Preface

On behalf of the Organizing Committee, it is our great pleasure to welcome you to the 2nd ACM International Conference on Multimedia in Asia (MMAsia 2020). MMAsia is a merge of the long-lasting experience of the former PCM and ICIMCS, which both have good history as well as attending experiences. Officially sponsored by ACM SIGMM, MMAsia is a newly established international conference to showcase the scientific achievements and industrial innovations in the multimedia field. Its mission is to illuminate the state of the art in multimedia computing by bringing together researchers and practitioners in this field.

In 2020, ACM Multimedia Asia is planned to hold in Dec 2020 Singapore with an extensive program that includes technical sessions covering all aspects of the multimedia field in forms of oral and poster presentations, tutorials, panels, brave new idea, doctoral symposium and grand challenge competitions. Unfortunately, due to the COVID19 pandemic, we have to postpone the event to March 2021 and at last to move online entirely.

Thanks to the efforts of our community, we still have received 93 submissions to the conference this year. These submissions cover widely in the areas of large-scale multimedia analysis and retrieval, multimedia systems and applications, multimedia communications and transmission, multimedia security and quality assessment, mobile multimedia computing, social multimedia analysis, computer vision/machine learning for multimedia application, and so on. We thank our 13 area chairs and 115 Technical Program Committee members who spent many efforts reviewing papers and providing valuable feedback to the authors. From the total of 93 submissions and based on at least three effective reviews per submission, the Program Chairs decided to accept 17 oral papers and 24 poster papers. We also selected high-quality papers (22) from the rest of the papers to be combined in the special sessions. Among the 17 oral papers, four papers are selected as the best paper candidate, to compete for the Best Paper and the Best Paper Runner-Up awards.

The technical program is an important aspect but only provides its full impact if surrounded by challenging keynotes. We are extremely pleased and grateful to have three exceptional keynote speakers, Prof. Jiebo Luo, Prof. Kristen Grauman, and Prof. Bernt Schiele, to accept our invitation to present their insightful ideas and prospects. We would also like to express our sincere gratitude to all the other committee members, to help organize exciting sessions including special sessions, demos, tutorials, finance, and publication. Their contributions are much appreciated. It is their outstanding effort in preparing this rich and complex program that is the first virtual event in MMAsia.

We sincerely hope that you will enjoy your virtual experience and value your participation in MM Asia 2020.

General Chairs

Tat-Seng Chua, National University of Singapore
Jingdong Wang, Microsoft Research

Qi Tian, Huawei Noah's Ark

Program Chairs

Cathal Gurrin, Dublin City University
Jia Jia, Tsinghua University
Hanwang Zhang, Nanyang Technological University
Qianru Sun, Singapore Management University

Organization

General Chairs

Tat-Seng Chua	National University of Singapore
Jingdong Wang	Microsoft Research
Qi Tian	Huawei Noah's Ark

Technical Program Chairs

Cathal Gurrin	Dublin City University
Jia Jia	Tsinghua University
Hanwang Zhang	Nanyang Technological University
Qianru Sun	Singapore Management University

Special Session Chairs

Bing-Kun Bao	Nanjing University of Posts and Telecommunications
Xiangnan He	University of Science and Technology of China
Svebor Karaman	Dataminr
Asako Kanezaki	Tokyo Institute of Technology

Doctor Symposium Chairs

Wei-Chen Chiu	National Chiao Tung University
Si Liu	Beihang University

Demo Chairs

Qiuhong Ke	The University of Melbourne
Liqiang Nie	Shandong University
Jingkuan Song	University of Electronic Science and Technology of China

Tutorial Chairs

Anna Khoreva	Bosch Center for Artificial Intelligence
Meng Wang	Hefei University of Technology

Panel Chairs

Jiaying Liu	Peking University
Qi Wu	The University of Adelaide

Brave New Idea Chairs

Chen Change Loy	Nanyang Technological University
Yi Yu	National Institute of Informatics
Zheng-Jun Zha	University of Science and Technology of China

Grand Challenge Chairs

Fanglin Wang	ADVANCE.AI
Joao Magalhaes	Universidade Nova De Lisboa

Publicity Chairs

Richang Hong	Hefei University of Technology
Roger Zimmermann	National University of Singapore
Benoit Huet	EURECOM
Changsheng Xu	Institute of Automation, Chinese Academy of Sciences

Finance Chairs

Jing Liu	Institute of Automation, Chinese Academy of Sciences
Lu Jin	Nanyang Technological University

Publication Chairs

Tian Gan	Shandong University
Tianzhu Zhang	University of Science and Technology of China

