Using Visual Relationship Detection and Ontology. Buildings, 2022, 12, 857.	1.4	8
1430	Conditional Embedding Pre-Training Language Model for Image Captioning. Neural Processing Letters, 2022, 54, 4987-5003.	2.0	2
1431	Talk-to-Resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot. Robotics and Autonomous Systems, 2022, 155, 104183.	3.0	6
1432	Learning multimodal relationship interaction for visual relationship detection. Pattern Recognition, 2022, 132, 108848.	5.1	3
1433	SPCA-Net: a based on spatial position relationship co-attention network for visual question answering. Visual Computer, 2022, 38, 3097-3108.	2.5	8
1434	Gender and Racial Bias in Visual Question Answering Datasets. , 2022, , .		4
1435	Scene graph generation with award-punishment strategy. Knowledge-Based Systems, 2022, , 109239.	4.0	0
1436	OCR-oriented Master Object for Text Image Captioning. , 2022, , .		2
1438	Depth-Aware and Semantic Guided Relational Attention Network for Visual Question Answering. IEEE Transactions on Multimedia, 2023, 25, 5344-5357.	5.2	2
1440	Adaptive Semantic-Enhanced Transformer for Image Captioning. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35, 1785-1796.	7.2	4
1441	Cross-Domain Multi-Style Merge for Image Captioning. SSRN Electronic Journal, 0, , .	0.4	0
1443	EKTVQA: Generalized Use of External Knowledge to Empower Scene Text in Text-VQA. IEEE Access, 2022, 10, 72092-72106.	2.6	1
1444	Towards Optimal Correlational Object Search. , 2022, , .		4
1445	Graph-based Cluttered Scene Generation and Interactive Exploration using Deep Reinforcement Learning. , 2022, , .		7

#	Article	IF	Citations
1446	Grounding Predicates through Actions. , 2022, , .		3
1447	StructFormer: Learning Spatial Structure for Language-Guided Semantic Rearrangement of Novel Objects. , 2022, , .		15
1448	OVIS: Open-Vocabulary Visual Instance Search via Visual-Semantic Aligned Representation Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36, 1773-1781.	3.6	1
1449	Playing Lottery Tickets with Vision and Language. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36, 652-660.	3.6	5
1450	Improving Scene Graph Classification by Exploiting Knowledge from Texts. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36, 2189-2197.	3.6	5
1451	Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36, 12262-12267.	3.6	5
1452	é™„åŠåè§é¢„æµ‹å™¨è¾¾åŠ©çš„å‡è¡¡åŒ–åœºæ™¯å›¾çš„Ÿæˆ• Scientia Sinica Informationis, 2022, , .	0.2	0
1453	RSSGG_CS: Remote Sensing Image Scene Graph Generation by Fusing Contextual Information and Statistical Knowledge. Remote Sensing, 2022, 14, 3118.	1.8	3
1454	GIAD-ST: Detecting anomalies in human monitoring based on generative inpainting via self-supervised multi-task learning. Journal of Intelligent Information Systems, 0, , .	2.8	1
1455	Improving text-image cross-modal retrieval with contrastive loss. Multimedia Systems, 0, , .	3.0	0
1456	Cross modification attention-based deliberation model for image captioning. Applied Intelligence, 0, , .	3.3	1
1457	Weakly Supervised Text-based Actor-Action Video Segmentation by Clip-level Multi-instance Learning. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-22.	3.0	3
1458	Adaptive Text Denoising Network for Image Caption Editing. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-18.	3.0	2
1459	Referring Expression Comprehension Via Enhanced Cross-modal Graph Attention Networks. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-21.	3.0	1
1460	CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing, 2022, 508, 293-304.	3.5	159
1461	Animating Images to Transfer CLIP for Video-Text Retrieval. , 2022, , .		5
1462	Incorporating External Knowledge Reasoning for Vision-and-Language Navigation with Assistantâ€™s Help. Applied Sciences (Switzerland), 2022, 12, 7053.	1.3	1
1463	Cross-Probe BERT for Fast Cross-Modal Search. , 2022, , .		5

#	Article	IF	Citations
1464	HVLM: Exploring Human-Like Visual Cognition and Language-Memory Network for Visual Dialog. Information Processing and Management, 2022, 59, 103008.	5.4	5
1465	Medical visual question answering based on question-type reasoning and semantic space constraint. Artificial Intelligence in Medicine, 2022, 131, 102346.	3.8	8
1466	FERNIE-ViL: Facial Expression Enhanced Vision-and-Language Model. , 2021, , .		0
1467	GA-SRN: graph attention based text-image semantic reasoning network for fine-grained image classification and retrieval. Neural Computing and Applications, 2022, 34, 21387-21401.	3.2	2
1468	Knowledge-Based Scene Graph Generation with Visual Contextual Dependency. Mathematics, 2022, 10, 2525.	1.1	3
1469	CAPTION: Caption Analysis with Proposed Terms, Image of Objects, and Natural Language Processing. SN Computer Science, 2022, 3, .	2.3	1
1470	Visual Relationship Detection With Deep Structural Ranking. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32, .	3.6	56
1471	Scene-Centric Joint Parsing of Cross-View Videos. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32, .	3.6	6
1472	Deep Structured Learning for Visual Relationship Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32, .	3.6	23
1473	HCVRD: A Benchmark for Large-Scale Human-Centered Visual Relationship Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32, .	3.6	16
1474	A General Formulation for Safely Exploiting Weakly Supervised Data. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32, .	3.6	5
1475	Improving Image Representations viaÂMoCo Pre-training forÂMultimodal CXR Classification. Lecture Notes in Computer Science, 2022, , 623-635.	1.0	0
1476	Debiased Scene Graph Generation for Dual Imbalance Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, , 1-15.	9.7	0
1477	Cross-Modal Graph With Meta Concepts for Video Captioning. IEEE Transactions on Image Processing, 2022, 31, 5150-5162.	6.0	3
1478	Hierarchical Context-Based Emotion Recognition With Scene Graphs. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35, 3725-3739.	7.2	5
1479	ViNTER: Image Narrative Generation with Emotion-Arc-Aware Transformer. , 2022, , .		3
1480	Flexible and scalable annotation tool to develop scene understanding datasets. , 2022, , .		0
1481	Embedding Arithmetic of Multimodal Queries for Image Retrieval. , 2022, , .		5

#	Article	IF	Citations
1482	Self-Supervised Road Layout Parsing with Graph Auto-Encoding. , 2022, , .		0
1483	The Topology and Language of Relationships in the Visual Genome Dataset. , 2022, , .		3
1484	Reasoning with Multi-Structure Commonsense Knowledge in Visual Dialog. , 2022, , .		4
1485	Semantically Grounded Visual Embeddings for Zero-Shot Learning. , 2022, , .		2
1486	NewsImages. , 2022, , .		1
1487	Modulating Bottom-Up and Top-Down Visual Processing via Language-Conditional Filters. , 2022, , .		1
1488	Multi-Scale Graph Attention Network for Scene Graph Generation. , 2022, , .		3
1489	High-Quality Image Generation from Scene Graphs with Transformer. , 2022, , .		1
1490	Automatic Pantograph Health Status Report Generation Based on Dense Captioning. , 2022, , .		0
1491	Hadamard Product Perceptron Attention for Image Captioning. Neural Processing Letters, 2023, 55, 2707-2724.	2.0	3
1492	Aligning Image Semantics and Label Concepts for Image Multi-Label Classification. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-23.	3.0	3
1493	Spatial relation learning in complementary scenarios with deep neural networks. Frontiers in Neurorobotics, 0, 16, .	1.6	0
1494	Integer Network for Cross Platform Graph Data Lossless Compression. , 2022, , .		0
1495	An Object Localization-based Dense Image Captioning Framework in Hindi. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22, 1-15.	1.3	1
1496	CommerceMM. , 2022, , .		13
1497	Long-term object search using incremental scene graph updating. Robotica, 2023, 41, 962-975.	1.3	2
1498	U-BERT for Fast and Scalable Text-Image Retrieval. , 2022, , .		3
1499	A Survey of Natural Language Generation. ACM Computing Surveys, 2023, 55, 1-38.	16.1	29

#	Article	IF	Citations
1500	Visuals to Text: A Comprehensive Review on Automatic Image Captioning. IEEE/CAA Journal of Automatica Sinica, 2022, 9, 1339-1365.	8.5	9
1501	Fine-grained label learning in object detection with weak supervision of captions. Multimedia Tools and Applications, 2023, 82, 6557-6579.	2.6	2
1502	Semantic interdisciplinary evaluation of image captioning models. Cogent Engineering, 2022, 9, .	1.1	7
1503	Global Local Fusion Neural Network for Multimodal Sentiment Analysis. Applied Sciences (Switzerland), 2022, 12, 8453.	1.3	2
1504	Transformer networks with adaptive inference for scene graph generation. Applied Intelligence, 0, , .	3.3	0
1505	Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence, 2022, 4, 669-677.	8.3	89
1506	Anomaly detection in surveillance videos: a thematic taxonomy of deep models, review and performance analysis. Artificial Intelligence Review, 2023, 56, 3319-3368.	9.7	6
1507	Incorporating retrieval-based method for feature enhanced image captioning. Applied Intelligence, 2023, 53, 9731-9743.	3.3	3
1508	Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge. Automation in Construction, 2022, 142, 104535.	4.8	12
1509	CAAN: Context-Aware attention network for visual question answering. Pattern Recognition, 2022, 132, 108980.	5.1	26
1510	Reasonable object detection guided by knowledge of global context and category relationship. Expert Systems With Applications, 2022, 209, 118285.	4.4	2
1512	Multilevel attention and relation network based image captioning model. Multimedia Tools and Applications, 2023, 82, 10981-11003.	2.6	9
1513	Multi-Granularity Semantic Collaborative Reasoning Network for Visual Dialog. Applied Sciences (Switzerland), 2022, 12, 8947.	1.3	2
1514	PU-GEN: Enhancing generative commonsense reasoning for language models with human-centered knowledge. Knowledge-Based Systems, 2022, 256, 109861.	4.0	4
1515	Rethinking referring relationships from a perspective of mask-level relational reasoning. Pattern Recognition, 2023, 133, 109044.	5.1	1
1516	Answering knowledge-based visual questions via the exploration of Question Purpose. Pattern Recognition, 2023, 133, 109015.	5.1	7
1517	Image captioning for effective use of language models in knowledge-based visual question answering. Expert Systems With Applications, 2023, 212, 118669.	4.4	14
1518	DFMKE: A dual fusion multi-modal knowledge graph embedding framework for entity alignment. Information Fusion, 2023, 90, 111-119.	11.7	10