Website Masters

Yaoyao Liu	Max Planck Institute for Informatics
Xin Fu	Beijing Jiaotong University

Program Committee Members

Ahmed Alateeq	DCU
Andreas Leibetseder	Alpen-Adria-Universitat Klagenfurt
Boi Mai Quach	DCU
Brian Chen	Columbia University
Cathal Gurrin	Dublin City University
Chang Tang	China University of Geosciences
Chaoqun Zheng	Shandong Normal University
Christian Beecks	University of Münster
Chunhui Bao	singapore management university
Dang Huynh	AISIA Lab
Daqing Liu	University of Science and Technology of China
Dejing Xu	Tencent
Dong Wang	Dalian University of Technology
Dong Zhang	Nanjing University of Science and Technology
Duc-Tien Dang-Nguyen	University of Bergen
Duy Nguyen Ho Minh	Max Planck Institute for Informatics/ Saarland University
Duyen LTran	Dublin City University
Fan Liu	Hohai University
Fan Ma	University of Technology Sydney
Guang Yu	National University of Defense Technology
Guansong Pang	University of Adelaide
Hao Zhang	City University of Hong Kong
Haozhe Wu	Tsinghua University
Hien D. Nguyen	University of Information Technology
Hoayu Tang	Xi'an Jiaotong University
Hongtao Xie	University of Science and Technology of China
Hualin Liu	Salesforce
Hui Cui	Shandong Normal Unversity
Huiyuan Yang	Binghamton University-SUNY
Hung T Nguyen	Tokyo University of Agriculture and Technology
Ichiro Ide	Nagoya University
Jakub Lokoc	Charles University in Prague
Jiali Xi	Shanghai Jiao Tong University
Jialiang Sun	University of Electronic Science and Technology of China
Jianjun Qian	Nanjing University of Science and Technology
Jiaxin Qi	Nanyang Technological University
Jiefu Chen	University of Electronic Science and Technology of China
Jingjing Li	University of Electronic Science and Technology of China
Jingran Zhang	University of Electronic Science and Technology of China
Jingyuan Chen	Damo Academy, Alibaba Group
Jiuxiang Gu	Adobe Research
Jiwei Wei	University of Electronic Science and Technology of China

Julia Dietlmeier	Insight SFI Research Centre for Data Analytics
Junbin Xiao	National University of Singapore
Junge Shen	Northwestern Polytechnical University
Kevin McGuinness	Insight Centre for Data Analytics
Kun Kuang	Zhejiang University
Liang Xie	WHUT
Long Chen	Columbia University
Luca Rossetto	University of Zurich
Manh-Duy Nguyen	Dublin City University
Maria Eskevich	CLARIN ERIC
Mario Taschwer	Klagenfurt University
Markus Fox	Klagenfurt University
Markus Fox	University of Klagenfurt
Meng Liu	Shandong Jianzhu University
Mingli Song	Zhejiang University
Mingyu Ding	The University of Hong Kong
Minh-Son Dao	National Institute of Information and Communications Technology
Natalia Sokolova	Klagenfurt University
Negin Ghamsarian	Alpen-Adria University of Klagenfurt
Ning Xu	Tianjin University
Ninh Van Tu	DCU
Quang Pham	SMU
Ralph Gasser	University of Basel
Runmin Cong	Beijing Jiaotong University
Shang-Liang Chen	National Cheng-Kung university
ShiKai Chen	Southeast University
Shuai Yang	Nanyang Technological University
Shuzhong Lin	Tianjin Polytechnic University
Sicheng Yu	Singapore Management University
Tan Wang	Nanyang Technological University
Tao He	Monash University
Tao Zhuo	National University of Singapore
Thanh Binh Nguyen	University of Science
Thuong Khanh Tran	University of Oulu
Tianyu Yang	Tencent AI Lab
Tu-Khiem Le	DCU
Wei Ji	Zhejiang University
Wei QIN	Hefei University of Technology
Weizhi Nie	Tianjin University
Wenqi Ren	Chinese Academy of Science
Wenqing Chu	Tencent
Werner Bailer	JOANNEUM RESEARCH
Xi Peng	Sichuan University
Xiaojiang Peng	ShenZhen Key Lab of Computer Vision and Pattern Recognition
Xin Fu	Beijing Jiaotong University