#	Article	IF	Citations
1519	Visual Relation Detection using Hybrid Analogical Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35, 801-808.	3.6	1
1520	Similarity Reasoning and Filtration for Image-Text Matching. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35, 1218-1226.	3.6	120
1521	Image-to-Image Retrieval by Learning Similarity between Scene Graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35, 10718-10726.	3.6	9
1522	Mutil-level Local Alignment andÂSemantic Matching Network forÂImage-Text Retrieval. Lecture Notes in Computer Science, 2022, , 212-224.	1.0	0
1523	Video Question Answering With Prior Knowledge and Object-Sensitive Learning. IEEE Transactions on Image Processing, 2022, 31, 5936-5948.	6.0	17
1524	A Comprehensive Framework for Long-Tailed Learning via Pretraining and Normalization. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35, 3437-3449.	7.2	1
1525	4D-OR: Semantic Scene Graphs forÂORÂDomain Modeling. Lecture Notes in Computer Science, 2022, , 475-485.	1.0	8
1526	Heterogeneous Knowledge Network for Visual Dialog. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 861-871.	5.6	3
1527	MLMG-SGG: Multilabel Scene Graph Generation With Multigrained Features. IEEE Transactions on Image Processing, 2024, 33, 1549-1559.	6.0	3
1528	SGT: Scene Graph-Guided Transformer forÂSurgical Report Generation. Lecture Notes in Computer Science, 2022, , 507-518.	1.0	4
1529	ADS-Cap: A Framework forÂAccurate andÂDiverse Stylized Captioning withÂUnpaired Stylistic Corpora. Lecture Notes in Computer Science, 2022, , 736-748.	1.0	0
1530	Decoupled Cross-Modal Phrase-Attention Network for Image-Sentence Matching. IEEE Transactions on Image Processing, 2024, 33, 1326-1337.	6.0	1
1531	What Happens in Crowd Scenes: A New Dataset About Crowd Scenes for Image Captioning. IEEE Transactions on Multimedia, 2023, 25, 5400-5412.	5.2	2
1532	Toward Driving Scene Understanding: A Paradigm and Benchmark Dataset for Ego-Centric Traffic Scene Graph Representation. IEEE Journal of Radio Frequency Identification, 2022, 6, 962-967.	1.5	2
1533	Medical visual question answering via corresponding feature fusion combined with semantic attention. Mathematical Biosciences and Engineering, 2022, 19, 10192-10212.	1.0	4
1534	CRIC: A VQA Dataset for Compositional Reasoning on Vision and Commonsense. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, , 1-18.	9.7	0
1535	BERTHop: An Effective Vision-and-Language Model forÂChest X-ray Disease Diagnosis. Lecture Notes in Computer Science, 2022, , 725-734.	1.0	5
1536	Graph Representation Learning Meets Computer Vision: A Survey. IEEE Transactions on Artificial Intelligence, 2023, 4, 2-22.	3.4	14

#	Article	IF	Citations
1537	Weakly-Supervised Video Object Grounding via Learning Uni-Modal Associations. IEEE Transactions on Multimedia, 2023, 25, 6329-6340.	5.2	1
1538	DT2I: Dense Text-to-Image Generation fromÂRegion Descriptions. Lecture Notes in Computer Science, 2022, , 395-406.	1.0	2
1539	KD-VLP: Improving End-to-End Vision-and-Language Pretraining with Object Knowledge Distillation. , 2022, , .		3
1540	An Open-Ended Web Knowledge Retrieval Framework for the Household Domain With Explanation and Learning Through Argumentation. International Journal on Semantic Web and Information Systems, 2022, 18, 1-34.	2.2	1
1541	â€˜Cool glasses, where did you get them?â€•Generating Visually Grounded Conversation Starters for Human-Robot Dialogue. , 2022, , .		2
1542	3DVQA: Visual Question Answering for 3D Environments. , 2022, , .		2
1543	Injecting Semantic Concepts into End-to-End Image Captioning. , 2022, , .		36
1544	VL-ADAPTER: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks. , 2022, , .		47
1545	RegionCLIP: Region-based Language-Image Pretraining. , 2022, , .		88
1546	Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering. , 2022, , .		14
1547	COTS: Collaborative Two-Stream Vision-Language Pre-Training Model for Cross-Modal Retrieval. , 2022, , .		21
1548	Bongard-HOI: Benchmarking Few-Shot Visual Reasoning for Human-Object Interactions. , 2022, , .		7
1549	Pseudo-Q: Generating Pseudo Language Queries for Visual Grounding. , 2022, , .		15
1550	Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks. , 2022, , .		17
1551	Vision-Language Pre-Training with Triple Contrastive Learning. , 2022, , .		74
1552	PPDL: Predicate Probability Distribution based Loss for Unbiased Scene Graph Generation. , 2022, , .		19
1553	SGTR: End-to-end Scene Graph Generation with Transformer. , 2022, , .		31
1554	Explaining Deep Convolutional Neural Networks via Latent Visual-Semantic Filter Attention. , 2022, , .		5

#	Article	IF	Citations
1555	Not All Relations are Equal: Mining Informative Labels for Scene Graph Generation. , 2022, , .		10
1556	Object-aware Video-language Pre-training for Retrieval. , 2022, , .		26
1557	Multi-modal Alignment using Representation Codebook. , 2022, , .		14
1558	The Devil is in the Labels: Noisy Label Correction for Robust Scene Graph Generation. , 2022, , .		31
1559	REX: Reasoning-aware and Grounded Explanation. , 2022, , .		8
1560	Query and Attention Augmentation for Knowledge-Based Explainable Reasoning. , 2022, , .		8
1561	Scaling Up Vision-Language Pretraining for Image Captioning. , 2022, , .		40
1562	An Empirical Study of Training End-to-End Vision-and-Language Transformers. , 2022, , .		88
1563	Hierarchical Modular Network for Video Captioning. , 2022, , .		30
1564	Grounded Language-Image Pre-training. , 2022, , .		130
1565	WebQA: Multihop and Multimodal QA. , 2022, , .		8
1566	M5Product: Self-harmonized Contrastive Learning for E-commercial Multi-modal Pretraining. , 2022, , .		11
1567	Grounding Answers for Visual Questions Asked by Visually Impaired People. , 2022, , .		12
1568	Interactive Image Synthesis with Panoptic Layout Generation. , 2022, , .		7
1569	Sim VQA: Exploring Simulated Environments for Visual Question Answering. , 2022, , .		5
1570	GlideNet: Global, Local and Intrinsic based Dense Embedding NETwork for Multi-category Attributes Prediction. , 2022, , .		5
1571	X -Trans2Cap: Cross-Modal Knowledge Transfer using Transformer for 3D Dense Captioning. , 2022, , .		19
1572	Structured Sparse R-CNN for Direct Scene Graph Generation. , 2022, , .		18

#	Article	IF	Citations
1573	Unsupervised Vision-Language Parsing: Seamlessly Bridging Visual Scene Graphs with Language Structures via Dependency Relationships. , 2022, , .		5
1574	Towards General Purpose Vision Systems: An End-to-End Task-Agnostic Vision-Language Architecture. , 2022, , .		9
1575	GroupViT: Semantic Segmentation Emerges from Text Supervision. , 2022, , .		101
1576	CoSSL: Co-Learning of Representation and Classifier for Imbalanced Semi-Supervised Learning. , 2022, , .		16
1577	Reinforced Structured State-Evolution for Vision-Language Navigation. , 2022, , .		14
1578	ViSTA: Vision and Scene Text Aggregation for Cross-Modal Retrieval. , 2022, , .		30
1579	Unsupervised Vision-and-Language Pretraining via Retrieval-based Multi-Granular Alignment. , 2022, , .		5
1580	HL-Net: Heterophily Learning Network for Scene Graph Generation. , 2022, , .		15
1581	Negative-Aware Attention Framework for Image-Text Matching. , 2022, , .		48
1582	Align and Prompt: Video-and-Language Pre-training with Entity Prompts. , 2022, , .		50
1583	NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks. , 2022, , .		16
1584	Spatial Commonsense Graph for Object Localisation in Partial Scenes. , 2022, , .		8
1585	LaTr: Layout-Aware Transformer for Scene-Text VQA. , 2022, , .		33
1586	Dual-Key Multimodal Backdoors for Visual Question Answering. , 2022, , .		9
1587	Measuring Compositional Consistency for Video Question Answering. , 2022, , .		3
1588	Scene Graph Expansion for Semantics-Guided Image Outpainting. , 2022, , .		9
1589	It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection. , 2022, , .		5
1590	Simple Multi-dataset Detection. , 2022, , .		21

#	Article	IF	Citations
1591	Modeling Image Composition for Complex Scene Generation. , 2022, , .		10
1592	From Representation to Reasoning: Towards both Evidence and Commonsense Reasoning for Video Question-Answering. , 2022, , .		14
1593	Trustworthy Long-Tailed Classification. , 2022, , .		19
1594	RelTransformer: A Transformer-Based Long-Tail Visual Relationship Recognition. , 2022, , .		2
1595	MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering. , 2022, , .		26
1596	RU-Net: Regularized Unrolling Network for Scene Graph Generation. , 2022, , .		11
1597	Uncertainty-based Visual Question Answering: Estimating Semantic Inconsistency between Image and Knowledge Base. , 2022, , .		1
1598	Region-based Cross-modal Retrieval. , 2022, , .		0
1599	TRICAN: Multi-Modal Hateful Memes Detection with Triplet-Relation Information Cross-Attention Network. , 2022, , .		0
1600	EdgeNet for efficient scene graph classification. , 2022, , .		0
1601	Multi-Attention Cascade Model Based on Multi-Head Structure for Image-Text Retrieval. , 2022, , .		1
1602	Image Caption Method Based on Graph Attention Network with Global Context. , 2022, , .		0
1603	TransNav: spatial sequential transformer network for visual navigation. Journal of Computational Design and Engineering, 2022, 9, 1866-1878.	1.5	1
1604	Hierarchical decoding with latent context for image captioning. Neural Computing and Applications, 2023, 35, 2429-2442.	3.2	1
1605	Automatic Defect Description of Railway Track Line Image Based on Dense Captioning. Sensors, 2022, 22, 6419.	2.1	5
1606	Semantic-aware visual scene representation. International Journal of Multimedia Information Retrieval, 2022, 11, 619-638.	3.6	2
1607	A Review of Multi-Modal Learning from the Text-Guided Visual Processing Viewpoint. Sensors, 2022, 22, 6816.	2.1	1
1608	TrollsWithOpinion: A taxonomy and dataset for predicting domain-specific opinion manipulation in troll memes. Multimedia Tools and Applications, 0, , .	2.6	1

#	Article	IF	Citations
1609	Guided Graph Attention Learning for Video-Text Matching. ACM Transactions on Multimedia Computing, Communications and Applications, 0, , .	3.0	0
1610	Scene Graph Semantic Inference for Image and Text Matching. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22, 1-23.	1.3	7
1611	A survey of transformer-based multimodal pre-trained modals. Neurocomputing, 2023, 515, 89-106.	3.5	14
1612	Describing UI Screenshots in Natural Language. ACM Transactions on Intelligent Systems and Technology, 2023, 14, 1-28.	2.9	1
1613	Explainable AI: A Neurally-Inspired Decision Stack Framework. Biomimetics, 2022, 7, 127.	1.5	3
1614	Socially situated artificial intelligence enables learning from human interaction. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119, .	3.3	19
1615	Research on visual question answering based on dynamic memory network model of multiple attention mechanisms. Scientific Reports, 2022, 12, .	1.6	4
1616	Learning to Collocate Visual-Linguistic Neural Modules for Image Captioning. International Journal of Computer Vision, 2023, 131, 82-100.	10.9	4
1617	On the Effectiveness of Images in Multi-modal Text Classification: An Annotation Study. ACM Transactions on Asian and Low-Resource Language Information Processing, 0, , .	1.3	0
1618	Prototype localâ€“global alignment network for imageâ€“text retrieval. International Journal of Multimedia Information Retrieval, 2022, 11, 525-538.	3.6	1
1619	2.5D visual relationship detection. Computer Vision and Image Understanding, 2022, 224, 103557.	3.0	1
1620	Safety compliance checking of construction behaviors using visual question answering. Automation in Construction, 2022, 144, 104580.	4.8	4
1621	ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35, 3208-3216.	3.6	91
1622	Classification by Attention: Scene Graph Classification with Prior Knowledge. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35, 5025-5033.	3.6	21
1626	Dynamic Graph Representation Learning for Video Dialog via Multi-Modal Shuffled Transformers. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35, 1415-1423.	3.6	15
1627	FIXMYPOSE: Pose Correctional Captioning and Retrieval. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35, 13161-13170.	3.6	4
1628	PlotShot: Generating Discourse-Constrained Stories Around Photos. Proceedings, 2016, 12, 2-8.	0.7	1
1629	Multi-Label Retinal Disease Classification Using Transformers. IEEE Journal of Biomedical and Health Informatics, 2023, 27, 2739-2750.	3.9	7