Xin Luo	Shandong University
Xingbo Liu	Shandong University
Xinggang Wang	Huazhong University of Science and Technology
Xinwang Liu	National University of Defense Technology
Xiuyi Jia	Nanjing University of Science and Technology
Xu Yang	Nanyang Technological University
Xu Yang	Nanjing University of Science & Technology
Yanbin Hao	University of Science and Technology of China
Yang Li	Zhejiang University
Yangyang Guo	Shandong University
Yaoyao Liu	Max Planck Institute for Informatics
Yawei Luo	Zhejiang University
Yi Jiang	Bytedance
Yifan Wang	University of Electronic Science and Technology of China
Yifeng Zhou	University of Electronic Science and Technology of China
Yongcheng Jing	The University of Sydney
Yuanen Zhou	Hefei University of Technology
Yue Liao	Beihang University
Zhanzhan Cheng	Zhejiang University & Hikvision Research Institute
Zhaoquan Yuan	School of Computing and Artificial Intelligence, Southwest Jiaotong University
Zhaozheng Chen	Singapore Management University
Zheng Wang	UESTC
Zhifan Gao	Sun Yat-sen University
Zhongqi Yue	Nanyang Technological University
Zijie Ye	Tsinghua University
Fudong Nian	Hefei University
Junyu Gao	CASIA
Li Su	University of Chinese Academy of Sciences
Weiqing Min	Institute of Computing Technology, Chinese Academy of Sciences

Table of Contents

16 Self-Supervised Adversarial Learning for Cross-Modal Retrieval

17 Multi-Level Expression Guided Attention Network for Referring Expression Comprehension

18 Adaptive Feature Aggregation Network for Nuclei Segmentation

19 Classification of Multimedia SNS Posts about Tourist Sites Based on Their Focus toward Predicting Eco-Friendly Users

20 Learning Intra-inter Semantic Aggregation for Video Object Detection

21 Robust Visual Tracking via Scale-Aware Localization and Peak Response Strength

22 Hungry Networks: 3D Mesh Reconstruction of a Dish and a Plate from a Single Dish Image for Estimating Food Volume

23 Scene Graph Generation via Multi-Relation Classification and Cross-modal Atention Coordinator

24 A Novel System Architecture and an Automatic Monitoring Method for Remote Production

25 Graph Convolution Network with Node Feature Optimization Using Cross Attention for Few-shot Learning

26 A Multi-Scale Language Embedding Network for Proposal-Free Referring Expression Comprehension

27 Similar Scene Retrieval in Soccer Videos with Weak Annotations by Multimodal Use of Bidirectional LSTM

28 Patch Assembly for Real-time Instance Segmentation

29 Full-Resolution Encoder–Decoder Networks with Multi-Scale Feature Fusion for Human Pose Estimation

30 Graph-based Variational Auto-Encoder for Generalized Zero-Shot Learning

46 Intermediate Coordinate based Pose Non-perspective Estimation from Line Correspondences

47 An Autoregressive Generation Model for Producing Instant Basketball Defensive Trajectory

48 Real-Time Arbitrary Video Style Transfer

49 C3VQG: Category Consistent Cyclic Visual Question Generation

50 Determining Image Age with Rank-Consistent Ordinal Classification and Object-centered Ensemble

51 Cross-Modal Learning for Saliency Prediction in Mobile Environment

52 Objective Object Segmentation Visual Quality Evaluation based on Pixel-Level and Region-Level Characteristics

53 Text-based Visual Question Answering with Knowledge Base

54 Attention-Constraint Facial Expression Recognition

55 Defense for Adversarial Videos by Self-adaptive JPEG Compression and Optical Texture

56 Fusing CAMs-Weighted Features and Temporal Information for Robust Loop Closure Detection

57 Fixations Based Personal Target Objects Segmentation

58 Improving auto-encoder novelty detection using channel attention and entropy minimization

59 Relationship Graph Learning Network For Visual Relationship Detection

60 Local Structure Alignment Guided Domain Adaptation with Few Source Samples

61 Multiplicative Angular Margin Loss for Text-Based Person Search

62 Integrating Aspect-aware Interactive Attention and Emotional Position-aware for Multi-aspect Sentiment Analysis

63 Graph-Based Motion Prediction for Abnormal Action Detection

64 Attention Feature Matching for Weakly-supervised Video Relocalization

65 Pulse Localization Networks with Infrared Camera

66 Structure-Preserving Extremely Low Light Image Enhancement with Fractional Order Differential Mask Guidance

67 Change Detection from Synthetic Aperture Radar Images Based on Deformable Residual Convolutional Neural Networks

68 Efficient Inter-image Relation Graph Neural Network Hashing for Scalable Image Retrieval

69 Synthesized 3D Model Suggestions with Smartphone Based MR to Modify the PreBuilt Environment: Interior Design

70 SeekSuspect : Retrieving Suspects from Criminal Datasets using Visual Memory

71 A Large-Scale Image Retrieval System for Everyday Scenes

72 Towards Annotation-Free Evaluation of Cross-Lingual Image Captioning

73 10 Years of Video Browser Showdown

MMAsia2020

Integrating aspect-aware interactive attention and emotional position-aware for multi-aspect sentiment analysis
Xiaoye Wang, Xiaowen Zhou, Zan Gao, Peng Yang, Xianbin Wen, Hongyun Ning
Proceeding • 2021