#	Article	IF	Citations
1630	Hierarchical Memory Learning forÂFine-Grained Scene Graph Generation. Lecture Notes in Computer Science, 2022, , 266-283.	1.0	6
1631	Unpaired Image Captioning by Image-Level Weakly-Supervised Visual Concept Recognition. IEEE Transactions on Multimedia, 2023, 25, 6702-6716.	5.2	2
1632	Geometric Features Informed Multi-person Human-Object Interaction Recognition inÂVideos. Lecture Notes in Computer Science, 2022, , 474-491.	1.0	3
1633	Webly Supervised Concept Expansion forÂGeneral Purpose Vision Models. Lecture Notes in Computer Science, 2022, , 662-681.	1.0	8
1634	Learning Audio-Video Modalities fromÂImage Captions. Lecture Notes in Computer Science, 2022, , 407-426.	1.0	10
1635	Object-Centric Unsupervised Image Captioning. Lecture Notes in Computer Science, 2022, , 219-235.	1.0	1
1636	ViGAT: Bottom-Up Event Recognition and Explanation in Video Using Factorized Graph Attention Network. IEEE Access, 2022, 10, 108797-108816.	2.6	7
1637	Panoptic Scene Graph Generation. Lecture Notes in Computer Science, 2022, , 178-196.	1.0	13
1638	Fine-Grained Scene Graph Generation withÂData Transfer. Lecture Notes in Computer Science, 2022, , 409-424.	1.0	21
1639	Towards Open-Vocabulary Scene Graph Generation withÂPrompt-Based Finetuning. Lecture Notes in Computer Science, 2022, , 56-73.	1.0	6
1640	Improving Closed andÂOpen-Vocabulary Attribute Prediction Using Transformers. Lecture Notes in Computer Science, 2022, , 201-219.	1.0	3
1641	X-DETR: A Versatile Architecture forÂInstance-wise Vision-Language Tasks. Lecture Notes in Computer Science, 2022, , 290-308.	1.0	6
1642	GRIT: Faster andÂBetter Image Captioning Transformer Using Dual Visual Features. Lecture Notes in Computer Science, 2022, , 167-184.	1.0	28
1643	Switch-BERT: Learning toÂModel Multimodal Interactions byÂSwitching Attention andÂInput. Lecture Notes in Computer Science, 2022, , 330-346.	1.0	0
1644	Neural Scene Decoration fromÂaÂSingle Photograph. Lecture Notes in Computer Science, 2022, , 136-152.	1.0	1
1645	Single-Stream Multi-level Alignment forÂVision-Language Pretraining. Lecture Notes in Computer Science, 2022, , 735-751.	1.0	2
1646	Video Graph Transformer forÂVideo Question Answering. Lecture Notes in Computer Science, 2022, , 39-58.	1.0	18
1647	PACS: A Dataset forÂPhysical Audiovisual CommonSense Reasoning. Lecture Notes in Computer Science, 2022, , 292-309.	1.0	1

#	Article	IF	Citations
1648	Open Vocabulary Object Detection with Pseudo Bounding-Box Labels. Lecture Notes in Computer Science, 2022, , 266-282.	1.0	9
1649	A Simple Approach andÂBenchmark forÂ21,000-Category Object Detection. Lecture Notes in Computer Science, 2022, , 1-18.	1.0	0
1650	EclipSE: Efficient Long-Range Video Retrieval Using Sight andÂSound. Lecture Notes in Computer Science, 2022, , 413-430.	1.0	8
1651	A Broad Study of Pre-training for Domain Generalization and Adaptation. Lecture Notes in Computer Science, 2022, , 621-638.	1.0	10
1652	Contrastive Vision-Language Pre-training withÂLimited Resources. Lecture Notes in Computer Science, 2022, , 236-253.	1.0	6
1653	Bottom UpÂTop DownÂDetection Transformers forÂLanguage Grounding inÂImages andÂPoint Clouds. Lecture Notes in Computer Science, 2022, , 417-433.	1.0	7
1654	Scaling Open-Vocabulary Image Segmentation withÂImage-Level Labels. Lecture Notes in Computer Science, 2022, , 540-557.	1.0	31
1655	The Abduction ofÂSherlock Holmes: A Dataset forÂVisual Abductive Reasoning. Lecture Notes in Computer Science, 2022, , 558-575.	1.0	2
1656	Mining Cross-Person Cues forÂBody-Part Interactiveness Learning inÂHOI Detection. Lecture Notes in Computer Science, 2022, , 121-136.	1.0	12
1657	Object Detection Based on Embedding Internal and External Knowledge. Lecture Notes in Computer Science, 2022, , 351-365.	1.0	0
1658	Rethinking Data Augmentation forÂRobust Visual Question Answering. Lecture Notes in Computer Science, 2022, , 95-112.	1.0	9
1659	UniTAB: Unifying Text andÂBox Outputs forÂGrounded Vision-Language Modeling. Lecture Notes in Computer Science, 2022, , 521-539.	1.0	16
1660	Class-Agnostic Object Detection withÂMulti-modal Transformer. Lecture Notes in Computer Science, 2022, , 512-531.	1.0	7
1661	Simple Open-Vocabulary Object Detection. Lecture Notes in Computer Science, 2022, , 728-755.	1.0	21
1662	Meta Spatio-Temporal Debiasing forÂVideo Scene Graph Generation. Lecture Notes in Computer Science, 2022, , 374-390.	1.0	8
1663	LocVTP: Video-Text Pre-training forÂTemporal Localization. Lecture Notes in Computer Science, 2022, , 38-56.	1.0	12
1664	Relationship Spatialization forÂDepth Estimation. Lecture Notes in Computer Science, 2022, , 615-637.	1.0	1
1665	Improving generalizability of ML-enabled software through domain specification. , 2022, , .		2

#	Article	IF	Citations
1666	CADE: The Missing Benchmark in Evaluating Dataset Requirements of AI-enabled Software. , 2022, , .		1
1667	LRB-Net: Improving VQA via division of labor strategy and multimodal classifiers. Displays, 2022, 75, 102329.	2.0	5
1668	AVR: attention based salient visual relationship detection. , 2022, , .		3
1669	Ques-to-Visual Guided Visual Question Answering. , 2022, , .		1
1670	IDEA: Increasing Text Diversity via Online Multi-Label Recognition for Vision-Language Pre-training. , 2022, , .		2
1671	Compute to Tell the Tale: Goal-Driven Narrative Generation. , 2022, , .		5
1672	Heterogeneous Learning for Scene Graph Generation. , 2022, , .		1
1673	ArCo: Attention-reinforced transformer with contrastive learning for image captioning. Image and Vision Computing, 2022, 128, 104570.	2.7	5
1674	Semantic embedding: scene image classification using scene-specific objects. Multimedia Systems, 0, , .	3.0	0
1675	A survey of Semantic Reasoning frameworks for robotic systems. Robotics and Autonomous Systems, 2023, 159, 104294.	3.0	2
1676	The Core of Smart Cities: Knowledge Representation and Descriptive Framework Construction in Knowledge-Based Visual Question Answering. Sustainability, 2022, 14, 13236.	1.6	1
1677	A Combination of Visual-Semantic Reasoning and Text Entailment-based Boosting Algorithm for Cheapfake Detection. , 2022, , .		4
1678	Search-oriented Micro-video Captioning. , 2022, , .		15
1679	Image Understanding by Captioning with Differentiable Architecture Search. , 2022, , .		2
1680	Towards Further Comprehension on Referring Expression with Rationale. , 2022, , .		0
1681	Leveraging Text Representation and Face-head Tracking for Long-form Multimodal Semantic Relation Understanding. , 2022, , .		2
1682	Understanding News Text and Images Connection with Context-enriched Multimodal Transformers. , 2022, , .		0
1683	Prompt-based Zero-shot Video Moment Retrieval. , 2022, , .		4

#	Article	IF	Citations
1684	Unsupervised and Pseudo-Supervised Vision-Language Alignment in Visual Dialog. , 2022, , .		4
1685	CMAL: A Novel Cross-Modal Associative Learning Framework for Vision-Language Pre-Training. , 2022, , .		0
1686	Multi-feature fusion enhanced transformer with multi-layer fused decoding for image captioning. Applied Intelligence, 2023, 53, 13398-13414.	3.3	2
1687	Residual Graph Attention Network and Expression-Respect Data Augmentation Aided Visual Grounding. , 2022, , .		1
1688	Relation Enhanced Vision Language Pre-Training. , 2022, , .		0
1689	Token Embeddings Alignment for Cross-Modal Retrieval. , 2022, , .		2
1690	RefCrowd: Grounding the Target in Crowd with Referring Expressions. , 2022, , .		0
1691	CAliC: Accurate and Efficient Image-Text Retrieval via Contrastive Alignment and Visual Contexts Modeling. , 2022, , .		3
1692	Progressive Tree-Structured Prototype Network for End-to-End Image Captioning. , 2022, , .		4
1693	Box-FaceS. , 2022, , .		0
1694	A Review on Methods and Applications in Multimodal Deep Learning. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-41.	3.0	18
1695	ChiQA. , 2022, , .		1
1696	Multi-modal co-attention relation networks for visual question answering. Visual Computer, 2023, 39, 5783-5795.	2.5	3
1697	A Multi-level Mesh Mutual Attention Model for Visual Question Answering. Data Science and Engineering, 2022, 7, 339-353.	4.6	3
1698	Integrating Object-aware and Interaction-aware Knowledge for Weakly Supervised Scene Graph Generation. , 2022, , .		8
1699	Multi Label Image Classification using Adaptive Graph Convolutional Networks (ML-AGCN). , 2022, , .		4
1700	Inferential Visual Question Generation. , 2022, , .		3
1701	AI-VQA. , 2022, , .		1

#	Article	IF	Citations
1702	AdsCVLR: Commercial Visual-Linguistic Representation Modeling in Sponsored Search. , 2022, , .		1
1703	Research Progress on Visionâ€“Language Multimodal Pretraining Model Technology. Electronics (Switzerland), 2022, 11, 3556.	1.8	1
1704	Improving weakly supervised phrase grounding via visual representation contextualization with contrastive learning. Applied Intelligence, 0, , .	3.3	0
1705	KBHN: A knowledge-aware bi-hypergraph network based on visual-knowledge features fusion for teaching image annotation. Information Processing and Management, 2023, 60, 103106.	5.4	4
1706	An NLP-guided ontology development and refinement approach to represent and query visual information. Expert Systems With Applications, 2023, 213, 118998.	4.4	5
1711	New Trends in Image Recognition Competitions. Kyokai Joho Imeji Zasshi/Journal of the Institute of Image Information and Television Engineers, 2019, 73, 1084-1089.	0.0	0
1712	Conformed Thoughts, Representational Systems, and Creative Procedures. Springer Series on Cultural Computing, 2022, , 167-188.	0.4	0
1713	Multimodal Emotion Classification with Multi-level Semantic Reasoning Network. IEEE Transactions on Multimedia, 2022, , 1-13.	5.2	2
1714	A-OKVQA: A Benchmark forÂVisual Question Answering Using World Knowledge. Lecture Notes in Computer Science, 2022, , 146-162.	1.0	15
1715	SeqTR: A Simple Yet Universal Network for Visual Grounding. Lecture Notes in Computer Science, 2022, , 598-615.	1.0	15
1716	OSANet: Object Semantic Attention Network for Visual Sentiment Analysis. IEEE Transactions on Multimedia, 2023, 25, 7139-7148.	5.2	4
1717	Human-Centric Image Cropping withÂPartition-Aware andÂContent-Preserving Features. Lecture Notes in Computer Science, 2022, , 181-197.	1.0	4
1718	Rethinking Feature Extraction: Gradient-Based Localized Feature Extraction for End-To-End Surgical Downstream Tasks. IEEE Robotics and Automation Letters, 2022, 7, 12623-12630.	3.3	1
1719	GRIT-VLP: Grouped Mini-batch Sampling forÂEfficient Vision andÂLanguage Pre-training. Lecture Notes in Computer Science, 2022, , 395-412.	1.0	1
1720	Latent Space Semantic Supervision Based on Knowledge Distillation for Cross-Modal Retrieval. IEEE Transactions on Image Processing, 2022, 31, 7154-7164.	6.0	4
1721	Scene-Text Oriented Referring Expression Comprehension. IEEE Transactions on Multimedia, 2023, 25, 7208-7221.	5.2	3
1722	Bayesian Tracking ofÂVideo Graphs Using Joint Kalman Smoothing andÂRegistration. Lecture Notes in Computer Science, 2022, , 440-456.	1.0	1
1723	Weakly Supervised Grounding for VQA in Vision-Language Transformers. Lecture Notes in Computer Science, 2022, , 652-670.	1.0	3

#	Article	IF	Citations
1724	Positional Attention Guided Transformer-like Architecture for Visual Question Answering. IEEE Transactions on Multimedia, 2022, , 1-13.	5.2	1
1725	Concept-Enhanced Relation Network for Video Visual Relation Inference. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 2233-2244.	5.6	1
1726	ECCV Caption: Correcting False Negatives byÂCollecting Machine-and-Human-verified Image-Caption Associations forÂMS-COCO. Lecture Notes in Computer Science, 2022, , 1-19.	1.0	5
1727	Unpaired referring expression grounding via bidirectional cross-modal matching. Neurocomputing, 2023, 518, 39-49.	3.5	1
1728	Unifying knowledge iterative dissemination and relational reconstruction network for imageâ€“text matching. Information Processing and Management, 2023, 60, 103154.	5.4	13
1729	Interactive Drawing Interface for Editing Scene Graph. , 2022, , .		1
1730	RFE-SRN: Image-text similarity reasoning network based on regional feature enhancement. Neurocomputing, 2023, 518, 593-601.	3.5	1
1731	An Enhanced Object Detection Model for Scene Graph Generation. Lecture Notes on Data Engineering and Communications Technologies, 2023, , 333-343.	0.5	1
1732	FNet withÂCross-Attention Encoder forÂVisual Question Answering. Lecture Notes on Data Engineering and Communications Technologies, 2023, , 602-611.	0.5	0
1733	A review of emerging research directions in Abstract Visual Reasoning. Information Fusion, 2023, 91, 713-736.	11.7	2
1734	Guide and interact: scene-graph based generation and control of video captions. Multimedia Systems, 2023, 29, 797-809.	3.0	1
1735	Quantitative Short-Term Precipitation Model Using Multimodal Data Fusion Based on a Cross-Attention Mechanism. Remote Sensing, 2022, 14, 5839.	1.8	3
1736	Personalized Saliency in Task-Oriented Semantic Communications: Image Transmission and Performance Analysis. IEEE Journal on Selected Areas in Communications, 2023, 41, 186-201.	9.7	22
1737	Element Information Enhancement forÂDiagram Question Answering withÂSynthetic Data. Communications in Computer and Information Science, 2022, , 78-86.	0.4	0
1738	DRAKE: Deep Pair-Wise Relation Alignment for Knowledge-Enhanced Multimodal Scene Graph Generation in Social Media Posts. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 3199-3213.	5.6	1
1739	Relative Position Relationship Learning Network for Scene Graph Generation. Lecture Notes in Computer Science, 2022, , 551-559.	1.0	0
1740	Semantic-Based Image Retrieval Using RS-Tree and Knowledge Graph. Lecture Notes in Computer Science, 2022, , 481-495.	1.0	0
1741	OG-SGG: Ontology-Guided Scene Graph Generationâ€”A Case Study in Transfer Learning for Telepresence Robotics. IEEE Access, 2022, 10, 132564-132583.	2.6	6

#	Article	IF	Citations
1742	Multi-Modal Knowledge Graph Construction and Application: A Survey. IEEE Transactions on Knowledge and Data Engineering, 2022, , 1-20.	4.0	37
1743	3D Question Answering. IEEE Transactions on Visualization and Computer Graphics, 2024, 30, 1772-1786.	2.9	4
1744	Catch Me if You Can: A Novel Task forÂDetection of Covert Geo-Locations (CGL). Lecture Notes in Electrical Engineering, 2022, , 199-217.	0.3	0
1745	One-Stage Visual Relationship Referring With Transformers and Adaptive Message Passing. IEEE Transactions on Image Processing, 2023, 32, 190-202.	6.0	0
1746	State-Aware Compositional Learning Toward Unbiased Training for Scene Graph Generation. IEEE Transactions on Image Processing, 2023, 32, 43-56.	6.0	3
1747	Reliable Extraction of Semantic Information and Rate of Innovation Estimation for Graph Signals. IEEE Journal on Selected Areas in Communications, 2023, 41, 119-140.	9.7	2
1748	Language Model Agnostic Gray-Box Adversarial Attack on Image Captioning. IEEE Transactions on Information Forensics and Security, 2023, 18, 626-638.	4.5	7
1749	ITeM: Image-to-Text Matching for Multimodal Documents. Journal of Natural Language Processing, 2022, 29, 1198-1232.	0.1	0
1750	Region Probability Map-Guided Fast Wide-Area Multiobject Detection. IEEE Transactions on Instrumentation and Measurement, 2023, 72, 1-13.	2.4	0
1751	Knowledge-Enriched Attention Network With Group-Wise Semantic for Visual Storytelling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 8634-8645.	9.7	3
1752	ArtCap: A Dataset for Image Captioning of Fine Art Paintings. IEEE Transactions on Computational Social Systems, 2024, 11, 576-587.	3.2	3
1753	Human-Object Interaction Detection: A Survey ofÂDeep Learning-Based Methods. Lecture Notes in Computer Science, 2022, , 441-452.	1.0	2
1754	Semantic Image Collection Summarization With Frequent Subgraph Mining. IEEE Access, 2022, 10, 131747-131764.	2.6	4
1755	Transforming Image Generation from Scene Graphs. , 2022, , .		1
1756	Transformer-based Scene Graph Generation Network With Relational Attention Module. , 2022, , .		0
1757	Improved Transformer with Parallel Encoders for Image Captioning. , 2022, , .		1
1758	Augmenting Vision Language Pretraining by Learning Codebook with Visual Semantics. , 2022, , .		0
1759	Improving Weakly Supervised Scene Graph Parsing through Object Grounding. , 2022, , .		0

#	Article	IF	Citations
1760	CaMEL: Mean Teacher Learning for Image Captioning. , 2022, , .		15
1761	GraphMapper: Efficient Visual Navigation by Scene Graph Generation. , 2022, , .		1
1762	Local Alignment with Global Semantic Consistence Network for Imageâ€“Text Matching. , 2022, , .		0
1763	High-Dimensional Indexing Scheme for Scene Graph Retrieval. , 2022, , .		0
1764	Semantic-aware multi-branch interaction network for deep multimodal learning. Neural Computing and Applications, 0, , .	3.2	0
1765	The Tensor Brain: A Unified Theory of Perception, Memory, and Semantic Decoding. Neural Computation, 2023, 35, 156-227.	1.3	1
1766	Self-Supervised Learning for Videos: A Survey. ACM Computing Surveys, 2023, 55, 1-37.	16.1	25
1767	Local self-attention in transformer for visual question answering. Applied Intelligence, 2023, 53, 16706-16723.	3.3	5
1768	CE-BART: Cause-and-Effect BART for Visual Commonsense Generation. Sensors, 2022, 22, 9399.	2.1	1
1769	Achieving Human Parity on Visual Question Answering. ACM Transactions on Information Systems, 2023, 41, 1-40.	3.8	2
1770	Generating comprehensive scene graphs with integrated multiple attribute detection. Machine Vision and Applications, 2023, 34, .	1.7	0
1771	Deep Learning Methods of Cross-Modal Tasks for Conceptual Design of Product Shapes: A Review. Journal of Mechanical Design, Transactions of the ASME, 2023, 145, .	1.7	6
1772	Text-dominated multimodal text classification with quadruplet attention. , 2022, , .		0
1773	A Framework for Image Captioning Based on Relation Network and Multilevel Attention Mechanism. Neural Processing Letters, 0, , .	2.0	0
1774	Improving visual-semantic embeddings by learning semantically-enhanced hard negatives for cross-modal information retrieval. Pattern Recognition, 2023, 137, 109272.	5.1	4
1775	Uncertainty-Aware Scene Graph Generation. Pattern Recognition Letters, 2022, , .	2.6	0
1776	Effective End-to-End Vision Language Pretraining With Semantic Visual Loss. IEEE Transactions on Multimedia, 2023, 25, 8408-8417.	5.2	0
1777	Neuron-Based Spiking Transmission and Reasoning Network for Robust Image-Text Retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 3516-3528.	5.6	4

#	Article	IF	Citations
1778	Intelligent Computing: The Latest Advances, Challenges, and Future. , 2023, 2, .		26
1779	Simultaneously Training and Compressing Vision-and-Language Pre-Training Model. IEEE Transactions on Multimedia, 2023, 25, 8194-8203.	5.2	0
1780	Universal Multimodal Representation for Language Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, , 1-18.	9.7	2
1781	Self-supervised Visual-Semantic Embedding Network Based onÂLocal Label Optimization. Lecture Notes in Computer Science, 2023, , 400-412.	1.0	0
1782	VLP: A Survey on Vision-language Pre-training. , 2023, 20, 38-56.		35
1783	Self-Training Vision Language BERTs With a Unified Conditional Model. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 3560-3569.	5.6	2
1784	Boosting Generic Visual-Linguistic Representation with Dynamic Contexts. IEEE Transactions on Multimedia, 2023, , 1-13.	5.2	0
1785	Text-based person search via local-relational-global fine grained alignment. Knowledge-Based Systems, 2023, 262, 110253.	4.0	3
1786	Cross-domain multi-style merge for image captioning. Computer Vision and Image Understanding, 2023, 228, 103617.	3.0	1
1787	Generative label fused network for imageâ€“text matching. Knowledge-Based Systems, 2023, 263, 110280.	4.0	7
1788	Cross-modal text and visual generation: A systematic review. Part 1: Image to text. Information Fusion, 2023, 93, 302-329.	11.7	3
1789	Image Generation from Scene Graph with Object Edges. , 2022, , .		3
1790	OVLN: Object-aware Vision and Language Navigation for Domestic Robots. , 2022, , .		0
1791	A Multimodal Fusion Scene Graph Generation Method Based on Semantic Description. , 2022, , .		0
1792	Egocentric Scene Description for the Blind and Visually Impaired. , 2022, , .		0
1793	Bottom-up and Top-down Object Inference Networks for Image Captioning. ACM Transactions on Multimedia Computing, Communications and Applications, 2022, 19, 1-18.	3.0	0
1794	Long-term robot manipulation task planning with scene graph and semantic knowledge. , 2023, 43, 12-22.		3
1795	Exploiting Long-Term Dependencies for Generating Dynamic Scene Graphs. , 2023, , .		3