Graph-based motion prediction for abnormal action detection
Yao Tang, Lin Zhao, Zhaoliang Yao, Chen Gong, Jian Yang
Proceeding • 2021

Attention feature matching for weakly-supervised video relocalization
Haoyu Tang, Jihua Zhu, Zan Gao, Tao Zhuo, Zhiyong Cheng
Proceeding • 2021

Pulse localization networks with infrared camera
Bohong Yang, Kai Meng, Hong Lu, Xinyao Nie, Guanhao Huang, Jingjing Luo, Xing Zhu
Proceeding • 2021

Structure-preserving extremely low light image enhancement with fractional order differential mask guidance
Yijun Liu, Zhengning Wang, Ruixu Geng, Hao Zeng, Yi Zeng
Proceeding • 2021

Change detection from SAR images based on deformable residual convolutional neural networks
Junjie Wang, Feng Gao, Junyu Dong
Proceeding • 2021

Local structure alignment guided domain adaptation with few source samples
Yuying Cai, Jinfeng Li, Baodi Liu, Weifeng Liu, Kai Zhang, Changsheng Xu

Proceeding • 2021

Semantic feature augmentation for fine-grained visual categorization with few-sample training

Xiang Guan, Yang Yang, Zheng Wang, Jingjing Li

Proceeding • 2021

Unsupervised learning of co-occurrences for face images retrieval

Thomas Petit, Pierre Letessier, Stefan Duffner, Christophe Garcia

Proceeding • 2021

Hierarchical clustering via mutual learning for unsupervised person re-identification

Xu Xu, Liyan Zhang, Zhaomeng Huang, Guodong Du

Proceeding • 2021

Self-supervised adversarial learning for cross-modal retrieval

Yangchao Wang, Shiyuan He, Xing Xu, Yang Yang, Jingjing Li, Heng Tao Shen

Proceeding • 2021

Multi-level expression guided attention network for referring expression comprehension

Liang Peng, Yang Yang, Xing Xu, Jingjing Li, Xiaofeng Zhu

Proceeding • 2021

Adaptive feature aggregation network for nuclei segmentation

Ruizhe Geng, Zhongyi Huang, Jie Chen

Proceeding • 2021

Classification of multimedia SNS posts about tourist sites based on their focus toward predicting eco-friendly users
Naoto Kashiwagi, Tokinori Suzuki, Jounghun Lee, Daisuke Ikeda
Proceeding • 2021

Learning intra-inter semantic aggregation for video object detection
Jun Liang, Haosheng Chen, Kaiwen Du, Yan Yan, Hanzi Wang
Proceeding • 2021

Robust visual tracking via scale-aware localization and peak response strength
Ying Wang, Luo Xiong, Kaiwen Du, Yan Yan, Hanzi Wang
Proceeding • 2021

Hungry networks: 3D mesh reconstruction of a dish and a plate from a single dish image for estimating food volume
Shu Naritomi, Keiji Yanai
Proceeding • 2021

Scene graph generation via multi-relation classification and cross-modal attention coordinator
Xiaoyi Zhang, Zheng Wang, Xing Xu, Jiwei Wei, Yang Yang
Proceeding • 2021

A novel system architecture and an automatic monitoring method for remote production
Yasuhiro Mochida, Daisuke Shirai, Takahiro Yamaguchi, Seiki Kuwabara, Hideki Nishizawa
Proceeding • 2021

Graph convolution network with node feature optimization using cross attention for few-shot learning
Ying Liu, Yanbo Lei, Sheikh Faisal Rashid
Proceeding • 2021

A multi-scale language embedding network for proposal-free referring expression comprehension
Taijin Zhao, Hongliang Li, Heqian Qiu, Qingbo Wu, King Ngi Ngan
Proceeding • 2021

Similar scene retrieval in soccer videos with weak annotations by multimodal use of bidirectional LSTM
Tomoki Haruyama, Sho Takahashi, Takahiro Ogawa, Miki Haseyama
Proceeding • 2021