#	Article	IF	Citations
1796	Composite Relationship Fields with Transformers for Scene Graph Generation. , 2023, , .		0
1797	Watching the News: Towards VideoQA Models that can Read. , 2023, , .		3
1798	PathNarratives: Data annotation for pathological human-AI collaborative diagnosis. Frontiers in Medicine, 0, 9, .	1.2	3
1799	Switching to Discriminative Image Captioning by Relieving a Bottleneck of Reinforcement Learning. , 2023, , .		3
1800	Dense but Efficient VideoQA for Intricate Compositional Reasoning. , 2023, , .		0
1801	From Less to More: Common-Sense Semantic Perception Benefits Image Captioning. Lecture Notes in Computer Science, 2023, , 356-368.	1.0	0
1802	Region Anomaly Detection via Spatial and Semantic Attributed Graph in Human Monitoring. Sensors, 2023, 23, 1307.	2.1	1
1803	Cross on Cross Attention: Deep Fusion Transformer for Image Captioning. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 4257-4268.	5.6	5
1804	Grounding human-object interaction to affordance behavior in multimodal datasets. Frontiers in Artificial Intelligence, 0, 6, .	2.0	2
1805	Automatic Construction Hazard Identification Integrating On-Site Scene Graphs with Information Extraction in Outfield Test. Buildings, 2023, 13, 377.	1.4	3
1806	K-VQG: Knowledge-aware Visual Question Generation for Common-sense Acquisition. , 2023, , .		1
1807	Learning by Hallucinating: Vision-Language Pre-training with Weak Supervision. , 2023, , .		0
1808	VirtualHome Action Genome: A Simulated Spatio-Temporal Scene Graph Dataset with Consistent Relationship Labels. , 2023, , .		1
1809	Scaling Novel Object Detection with Weakly Supervised Detection Transformers. , 2023, , .		2
1810	More Than Just Attention: Improving Cross-Modal Attentions with Contrastive Constraints for Image-Text Matching. , 2023, , .		2
1811	Exploring Optical-Flow-Guided Motion and Detection-Based Appearance for Temporal Sentence Grounding. IEEE Transactions on Multimedia, 2023, 25, 8539-8553.	5.2	9
1812	DRAMA: Joint Risk Localization and Captioning in Driving. , 2023, , .		1
1813	More Knowledge, Less Bias: Unbiasing Scene Graph Generation with Explicit Ontological Adjustment. , 2023, , .		4

#	Article	IF	Citations
1814	Textual Context-Aware Dense Captioning With Diverse Words. IEEE Transactions on Multimedia, 2023, 25, 8753-8766.	5.2	19
1815	MixGen: A New Multi-Modal Data Augmentation. , 2023, , .		12
1816	MKVSE: Multimodal Knowledge Enhanced Visual-semantic Embedding for Image-text Retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 2022, 19, 1-21.	3.0	3
1817	Quaternion Relation Embedding for Scene Graph Generation. IEEE Transactions on Multimedia, 2023, 25, 8646-8656.	5.2	3
1818	Knowledge-based Visual Context-Aware Framework for Applications in Robotic Services. , 2023, , .		0
1819	Semantic-Guided Selective Representation for Image Captioning. IEEE Access, 2023, 11, 14500-14510.	2.6	0
1820	Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. Progress in Artificial Intelligence, 2023, 12, 1-32.	1.5	6
1821	Zero-shot Object Detection Through Vision-Language Embedding Alignment. , 2022, , .		2
1822	Activity-oriented visual relationship detection technique for context recognition. Journal of Advanced Marine Engineering and Technology, 2022, 46, 422-429.	0.1	0
1823	Visual Question Answering System for Indian Regional Languages. , 2022, , .		0
1824	TECMH: Transformer-Based Cross-Modal Hashing For Fine-Grained Image-Text Retrieval. Computers, Materials and Continua, 2023, 75, 3713-3728.	1.5	0
1825	Visual Question Answering reasoning with external knowledge based on bimodal graph neural network. Electronic Research Archive, 2023, 31, 1948-1965.	0.4	0
1826	Independent Relationship Detection forÂReal-Time Scene Graph Generation. Communications in Computer and Information Science, 2023, , 106-118.	0.4	0
1827	Event-Oriented Visual Question Answering: The E-VQA Dataset and Benchmark. IEEE Transactions on Knowledge and Data Engineering, 2023, , 1-14.	4.0	0
1828	EAGLE: An Enhanced Attention-Based Strategy byÂGenerating Answers fromÂLearning Questions toÂaÂRemote Sensing Image. Lecture Notes in Computer Science, 2023, , 558-572.	1.0	0
1829	Causal-SETR: A SEgmentation TRansformer Variant Based onÂCausal Intervention. Lecture Notes in Computer Science, 2023, , 414-430.	1.0	1
1830	Nemesis: Neural Mean Teacher Learning-Based Emotion-Centric Speaker. Algorithms, 2023, 16, 97.	1.2	0
1831	FTN-VQA: Multimodal Reasoning by Leveraging a Fully Transformer-based Network for Visual Question Answering. Fractals, 0, , .	1.8	0

#	Article	IF	Citations
1832	Image generation models from scene graphs and layouts: A comparative analysis. Journal of King Saud University - Computer and Information Sciences, 2023, 35, 101543.	2.7	3
1833	What Is a Multi-Modal Knowledge Graph: A Survey. Big Data Research, 2023, 32, 100380.	2.6	3
1834	Towards local visual modeling for image captioning. Pattern Recognition, 2023, 138, 109420.	5.1	15
1835	Quaternion Representation Learning for cross-modal matching. Knowledge-Based Systems, 2023, 270, 110505.	4.0	2
1836	Balanced image captioning with task-aware decoupled learning and fusion. Neurocomputing, 2023, 538, 126159.	3.5	0
1837	Bi-Attention enhanced representation learning for image-text matching. Pattern Recognition, 2023, 140, 109548.	5.1	3
1838	A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. Artificial Intelligence Review, 2023, 56, 13619-13661.	9.7	2
1839	Nemo. Proceedings of the VLDB Endowment, 2022, 15, 4093-4105.	2.1	2
1840	Optimal Graph Transformer Viterbi knowledge inference network for more successful visual navigation. Advanced Engineering Informatics, 2023, 55, 101889.	4.0	5
1841	Enhanced Video BERT for Fast Video Advertisement Retrieval. , 2022, , .		0
1842	Tree-based Text-Vision BERT for Video Search in Baidu Video Advertising. , 2022, , .		1
1843	A CNN-LSTM-based model for fashion image aesthetic captioning. , 2023, , .		0
1844	Grounding Scene Graphs on Natural Images via Visio-Lingual Message Passing. , 2023, , .		0
1845	Text-Guided Object Detector for Multi-modal Video Question Answering. , 2023, , .		0
1846	Improving Predicate Representation in Scene Graph Generation by Self-Supervised Learning. , 2023, , .		0
1847	Context understanding in computer vision: A survey. Computer Vision and Image Understanding, 2023, 229, 103646.	3.0	7
1848	Scalable and Accurate Self-supervised Multimodal Representation Learning without Aligned Video and Text Data. , 2023, , .		2
1849	Neural Belief Propagation for Scene Graph Generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 10161-10172.	9.7	2

#	Article	IF	Citations
1850	High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images. Computers in Biology and Medicine, 2023, 155, 106646.	3.9	37
1851	Im2Graph: A Weakly Supervised Approach for Generating Holistic Scene Graphs from Regional Dependencies. Future Internet, 2023, 15, 70.	2.4	0
1852	TraVL: Transferring Pre-trained Visual-Linguistic Models forÂCross-Lingual Image Captioning. Lecture Notes in Computer Science, 2023, , 341-355.	1.0	0
1853	Content-based and Knowledge-enriched Representations for Classification Across Modalities: A Survey. ACM Computing Surveys, 2023, 55, 1-40.	16.1	0
1854	Contextualized Scene Knowledge Graphs for XAI Benchmarking. , 2022, , .		2
1855	MUST-VQA: MUltilingual Scene-Text VQA. Lecture Notes in Computer Science, 2023, , 345-358.	1.0	2
1856	Knowledge-Enhanced Visual Question Answering with Multi-modal Joint Guidance. , 2022, , .		0
1857	Transformer-Enhanced Visual-Semantic Representation for Text-Image Retrieval. , 2022, , .		0
1858	Visual lifelog retrieval: humans and machines interpretation on first-person images. Multimedia Tools and Applications, 2023, 82, 37757-37787.	2.6	1
1859	Video inÂ10 Bits: Few-Bit VideoQA forÂEfficiency andÂPrivacy. Lecture Notes in Computer Science, 2023, , 738-754.	1.0	0
1860	See Finer, See More: Implicit Modality Alignment forÂText-Based Person Retrieval. Lecture Notes in Computer Science, 2023, , 624-641.	1.0	11
1861	Multimodal Fake News Analysis Based on Imageâ€“Text Similarity. IEEE Transactions on Computational Social Systems, 2024, 11, 959-972.	3.2	3
1862	AGREE: Aligning Cross-Modal Entities for Image-Text Retrieval Upon Vision-Language Pre-trained Models. , 2023, , .		1
1863	HGAN: Hierarchical Graph Alignment Network for Image-Text Retrieval. IEEE Transactions on Multimedia, 2023, 25, 9189-9202.	5.2	5
1864	TEVL: Trilinear Encoder for Video-language Representation Learning. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-20.	3.0	0
1865	Visual Paraphrase Generation with Key Information Retained. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-19.	3.0	1
1866	A Survey onÂ3D Scene Graphs: Definition, Generation andÂApplication. Lecture Notes in Networks and Systems, 2023, , 136-147.	0.5	0
1867	RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61, 1-13.	2.7	8

#	Article	IF	Citations
1868	SST-VLM: Sparse Sampling-Twice Inspired Video-Language Model. Lecture Notes in Computer Science, 2023, , 537-553.	1.0	0
1869	Causal Property Based Anti-conflict Modeling withÂHybrid Data Augmentation forÂUnbiased Scene Graph Generation. Lecture Notes in Computer Science, 2023, , 571-587.	1.0	0
1870	Is anÂObject-Centric Video Representation Beneficial forÂTransfer?. Lecture Notes in Computer Science, 2023, , 379-397.	1.0	2
1871	Unsupervised Cross-Modal Hashing With Modality-Interaction. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 5296-5308.	5.6	10
1872	A Unified Perspective of Multi-level Cross-Modal Similarity for Cross-Modal Retrieval. , 2022, , .		0
1873	Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey. Expert Systems With Applications, 2023, 221, 119773.	4.4	1
1874	Egocentric Image Captioning for Privacy-Preserved Passive Dietary Intake Monitoring. IEEE Transactions on Cybernetics, 2024, 54, 679-692.	6.2	6
1875	Affective Image Captioning for Visual Artworks Using Emotion-Based Cross-Attention Mechanisms. IEEE Access, 2023, 11, 24527-24534.	2.6	3
1876	A Survey of Vision and Language Related Multi-Modal Task. , 2022, 1, 111-136.		1
1877	Skew Class-Balanced Re-Weighting for Unbiased Scene Graph Generation. Machine Learning and Knowledge Extraction, 2023, 5, 287-303.	3.2	2
1878	Contrastive Adversarial Training for Multi-Modal Machine Translation. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22, 1-18.	1.3	0
1879	Spatial-aware topic-driven-based image Chinese caption for disaster news. Neural Computing and Applications, 2023, 35, 9481-9500.	3.2	0
1880	Bilaterally Slimmable Transformer for Elastic and Efficient Visual Question Answering. IEEE Transactions on Multimedia, 2023, 25, 9543-9556.	5.2	0
1881	Transfer Learning for the Visual Arts: The Multi-modal Retrieval of Iconclass Codes. Journal on Computing and Cultural Heritage, 2023, 16, 1-16.	1.2	0
1882	ELâ€GAN: Edgeâ€Enhanced Generative Adversarial Network for Layoutâ€toâ€Image Generation. Computer Graphics Forum, 2022, 41, 407-418.	1.8	1
1883	Fineâ€Grained Scene Graph Generation with Overlap Region and Geometrical Center. Computer Graphics Forum, 2022, 41, 359-370.	1.8	0
1884	Multi-source Interaction Network for TextVQA. , 2022, , .		0
1885	Rethinking symbolic and visual context in Referring Expression Generation. Frontiers in Artificial Intelligence, 0, 6, .	2.0	0

#	Article	IF	Citations
1886	MyscÃ©al: a deeper analysis of an interactive lifelog search engine. Multimedia Tools and Applications, 0, , .	2.6	1
1887	RSG-GCN: Predicting Semantic Relationships in Urban Traffic Scene With Map Geometric Prior. IEEE Open Journal of Intelligent Transportation Systems, 2023, 4, 244-260.	2.6	0
1888	Supervised Deep Learning Techniques for Image Description: A Systematic Review. Entropy, 2023, 25, 553.	1.1	1
1889	Video question answering supported by a multi-task learning objective. Multimedia Tools and Applications, 2023, 82, 38799-38826.	2.6	1
1890	VAQA: Visual Arabic Question Answering. Arabian Journal for Science and Engineering, 0, , .	1.7	1
1891	Integrating Language Guidance Into Image-Text Matching for Correcting False Negatives. IEEE Transactions on Multimedia, 2024, 26, 103-116.	5.2	1
1892	WATAA: Web Alternative Text Authoring Assistant for Improving Web Content Accessibility. , 2023, , .		3
1893	Double Graph Attention Networks for Visual Semantic Navigation. Neural Processing Letters, 0, , .	2.0	0
1894	Progressive Instance-Aware Feature Learning for Compositional Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 10317-10330.	9.7	9
1895	BPCN:A simple and efficient model for visual question answering. , 2022, , .		0
1897	Towards Captioning anÂImage Collection fromÂaÂCombined Scene Graph Representation Approach. Lecture Notes in Computer Science, 2023, , 178-190.	1.0	2
1898	Multimodal embodied attribute learning by robots for object-centric action policies. Autonomous Robots, 0, , .	3.2	0
1899	Effectively Filtering Images forÂBetter Multi-modal Knowledge Graph. Communications in Computer and Information Science, 2023, , 10-22.	0.4	0
1900	Multimodal Pretraining from Monolingual to Multilingual. , 2023, 20, 220-232.		0
1901	VISCOUNTH: A Large-scale Multilingual Visual Question Answering Dataset for Cultural Heritage. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19, 1-20.	3.0	3
1902	A Survey of Full-Cycle Cross-Modal Retrieval: From a Representation Learning Perspective. Applied Sciences (Switzerland), 2023, 13, 4571.	1.3	2
1903	Visual Relationship Detection for Workplace Safety Applications. IEEE Transactions on Artificial Intelligence, 2024, 5, 956-961.	3.4	1
1904	A Dual Reinforcement Learning Framework for Weakly Supervised Phrase Grounding. IEEE Transactions on Multimedia, 2024, 26, 394-405.	5.2	0

#	Article	IF	Citations
1905	A Multiview Text Imagination Network Based on Latent Alignment for Image-Text Matching. IEEE Intelligent Systems, 2023, 38, 41-50.	4.0	1
1906	Multi-Object Navigation Using Potential Target Position Policy Function. IEEE Transactions on Image Processing, 2023, 32, 2608-2619.	6.0	1
1907	Graph-Based Contextual Attention Network for Single Image Deraining. Lecture Notes in Computer Science, 2023, , 287-297.	1.0	0
1908	Context Matters: Distilling Knowledge Graph for Enhanced Object Detection. IEEE Transactions on Multimedia, 2024, 26, 487-500.	5.2	0
1909	Bottom-Up Transformer Reasoning Network for Text-Image Retrieval. Communications in Computer and Information Science, 2023, , 176-187.	0.4	0
1910	A Cross-Modal Object-Aware Transformer for Vision-and-Language Navigation. , 2022, , .		0
1911	Plug-and-Play Regulators for Image-Text Matching. IEEE Transactions on Image Processing, 2023, 32, 2322-2334.	6.0	1
1912	RelTR: Relation Transformer for Scene Graph Generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45, 11169-11183.	9.7	14
1913	A Pipeline for Story Visualization from Natural Language. Applied Sciences (Switzerland), 2023, 13, 5107.	1.3	0
1914	Semantic-Aware Graph Matching Mechanism for Multi-Label Image Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33, 6788-6803.	5.6	2
1915	Envisioning Narrative Intelligence: A Creative Visual Storytelling Anthology. , 2023, , .		3
1916	Semantic Representations With Attention Networks for Boosting Image Captioning. IEEE Access, 2023, 11, 40230-40239.	2.6	5
1917	Object semantic analysis for image captioning. Multimedia Tools and Applications, 0, , .	2.6	0
1919	Cleaner Categories Improve Object Detection andÂVisual-Textual Grounding. Lecture Notes in Computer Science, 2023, , 412-442.	1.0	0
1920	Similarity Contrastive Capsule Transformation for Image-Text Matching. , 2023, , .		0
1922	Video Captioning Using Deep Learning Approach-A Comprehensive Survey. Proceedings in Adaptation, Learning and Optimization, 2023, , 68-87.	1.5	0
1928	Visual Questions Answering Developments, Applications, Datasets and Opportunities: A State-of-the-Art Survey. , 2023, , .		4
1933	Jointly Visual- and Semantic-Aware Graph Memory Networks for Temporal Sentence Localization in Videos. , 2023, , .		2

#	Article	IF	Citations
1934	Embrace Smaller Attention: Efficient Cross-Modal Matching with Dual Gated Attention Fusion. , 2023, , .		0
1935	ERBNet: An Effective Representation Based Network for Unbiased Scene Graph Generation. , 2023, , .		0
1936	Divcon: Learning Concept Sequences for Semantically Diverse Image Captioning. , 2023, , .		0
1953	Attending to Transforms: A Survey on Transformer-based Image Captioning. , 2023, , .		0
1957	Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey. , 2023, 20, 447-482.		20
1961	Hate Speech Detection using Multimodal Meme Analysis. , 2023, , .		0
1963	Reference-Limited Compositional Zero-Shot Learning. , 2023, , .		0
1965	SkeletonGAN: Fine-Grained Pose Synthesis of Human-Object Interactions. , 2023, , .		0
1967	Image Dense Captioning of Irregular Regions Based on Visual Saliency. , 2023, , .		0
1971	Knowledge Distillation Across Vision and Language. Studies in Computational Intelligence, 2023, , 65-94.	0.7	0
1978	Combining Multi-vision Embedding inÂContextual Attention forÂVietnamese Visual Question Answering. Lecture Notes in Computer Science, 2023, , 172-185.	1.0	0
1982	Neuro-Psychological Approaches for Artificial Intelligence. Advances in Environmental Engineering and Green Technologies Book Series, 2023, , 29-43.	0.3	0
1983	Outside Knowledge Visual Question Answering Version 2.0. , 2023, , .		1
1984	I-Tuning: Tuning Frozen Language Models with Image for Lightweight Image Captioning. , 2023, , .		1
1985	Video Captioning via Relation-Aware Graph Learning. , 2023, , .		1
1988	VizOPS: A data-driven ontology to represent public place surveillance data. , 2022, , .		1
1997	Cross-Modality Time-Variant Relation Learning for Generating Dynamic Scene Graphs. , 2023, , .		0
1998	3D VSG: Long-term Semantic Scene Change Prediction through 3D Variable Scene Graphs. , 2023, , .		3

#	Article	IF	Citations
1999	Zero-Shot Object Goal Visual Navigation. , 2023, , .		3
2000	Towards Open-World Interactive Disambiguation for Robotic Grasping. , 2023, , .		1
2002	SRI-Graph: A Novel Scene-Robot Interaction Graph for Robust Scene Understanding. , 2023, , .		1
2003	FewSOL: A Dataset for Few-Shot Object Learning in Robotic Environments. , 2023, , .		0
2008	Local and Global Multimodal Interaction for Image Caption. , 2023, , .		0
2011	Dimension-Prompts Boost Commonsense Consolidation. , 2023, , .		0
2012	MAMO: Fine-Grained Vision-Language Representations Learning with Masked Multimodal Modeling. , 2023, , .		2
2015	Automatic Sentiment Labelling ofÂMultimodal Data. Communications in Computer and Information Science, 2023, , 154-175.	0.4	1
2018	Efficient Augmentation for Imbalanced Deep Learning. , 2023, , .		1
2021	A Symmetric Dual Encoding Dense Retrieval Framework for Knowledge-Intensive Visual Question Answering. , 2023, , .		2
2031	CSA-BERT: Video Question Answering. , 2023, , .		0
2033	FashionVQA: A Domain-Specific Visual Question Answering System. , 2023, , .		0
2034	SSGVS: Semantic Scene Graph-to-Video Synthesis. , 2023, , .		1
2035	Visual Semantic Relatedness Dataset for Image Captioning. , 2023, , .		0
2036	Is Multimodal Vision Supervision Beneficial to Language?. , 2023, , .		0
2039	Multi-Label Ranking: Mining Multi-Label and Label Ranking Data. , 2023, , 511-535.		0
2040	MemeGraphs: Linking Memes toÂKnowledge Graphs. Lecture Notes in Computer Science, 2023, , 534-551.	1.0	1
2041	TextREC: A Dataset forÂReferring Expression Comprehension withÂReading Comprehension. Lecture Notes in Computer Science, 2023, , 402-420.	1.0	0

#	Article	IF	Citations
2045	Towards Confidence-Aware Commonsense Knowledge Integration for Scene Graph Generation. , 2023, , .		0
2046	Knowledge Prompt Makes Composed Pre-Trained Models Zero-Shot News Captioner. , 2023, , .		0
2047	Difference-Aware Iterative Reasoning Network for Key Relation Detection. , 2023, , .		0
2048	Addressing Predicate Overlap in Scene Graph Generation with Semantic Granularity Controller. , 2023, , .		0
2050	Leveraging Knowledge Graphs for CheapFakes Detection: Beyond Dataset Evaluation. , 2023, , .		0
2051	Aesthetic Visual Question Answering of Photographs. , 2023, , .		0
2054	Naive Scene Graphs: How Visual is Modern Visual Relationship Detection?. , 2023, , .		0
2062	Anime Character Identification and Tag Prediction by Multimodality Modeling: Dataset and Model. , 2023, , .		0
2071	Enhancing Dynamic Image Advertising with Vision-Language Pre-training. , 2023, , .		1
2073	Contextual andÂSemantic Novelty inÂText. Synthesis Lectures on Computer Vision, 2024, , 71-95.	0.4	0
2074	BIT: Improving Image-text Sentiment Analysis via Learning Bidirectional Image-text Interaction. , 2023, , .		1
2075	A Novel End-to-End Transformer for Scene Graph Generation. , 2023, , .		0
2076	A Novel Cross-Fusion Method of Different Types of Features for Image Captioning. , 2023, , .		0
2077	Generating Questions via Unexploited OCR Texts: Prompt-Based Data Augmentation for TextVQA. , 2023, , .		1
2078	Pre-training A Prompt Pool for Vision-Language Model. , 2023, , .		0
2080	Improving Visual Question Answering by Multimodal Gate Fusion Network. , 2023, , .		0
2081	Augmented Spatial Context Fusion Network for Scene Graph Generation. , 2023, , .		0
2082	Towards Few-shot Image Captioning with Cycle-based Compositional Semantic Enhancement Framework. , 2023, , .		0