Patch assembly for real-time instance segmentation
Yutao Xu, Hanli Wang, Jian Zhu
Proceeding • 2021

Full-resolution encoder-decoder networks with multi-scale feature fusion for human pose estimation
Jie Ou, Mingjian Chen, Hong Wu
Proceeding • 2021

Graph-based variational auto-encoder for generalized zero-shot learning
Jiwei Wei, Yang Yang, Xing Xu, Yanli Ji, Xiaofeng Zhu, Heng Tao Shen
Proceeding • 2021

A multi-scale human action recognition method based on Laplacian pyramid depth motion images

Chang Li, Qian Huang, Xing Li, Qianhan Wu
Proceeding • 2021

Fixed-size video summarization over streaming data via non-monotone submodular maximization
Ganfeng Lu, Jiping Zheng
Proceeding • 2021

Overlap classification mechanism for skeletal bone age assessment
Pengyi Hao, Xuhang Xie, Tianxing Han, Cong Bai
Proceeding • 2021

Fixation guided network for salient object detection
Zhe Cui, Li Su, Weigang Zhang, Qingming Huang
Proceeding • 2021

Motion-transformer: self-supervised pre-training for skeleton-based action recognition
Yi-Bin Cheng, Xipeng Chen, Dongyu Zhang, Liang Lin
Proceeding • 2021

Interactive re-ranking for cross-modal retrieval based on object-wise question answering
Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama
Proceeding • 2021

A background-induced generative network with multi-level discriminator for text-to-image generation
Ping Wang, Li Liu, Huaxiang Zhang, Tianshi Wang
Proceeding • 2021

WFN-PSC: weighted-fusion network with poly-scale convolution for image dehazing
Lexuan Sun, Xueliang Liu, Zhenzhen Hu, Richang Hong
Proceeding • 2021

Video scene detection based on link prediction using graph convolution network
Yingjiao Pei, Zhongyuan Wang, Heling Chen, Baojin Huang, Weiping Tu
Proceeding • 2021

Cross-cultural design of facial expressions for humanoids: is there cultural difference between Japan and Denmark?
Ichi Kanaya, Meina Tawaki, Keiko Yamamoto
Proceeding • 2021

Table detection and cell segmentation in online handwritten documents with graph attention networks
Ying Liu, Heng Zhang, Xiao-Long Yun, Jun-Yu Ye, Cheng-Lin Liu
Proceeding • 2021

RICAPS: residual inception and cascaded capsule network for broadcast sports video classification
Abdullah Aman Khan, Saifullah Tumrani, Chunlin Jiang, Jie Shao
Proceeding • 2021

Transfer non-stationary texture with complex appearance
Cheng Peng, Na Qi, Qing Zhu
Proceeding • 2021

Story segmentation for news broadcast based on primary caption
Heling Chen, Zhongyuan Wang, Yingjiao Pei, Baojin Huang, Weiping Tu

Proceeding • 2021

Intermediate coordinate based pose non-perspective estimation from line correspondences

Yujia Cao, Zhichao Cui, Yuehu Liu, Xiaojun Lv, Kaibei Peng

Proceeding • 2021

An autoregressive generation model for producing instant basketball defensive trajectory

Huan-Hua Chang, Wen-Cheng Chen, Wan-Lun Tsai, Min-Chun Hu, Wei-Ta Chu

Proceeding • 2021

Real-time arbitrary video style transfer

Xingyu Liu, Zongxing Ji, Piao Huang, Tongwei Ren

Proceeding • 2021

C3VQG: category consistent cyclic visual question generation

Shagun Uppal, Anish Madan, Sarthak Bhagat, Yi Yu, Rajiv Ratn Shah

Proceeding • 2021

Cross-modal learning for saliency prediction in mobile environment

Dakai Ren, Xiangming Wen, Xiaoya Liu, Shuai Huang, Jiazhong Chen

Proceeding • 2021

Objective object segmentation visual quality evaluation based on pixel-level and region-level characteristics

Ran Shi, Jian Xiong, Tong Qiao

Proceeding • 2021

Text-based visual question answering with knowledge base

Fang Zhou, Bei Yin, Zanxia Jin, Heran Wu, Dongyan Zhang
Proceeding • 2021

Attention-constraint facial expression recognition
Qisheng Jiang
Proceeding • 2021

EvoGAN: an evolutionary GAN for face aging and rejuvenation
Lianli Gao, Jingqiu Zhang, Jingkuan Song, HengTao Shen
Proceeding • 2021