#	Article	IF	Citations
2084	3DSSR: 3D Subscene Retrieval. , 2023, , .		0
2085	Scene Graph Driven Text-Prompt Generation for Image Inpainting. , 2023, , .		1
2086	Reading Between theÂLanes: Text VideoQA onÂtheÂRoad. Lecture Notes in Computer Science, 2023, , 137-154.	1.0	1
2087	Text-Visual Prompting for Efficient 2D Temporal Video Grounding. , 2023, , .		4
2088	Text with Knowledge Graph Augmented Transformer for Video Captioning. , 2023, , .		3
2089	Understanding and Constructing Latent Modality Structures in Multi-Modal Representation Learning. , 2023, , .		0
2090	Open-vocabulary Attribute Detection. , 2023, , .		3
2091	ViLEM: Visual-Language Error Modeling for Image-Text Retrieval. , 2023, , .		0
2092	Uni-Perceiver v2: A Generalist Model for Large-Scale Vision and Vision-Language Tasks. , 2023, , .		3
2093	Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning. , 2023, , .		2
2094	Clover: Towards A Unified Video-Language Alignment and Fusion Model. , 2023, , .		3
2095	Connecting Vision and Language with Video Localized Narratives. , 2023, , .		1
2096	3D Spatial Multimodal Knowledge Accumulation for Scene Graph Prediction in Point Cloud. , 2023, , .		1
2097	Non-Contrastive Learning Meets Language-Image Pre-Training. , 2023, , .		0
2098	OvarNet: Towards Open-Vocabulary Object Attribute Recognition. , 2023, , .		2
2099	Position-Guided Text Prompt for Vision-Language Pre-Training. , 2023, , .		2
2100	Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training. , 2023, , .		6
2101	VindLU: A Recipe for Effective Video-and-Language Pretraining. , 2023, , .		4

#	Article	IF	Citations
2102	Seeing What You Miss: Vision-Language Pre-training with Semantic Completion Learning. , 2023, , .		2
2103	CNVid-3.5M: Build, Filter, and Pre-Train the Large-Scale Public Chinese Video-Text Dataset. , 2023, , .		1
2104	Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding. , 2023, , .		0
2105	Prompting Large Language Models with Answer Heuristics for Knowledge-Based Visual Question Answering. , 2023, , .		18
2106	ScaleDet: A Scalable Multi-Dataset Object Detector. , 2023, , .		0
2107	Divide and Conquer: Answering Questions with Object Factorization and Compositional Reasoning. , 2023, , .		2
2108	Reproducible Scaling Laws for Contrastive Language-Image Learning. , 2023, , .		21
2109	An Empirical Study of End-to-End Video-Language Transformers with Masked Visual Modeling. , 2023, , .		4
2110	Probabilistic Debiasing of Scene Graphs. , 2023, , .		2
2111	LayoutDiffusion: Controllable Diffusion Model for Layout-to-Image Generation. , 2023, , .		4
2112	Panoptic Video Scene Graph Generation. , 2023, , .		3
2113	KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation. , 2023, , .		2
2114	Exploring Structured Semantic Prior for Multi Label Recognition with Incomplete Labels. , 2023, , .		0
2115	MVImgNet: A Large-scale Dataset of Multi-view Images. , 2023, , .		4
2116	All in One: Exploring Unified Video-Language Pre-Training. , 2023, , .		21
2117	FlexiViT: One Model for All Patch Sizes. , 2023, , .		4
2118	RefTeacher: A Strong Baseline for Semi-Supervised Referring Expression Comprehension. , 2023, , .		3
2119	MAP: Multimodal Uncertainty-Aware Vision-Language Pre-training Model. , 2023, , .		1

#	Article	IF	Citations
2120	ConStruct-VL: Data-Free Continual Structured VL Concepts Learning*. , 2023, , .		1
2121	Unbiased Scene Graph Generation in Videos. , 2023, , .		2
2122	What Can Human Sketches Do for Object Detection?. , 2023, , .		13
2123	Advancing Visual Grounding with Scene Knowledge: Benchmark and Method. , 2023, , .		2
2124	Few-Shot Referring Relationships in Videos. , 2023, , .		1
2125	Multimodal Prompting with Missing Modalities for Visual Recognition. , 2023, , .		4
2126	RefCLIP: A Universal Teacher for Weakly Supervised Referring Expression Comprehension. , 2023, , .		3
2127	Detection Hub: Unifying Object Detection Datasets via Query Adaptation on Language Embedding. , 2023, , .		2
2128	Learning Emotion Representations from Verbal and Nonverbal Communication. , 2023, , .		1
2129	Learning to Generate Language-Supervised and Open-Vocabulary Scene Graph Using Pre-Trained Visual-Semantic Space. , 2023, , .		2
2130	VQACL: A Novel Visual Question Answering Continual Learning Setting. , 2023, , .		0
2131	Semantic-Conditional Diffusion Networks for Image Captioning*. , 2023, , .		7
2132	Improving Visual Grounding by Encouraging Consistent Gradient-Based Explanations. , 2023, , .		2
2133	Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles. , 2023, , .		1
2134	Fine-grained Image-text Matching by Cross-modal Hard Aligning Network. , 2023, , .		5
2135	Teaching Structured Vision & Language Concepts to Vision & Language Models. , 2023, , .		2
2136	Dynamic Inference with Grounding Based Vision and Language Models. , 2023, , .		0
2137	IS-GGT: Iterative Scene Graph Generation with Generative Transformers. , 2023, , .		2

#	Article	IF	Citations
2138	CapDet: Unifying Dense Captioning and Open-World Detection Pretraining. , 2023, , .		4
2139	PolyFormer: Referring Image Segmentation as Sequential Polygon Generation. , 2023, , .		5
2140	Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. , 2023, , .		13
2141	PACO: Parts and Attributes of Common Objects. , 2023, , .		2
2142	MIST : Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering. , 2023, , .		5
2143	Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks. , 2023, , .		73
2144	Affection: Learning Affective Explanations for Real-World Visual Data. , 2023, , .		2
2145	S³C: Semi-Supervised VQA Natural Language Explanation via Self-Critical Learning. , 2023, , .		0
2146	GRES: Generalized Referring Expression Segmentation. , 2023, , .		6
2147	Accelerating Vision-Language Pretraining with Free Language Modeling. , 2023, , .		1
2148	Multi-Modal Representation Learning with Text-Driven Soft Masks. , 2023, , .		2
2149	Open-Category Human-Object Interaction Pre-training via Language Modeling Framework. , 2023, , .		0
2150	HAAV: Hierarchical Aggregation of Augmented Views for Image Captioning. , 2023, , .		2
2151	BBDM: Image-to-Image Translation with Brownian Bridge Diffusion Models. , 2023, , .		4
2152	Generalized Decoding for Pixel, Image, and Language. , 2023, , .		14
2153	Leveraging per Image-Token Consistency for Vision-Language Pre-training. , 2023, , .		1
2154	NÃœWA-LIP: Language-guided Image Inpainting with Defect-free VQGAN. , 2023, , .		4
2155	Hyperbolic Contrastive Learning for Visual Representations beyond Objects. , 2023, , .		3

#	Article	IF	Citations
2156	Scaling Language-Image Pre-Training via Masking. , 2023, , .		21
2157	GLIGEN: Open-Set Grounded Text-to-Image Generation. , 2023, , .		28
2158	Detecting Everything in the Open World: Towards Universal Object Detection. , 2023, , .		8
2159	Uncurated Image-Text Datasets: Shedding Light on Demographic Bias. , 2023, , .		2
2160	LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling. , 2023, , .		9
2161	Fast Contextual Scene Graph Generation with Unbiased Context Augmentation. , 2023, , .		1
2162	Cross-Modal Representation Learning. , 2023, , 211-240.		1
2164	Scene Graph Generation using Depth-based Multimodal Network. , 2023, , .		0
2169	USGG: Union Message Based Scene Graph Generation. , 2023, , .		0
2170	Interpretable Visual Question Answering Via Reasoning Supervision. , 2023, , .		0
2171	Consistent and Multi-Scale Scene Graph Transformer for Semantic-Guided Image Outpainting. , 2023, , .		0
2180	Image Caption withÂPrior Knowledge Graph andÂHeterogeneous Attention. Lecture Notes in Computer Science, 2023, , 344-356.	1.0	0
2181	A Balanced Relation Prediction Framework forÂScene Graph Generation. Lecture Notes in Computer Science, 2023, , 216-228.	1.0	0
2186	3D Scene Graph Prediction on Point Clouds Using Knowledge Graphs. , 2023, , .		0
2188	PMC-CLIP: Contrastive Language-Image Pre-training Using Biomedical Documents. Lecture Notes in Computer Science, 2023, , 525-536.	1.0	3
2190	Image Caption Method with Verb-Specific Scene Graph Decomposition. , 2023, , .		0
2195	Alleviating Training Bias with Less Cost via Multi-expert De-biasing Method in Scene Graph Generation. , 2023, , .		0
2198	Multi-View Predicate Recognition for Solving Semantic Ambiguity Problem in Scene Graph Generation. , 2023, , .		0

#	Article	IF	Citations
2201	The Potential of a Visual Dialogue Agent In a Tandem Automated Audio Description System for Videos. , 2023, , .		0
2204	Video Referring Expression Comprehension via Transformer with Content-conditioned Query. , 2023, , .		0
2208	MMpedia: A Large-Scale Multi-modal Knowledge Graph. Lecture Notes in Computer Science, 2023, , 18-37.	1.0	0
2209	Answer-Based Entity Extraction and Alignment for Visual Text Question Answering. , 2023, , .		0
2210	Enhanced CatBoost with Stacking Features for Social Media Prediction. , 2023, , .		0
2211	Food-500 Cap: A Fine-Grained Food Caption Benchmark for Evaluating Vision-Language Models. , 2023, , .		0
2212	Improving Scene Graph Generation with Superpixel-Based Interaction Learning. , 2023, , .		0
2213	Beware of Overcorrection: Scene-induced Commonsense Graph for Scene Graph Generation. , 2023, , .		0
2215	Multi-modal Context-Aware Network forÂScene Graph Generation. Lecture Notes in Computer Science, 2023, , 335-347.	1.0	0
2216	Disambiguation of Visual Representations. , 2023, , .		0
2218	Text-based Person Search in Full Images via Semantic-Driven Proposal Generation. , 2023, , .		1
2219	Multi-head Similarity Feature Representation andÂFiltration forÂImage-Text Matching. Lecture Notes in Computer Science, 2023, , 629-643.	1.0	0
2227	Bidirectional Edge-Based 3D Scene Graph Generation from Point Clouds. , 2023, , .		0
2233	Using Markov Random Field (MRF) Hypergraph Transformer Method for Visual Question Answering (VQA) Application. , 2023, , .		0
2239	Deep Learning Misconduct and How Conscious Learning Avoids it. Artificial Intelligence, 0, , .	2.0	0
2241	Cross-Lingual Transfer of Large Language Model by Visually-Derived Supervision Toward Low-Resource Languages. , 2023, , .		0
2242	SpaceCLIP: A Vision-Language Pretraining Framework With Spatial Reconstruction On Text. , 2023, , .		0
2243	OCF-Layout2Img: Object-Wise Constraint Free Layout to Image Generation. , 2023, , .		0

#	Article	IF	Citations
2244	Towards Deconfounded Image-Text Matching with Causal Inference. , 2023, , .		1
2245	Language-Guided Visual Aggregation Network for Video Question Answering. , 2023, , .		0
2246	Dark Knowledge Balance Learning for Unbiased Scene Graph Generation. , 2023, , .		0
2247	Progressive Positive Association Framework for Image and Text Retrieval. , 2023, , .		0
2248	All in One: Exploring Unified Vision-Language Tracking with Multi-Modal Alignment. , 2023, , .		0
2249	CCMB: A Large-scale Chinese Cross-modal Benchmark. , 2023, , .		0
2250	BLAT: Bootstrapping Language-Audio Pre-training based on AudioSet Tag-guided Synthetic Data. , 2023, , .		0
2251	A Symbolic Characters Aware Model for Solving Geometry Problems. , 2023, , .		0
2252	Open-Vocabulary Object Detection via Scene Graph Discovery. , 2023, , .		0
2253	A Baseline Investigation: Transformer-based Cross-view Baseline for Text-based Person Search. , 2023, , .		0
2254	Unlocking the Power of Cross-Dimensional Semantic Dependency for Image-Text Matching. , 2023, , .		0
2255	VTQAGen: BART-based Generative Model For Visual Text Question Answering. , 2023, , .		0
2256	ChinaOpen: A Dataset for Open-world Multimodal Learning. , 2023, , .		0
2257	Double-Fine-Tuning Multi-Objective Vision-and-Language Transformer for Social Media Popularity Prediction. , 2023, , .		0
2258	LUNA: Language as Continuing Anchors for Referring Expression Comprehension. , 2023, , .		0
2259	COPA : Efficient Vision-Language Pre-training through Collaborative Object- and Patch-Text Alignment. , 2023, , .		0
2260	Focusing on Flexible Masks: A Novel Framework for Panoptic Scene Graph Generation with Relation Constraints. , 2023, , .		0
2261	Enhancing Vision-Language Pre-Training with Jointly Learned Questioner and Dense Captioner. , 2023, , .		0

#	Article	IF	Citations
2264	SGDraw: Scene Graph Drawing Interface Using Object-Oriented Representation. Lecture Notes in Computer Science, 2023, , 211-226.	1.0	0
2265	GVCCI: Lifelong Learning of Visual Grounding for Language-Guided Robotic Manipulation. , 2023, , .		0
2266	Trust perception based human-robot collaborative indoor target search. , 2023, , .		0
2271	A Review on VQA: Methods, Tools and Datasets. , 2023, , .		0
2274	An Effective Dynamic Reweighting Method forÂUnbiased Scene Graph Generation. Lecture Notes in Computer Science, 2024, , 345-356.	1.0	0
2276	ClipCrop: Conditioned Cropping Driven by Vision-Language Model. , 2023, , .		0
2278	Haystack: A Panoptic Scene Graph Dataset to Evaluate Rare Predicate Classes. , 2023, , .		1
2279	Chest X-Ray Feature Pyramid Sum Model with Diseased Area Data Augmentation Method. , 2023, , .		0
2280	Painter: Teaching Auto-regressive Language Models to Draw Sketches. , 2023, , .		0
2281	Video-and-Language (VidL) models and their cognitive relevance. , 2023, , .		0
2282	Black-Box Attacks on Image Activity Prediction and its Natural Language Explanations. , 2023, , .		0
2283	MERGE: Multi-Entity Relational Reasoning Based Explanation in Visual Question Answering. , 2023, , .		0
2284	Knowledge Informed Sequential Scene Graph Verification Using VQA. , 2023, , .		0
2286	Fine-Grained is Too Coarse: A Novel Data-Centric Approach for Efficient Scene Graph Generation. , 2023, , .		0
2287	VLMAH: Visual-Linguistic Modeling of Action History for Effective Action Anticipation. , 2023, , .		0
2288	Which Tokens to Use? Investigating Token Reduction in Vision Transformers. , 2023, , .		1
2289	SceneGenie: Scene Graph Guided Diffusion Models for Image Synthesis. , 2023, , .		1
2290	Relational Prior Knowledge Graphs for Detection and Instance Segmentation. , 2023, , .		0

#	Article	IF	Citations
2291	UDA-HOID: Unsupervised Domain Adaptation for Human-Object Interaction Detection. , 2023, , .		0
2296	Transformer-based Deep Embedding Network for Scene Graph Generation. , 2023, , .		0
2300	A Knowledge Acquisition Framework for Autonomous Decision Making in Service Robots. , 2023, , .		0
2301	Neural-Based Cross-Modal Search and Retrieval of Artwork. , 2023, , .		0
2303	A Fine-Grained Image Description Generation Method Based onÂJoint Objectives. Communications in Computer and Information Science, 2024, , 32-46.	0.4	0
2310	TextPSG: Panoptic Scene Graph Generation from Textual Descriptions. , 2023, , .		0
2311	Saliency Regularization for Self-Training with Partial Annotations. , 2023, , .		0
2312	EdaDet: Open-Vocabulary Object Detection Using Early Dense Alignment. , 2023, , .		0
2313	Learning Trajectory-Word Alignments for Video-Language Tasks. , 2023, , .		0
2314	VQA Therapy: Exploring Answer Differences by Visually Grounding Answers. , 2023, , .		0
2315	SMAUG: Sparse Masked Autoencoder for Efficient Video-Language Pre-training. , 2023, , .		0
2316	Unified Visual Relationship Detection with Vision and Language Models. , 2023, , .		1
2317	OmniLabel: A Challenging Benchmark for Language-Based Object Detection. , 2023, , .		0
2318	Noise-aware Learning from Web-crawled Image-Text Data for Image Captioning. , 2023, , .		0
2319	Equivariant Similarity for Vision-Language Foundation Models. , 2023, , .		0
2320	Confidence-aware Pseudo-label Learning for Weakly Supervised Visual Grounding. , 2023, , .		0
2321	Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks. , 2023, , .		0
2322	Visually-Prompted Language Model for Fine-Grained Scene Graph Generation in an Open World. , 2023, , .		2

#	Article	IF	Citations
2323	V3Det: Vast Vocabulary Visual Detection Dataset. , 2023, , .		0
2324	Too Large; Data Reduction for Vision-Language Pre-Training. , 2023, , .		0
2325	Learning Human-Human Interactions in Images from Weak Textual Supervision. , 2023, , .		0
2326	Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. , 2023, , .		0
2327	Decouple Before Interact: Multi-Modal Prompt Learning for Continual Visual Question Answering. , 2023, , .		0
2328	HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. , 2023, , .		0
2329	Video OWL-ViT: Temporally-consistent open-world localization in video. , 2023, , .		0
2330	MOST: Multiple Object localization with Self-supervised Transformers for object discovery. , 2023, , .		0
2331	Tracking by Natural Language Specification with Long Short-term Context Decoupling. , 2023, , .		0
2332	Beyond Object Recognition: A New Benchmark towards Object Concept Learning. , 2023, , .		0
2333	In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval. , 2023, , .		0
2334	Unmasked Teacher: Towards Training-Efficient Video Foundation Models. , 2023, , .		2
2335	Helping Hands: An Object-Aware Ego-Centric Video Recognition Model. , 2023, , .		0
2336	Sentence Attention Blocks for Answer Grounding. , 2023, , .		0
2337	Compositional Feature Augmentation for Unbiased Scene Graph Generation. , 2023, , .		0
2338	Environment-Invariant Curriculum Relation Learning for Fine-Grained Scene Graph Generation. , 2023, , .		0
2339	Towards Models that Can See and Read. , 2023, , .		1
2340	Learning Navigational Visual Representations with Semantic Map Supervision. , 2023, , .		1

#	Article	IF	Citations
2341	Detecting Objects with Context-Likelihood Graphs and Graph Refinement. , 2023, , .		0
2342	ViLTA: Enhancing Vision-Language Pre-training through Textual Augmentation. , 2023, , .		1
2343	Going Denser with Open-Vocabulary Part Segmentation. , 2023, , .		0
2344	Going Beyond Nouns With Vision & Language Models Using Synthetic Data. , 2023, , .		0
2345	Vision Relation Transformer for Unbiased Scene Graph Generation. , 2023, , .		0
2346	Advancing Referring Expression Segmentation Beyond Single Image. , 2023, , .		1
2347	Weakly Supervised Referring Image Segmentation with Intra-Chunk and Inter-Chunk Consistency. , 2023, , .		0
2348	ViM: Vision Middleware for Unified Downstream Transferring. , 2023, , .		0
2349	RealGraph: A Multiview Dataset for 4D Real-world Context Graph Generation. , 2023, , .		0
2350	Hierarchical Visual Primitive Experts for Compositional Zero-Shot Learning. , 2023, , .		0
2351	Segment Every Reference Object in Spatial and Temporal Spaces. , 2023, , .		0
2352	VL-Match: Enhancing Vision-Language Pretraining with Token-Level and Instance-Level Matching. , 2023, , .		0
2353	HiLo: Exploiting High Low Frequency Relations for Unbiased Panoptic Scene Graph Generation. , 2023, , .		0
2354	HiVLP: Hierarchical Interactive Video-Language Pre-Training. , 2023, , .		1
2355	Toward Multi-Granularity Decision-Making: Explicit Visual Reasoning with Hierarchical Knowledge. , 2023, , .		0
2356	SGAligner: 3D Scene Alignment with Scene Graphs. , 2023, , .		0
2357	RLIPv2: Fast Scaling of Relational Language-Image Pre-training. , 2023, , .		0
2358	Shatter and Gather: Learning Referring Image Segmentation with Text Supervision. , 2023, , .		0

#	Article	IF	Citations
2359	Multi3DRefer: Grounding Text Description to Multiple 3D Objects. , 2023, , .		0
2360	LAW-Diffusion: Complex Scene Generation by Diffusion with Layouts. , 2023, , .		0
2361	SLAN: Self-Locator Aided Network for Vision-Language Understanding. , 2023, , .		0
2362	PreSTU: Pre-Training for Scene-Text Understanding. , 2023, , .		0
2363	SINC: Self-Supervised In-Context Learning for Vision-Language Tasks. , 2023, , .		0
2364	BUS : Efficient and Effective Vision-language Pre-training with Bottom-Up Patch Summarization. , 2023, , .		0
2365	Who are you referring to? Coreference resolution in image narrations. , 2023, , .		0
2368	Evaluating CLIPâ€™s Understanding on Relationships in a Blocks World. , 2023, , .		0
2369	Multimodal Large Language Models: A Survey. , 2023, , .		2
2374	Appearance-Motion Dual-Stream Heterogeneous Network forÂVideoQA. Lecture Notes in Computer Science, 2024, , 212-227.	1.0	0
2375	Structure-Aware Adaptive Hybrid Interaction Modeling forÂImage-Text Matching. Lecture Notes in Computer Science, 2024, , 327-341.	1.0	0
2377	Noise-Augmented Missing Modality Aware Prompt Based Learning for Robust Visual Recognition. , 2023, , .		0
2379	Phase Discriminated Multi-Policy for Visual Room Rearrangement. , 2023, , .		0
2380	Multi-modal Domain Adaptation for Text Visual Question Answering Tasks. , 2023, , .		0
2384	S-VQA: Sentence-Based Visual Question Answering. , 2023, , .		0
2387	Cross-Modal Retrieval Based onÂSemantic Filtering andÂAdaptive Pooling. Lecture Notes in Electrical Engineering, 2024, , 296-310.	0.3	0
2388	Cross Modal Retrieval Algorithm Based on Iterative Queries. Lecture Notes in Electrical Engineering, 2024, , 332-344.	0.3	0
2390	Weakly-Supervised Grounding forÂVQA withÂDual Visual-Linguistic Interaction. Lecture Notes in Computer Science, 2024, , 156-169.	1.0	0

#	Article	IF	Citations
2392	DSG: An End-to-End Document Structure Generator. , 2023, , .		0
2394	Semi-Supervised Implicit Augmentation for Data-Scarce VQA. , 0, , .		0
2399	Visual Question Generation Using Deep Learning. , 2023, , .		0
2406	MSAM: Deep Semantic Interaction Network forÂVisual Question Answering. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 2024, , 39-56.	0.2	0
2411	Does theÂPerformance ofÂText-to-Image Retrieval Models Generalize Beyond Captions-as-a-Query?. Lecture Notes in Computer Science, 2024, , 161-176.	1.0	0
2412	Exploring Multimodal Features for Sentiment Classification of Social Media Data. Lecture Notes in Networks and Systems, 2024, , 527-537.	0.5	0
2413	Natural Language Navigation for Robotic Systems: Integrating GPT and Dense Captioning Models with Object Detection in Autonomous Inspections. , 2024, , .		0
2416	PoseTron: Enabling Close-Proximity Human-Robot Collaboration Through Multi-human Motion Prediction. , 2024, , .		0
2420	In Defense ofÂScene Graph Generation forÂHuman-Robot Open-Ended Interaction inÂService Robotics. Lecture Notes in Computer Science, 2024, , 299-310.	1.0	0
2422	A Balanced Counting Visual Question Answering Dataset. Lecture Notes in Networks and Systems, 2024, , 509-518.	0.5	0