Destylization of text with decorative elements
Yuting Ma, Fan Tang, Weiming Dong, Changsheng Xu
Proceeding • 2021

Multi-focus noisy image fusion based on gradient regularized convolutional sparse representatione
Xuanjing Shen, Yunqi Zhang, Haipeng Chen, Di Gai
Proceeding • 2021

Determining image age with rank-consistent ordinal classification and object-centered ensemble
Shota Ashida, Adam Jatowt, Antoine Doucet, Masatoshi Yoshikawa
Proceeding • 2021

A treatment engine by multimodal EMR data
Zhaomeng Huang, Liyan Zhang, Xu Xu
Proceeding • 2021

Storyboard relational model for group activity recognition

Boning Li, Xiangbo Shu, Rui Yan
Proceeding • 2021

Distilling knowledge in causal inference for unbiased visual question answering
Yonghua Pan, Zechao Li, Liyan Zhang, Jinhui Tang
Proceeding • 2021

Incremental multi-view object detection from a moving camera
Takashi Konno, Ayako Amma, Asako Kanezaki
Proceeding • 2021

An automated method with anchor-free detection and U-shaped segmentation for nuclei instance segmentation
Xuan Feng, Lijuan Duan, Jie Chen
Proceeding • 2021

Improving face recognition in surveillance video with judicious selection and fusion of representative frames
Zhaozhen Ding, Qingfang Zheng, Chunhua Hou, Guang Shen
Proceeding • 2021

Two-stage structure aware image inpainting based on generative adversarial networks
Jin Wang, Xi Zhang, Chen Wang, Qing Zhu, Baocai Yin
Proceeding • 2021

Low-quality watermarked face inpainting with discriminative residual learning
Zheng He, Xueli Wei, Kangli Zeng, Zhen Han, Qin Zou, Zhongyuan Wang
Proceeding • 2021

A multimedia solution to motivate childhood cancer patients to keep up with cancer treatment

Carmen Wang Er Chai, Bee Theng Lau, Abdullah Al Mahmud, Mark Kit Tsun Tee

Proceeding • 2021

Global and local feature alignment for video object detection

Haihui Ye, Qiang Qi, Ying Wang, Yang Lu, Hanzi Wang

Proceeding • 2021

Defense for adversarial videos by self-adaptive JPEG compression and optical texture

Authors:

Yupeng Cheng • Nanyang Technological University, Singapore

Xingxing Wei • Beihang University, Beijing, China

Huazhu Fu • Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

Shang-Wei Lin • Nanyang Technological University, Singapore

Weisi Lin • Nanyang Technological University, Singapore

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446308

Despite demonstrated outstanding effectiveness in various computer vision tasks, Deep Neural Networks (DNNs) are known to be vulnerable to adversarial examples. Nowadays, adversarial attacks as well as their defenses w.r.t. DNNs in image domain have been intensively studied, and there are some recent works starting to explore adversarial attacks w.r.t. DNNs in video domain. However, the corresponding defense is rarely studied. In this paper, we propose a new two-stage framework for defending video adversarial attack. It contains two main components, namely self-adaptive Joint Photographic Experts Group (JPEG) compression defense and optical texture based defense (OTD). In self-adaptive JPEG compression defense, we propose to adaptively choose an appropriate JPEG quality based on an estimation of moving foreground object, such that the JPEG compression could depress most impact of adversarial noise without losing too much video quality. In OTD, we generate "optical texture" containing high-frequency information based on the optical flow map, and use it to edit Y channel (in YCrCb color space) of input frames, thus further reducing the influence of adversarial perturbation. Experimental results on a benchmark dataset demonstrate the effectiveness of our framework in recovering the classification performance on perturbed videos.

Fusing CAMs-weighted features and temporal information for robust loop closure detection

Authors:

Yaoqing Li • Shenzhen University, Shenzhen, China

Sheng-hua Zhong • Shenzhen University, Shenzhen, China

Tongwei Ren • Nanjing University, Nanjing, China

Yan Liu • The Hong Kong Polytechnic University, Hong Kong, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446309

As a key component in simultaneous localization and mapping (SLAM) system, loop closure detection (LCD) eliminates the accumulated errors by recognizing previously visited places. In recent years, deep learning methods have been proved effective in LCD. However, most of the existing methods do not make good use of the useful information provided by monocular images, which tends to limit their performance in challenging dynamic scenarios with partial occlusion by moving objects. To this end, we propose a novel workflow, which is able to combine multiple information provided by images. We first introduce semantic information into LCD by developing a local-aware Class Activation Maps (CAMs) weighting method for extracting features, which can reduce the adverse effects of moving objects. Compared with previous methods based on semantic segmentation, our method has the advantage of not requiring additional models or other complex operations. In addition, we propose two effective temporal constraint strategies, which utilize the relationship of image sequences to improve the detection performance. Moreover, we propose to use the keypoint matching strategy as the final detector to further refuse false positives. Experiments on four publicly available datasets indicate that our approach can achieve higher accuracy and better robustness than the state-of-the-art methods.

Fixations based personal target objects segmentation

Authors:

Ran Shi • Nanjing University of Science and Technology, Nanjing, China

Gongyang Li • Shanghai University, Shanghai, China

Weijie Wei • Shanghai University, Shanghai, China

Zhi Liu • Shanghai University, Shanghai, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446310

With the development of the eye-tracking technique, the fixation becomes an emergent interactive mode in many human-computer interaction study field. For a personal target objects segmentation task, although the fixation can be taken as a novel and more convenient interactive input, it induces a heavy ambiguity problem of the input's indication so that the segmentation quality is severely degraded. In this paper, to address this challenge, we develop an "extraction-to-fusion" strategy based iterative lightweight neural network, whose input is composed by an original image, a fixation map and a position map. Our neural network consists of two main parts: The first extraction part is a concise interlaced structure of standard convolution layers and progressively higher dilated convolution layers to better extract and integrate local and global features of target objects. The second fusion part is a convolutional long short-term memory component to refine the extracted features and store them. Depending on the iteration framework, current extracted features are refined by fusing them with stored features extracted in the previous iterations, which is a feature transmission mechanism in our neural network. Then, current improved segmentation result is generated to further adjust the fixation map and the position map in the next iteration. Thus, the ambiguity problem induced by the fixations can be alleviated. Experiments demonstrate better segmentation performance of our method and effectiveness of each part in our model.

Improving auto-encoder novelty detection using channel attention and entropy minimization

Authors:

Miao Tian • Zhejiang University of Technology, Hangzhou, China

Dongyan Guo • Zhejiang University of Technology, Hangzhou, China

Ying Cui • Zhejiang University of Technology, Hangzhou, China

Xiang Pan • Zhejiang University of Technology, Hangzhou, China

Shengyong Chen • Tianjin University of Technology, Tianjin, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446311

Novelty detection is a important research area which mainly solves the classification problem of inliers which usually consists of normal samples and outliers composed of abnormal samples. Auto-encoder is often used for novelty detection. However, the generalization ability of the auto-encoder may cause the undesirable reconstruction of abnormal elements and reduce the identification ability of the model. To solve the problem, we focus on the perspective of better reconstructing the normal samples as well as retaining the unique information of normal samples to improve the performance of auto-encoder for novelty detection. Firstly, we introduce attention mechanism into the task. Under the action of attention mechanism, auto-encoder can pay more attention to the representation of inlier samples through adversarial training. Secondly, we apply the information entropy into the latent layer to make it sparse and constrain the expression of diversity. Experimental results on three public datasets show that the proposed method achieves comparable performance compared with previous popular approaches.

Relationship graph learning network for visual relationship detection

Authors:

Yanan Li • Hangzhou Dianzi University, Hangzhou, China

Jun Yu • Hangzhou Dianzi University, Hangzhou, China

Yibing Zhan • Hangzhou Dianzi University, Hangzhou, China

Zhi Chen • Hangzhou Dianzi University, Hangzhou, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446312

Visual relationship detection aims to predict the relationships between detected object pairs. It is well believed that the correlations between image components (i.e., objects and relationships between objects) are significant considerations when predicting objects' relationships. However, most current visual relationship detection methods only exploited the correlations among objects, and the correlations among objects' relationships remained underexplored. This paper proposes a relationship graph learning network (RGLN) to explore the correlations among objects' relationships for visual relationship detection. Specifically, RGLN obtains image objects using an object detector, and then, every pair of objects constitutes a relationship proposal. All relationship proposals construct a relationship graph, in which the proposals are treated as nodes. Accordingly, RGLN designs bi-stream graph attention subnetworks to detect relationship proposals, in which one graph attention subnetwork analyzes correlations among relationships based on visual and spatial information, and the other analyzes correlations based on semantic and spatial information. Besides, RGLN exploits a relationship selection subnetwork to ignore redundant information of object pairs with no relationships. We conduct extensive experiments on two public datasets: the VRD and the VG datasets. The experimental results compared with the state-of-the-art demonstrate the competitiveness of RGLN.

Multiplicative angular margin loss for text-based person search

Authors:

Peng Zhang • University of Electronic Science and Technology of China, Chengdu, China

Deqiang Ouyang • University of Electronic Science and Technology of China, Chengdu, China

Feiyu Chen • University of Electronic Science and Technology of China, Chengdu, China and Sichuan Artificial Intelligence Research Institute, Yibin, China

Jie Shao • University of Electronic Science and Technology of China, Chengdu, China and Sichuan Artificial Intelligence Research Institute, Yibin, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446314

Text-based person search aims at retrieving the most relevant pedestrian images from database in response to a query in form of natural language description. Existing algorithms mainly focus on embedding textual and visual features into a common semantic space so that the similarity score of features from different modalities can be computed directly. Softmax loss is widely adopted to classify textual and visual features into a correct category in the joint embedding space. However, softmax loss can only help classify features but not increase the intra-class compactness and inter-class discrepancy. To this end, we propose multiplicative angular margin (MAM) loss to learn angularly discriminative features for each identity. The multiplicative angular margin loss penalizes the angle between feature vector and its corresponding classifier vector to learn more discriminative feature. Moreover, to focus more on informative image-text pair, we propose pairwise similarity weighting (PSW) loss to assign higher weight to informative pairs. Extensive experimental evaluations have been conducted on the CUHK-PEDES dataset over our proposed losses. The results show the superiority of our proposed method. Code is available at https://github.com/pengzhanguestc/MAM_loss.

Integrating aspect-aware interactive attention and emotional position-aware for multi-aspect sentiment analysis

Authors:

Xiaoye Wang • Tianjin University of Technology Key Laboratory of Computer Vision and System, Tianjin, P.R, China

Xiaowen Zhou • Tianjin University of Technology Key Laboratory of Computer Vision and System, Tianjin, P.R China

Zan Gao • Qilu University of Technology (Shandong Academy of Sciences), Jinan, P.R China

Peng Yang • Tianjin University of Technology Key Laboratory of Computer Vision and System, Tainjin, P.R China

Xianbin Wen • Tianjin University of Technology Key Laboratory of Computer Vision and System, Tainjin, P.R China

Hongyun Ning • Tianjin University of Technology Key Laboratory of Computer Vision and System, Tainjin, P.R China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446315

Aspect-level Sentiment Analysis is a fine-grained sentiment analysis task, which aims to infer the corresponding sentiment polarity with different aspects in an opinion sentence. Attention-based neural networks have proven to be effective in extracting aspect terms, but the prior models are based on context-dependent. Moreover, the prior works only attend aspect terms to detect the sentiment word and cannot consider the sentiment words that might be influenced by domain-specific knowledge. In this work, we proposed a novel integrating Aspect-aware Interactive Attention and Emotional Position-aware module for multi-aspect sentiment analysis (abbreviated to AIAEP) where the aspect-aware interactive attention is utilized to extract aspect terms, and it fuses the domain-specific information of an aspect and context and learns their relationship representations by global context and local context attention mechanisms. Specifically, in the sentiment lexicon, the syntactic parse is used to increase the prior domain knowledge. Then we propose a novel position-aware fusion scheme to compose aspect-sentiment pairs. It combines absolute distance and

relative distance from aspect terms and sentiment words, which can improve the accuracy of polarity classification. Extensive experimental results on SemEval2014 task4 restaurant and AIChallenge2018 datasets demonstrate that AIAEP can outperform state-of-the-art approaches, and it is very effective for aspect-level sentiment analysis.

Graph-based motion prediction for abnormal action detection

Authors:

Yao Tang • Nanjing University of Science and Technology

Lin Zhao • Nanjing University of Science and Technology

Zhaoliang Yao • Nanjing University of Science and Technology

Chen Gong • Nanjing University of Science and Technology

Jian Yang • Nanjing University of Science and Technology

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446316

Abnormal action detection is the most noteworthy part of anomaly detection, which tries to identify unusual human behaviors in videos. Previous methods typically utilize future frame prediction to detect frames deviating from the normal scenario. While this strategy enjoys success in the accuracy of anomaly detection, critical information such as the cause and location of the abnormality is unable to be acquired. This paper proposes human motion prediction for abnormal action detection. We employ sequence of human poses to represent human motion, and detect irregular behavior by comparing the predicted pose with the actual pose detected in the frame. Hence the proposed method is able to explain why the action is regarded as irregularity and locate where the anomaly happens. Moreover, pose sequence is robust to noise, complex background and small targets in videos. Since posture information is non-Euclidean data, graph convolutional network is adopted for future pose prediction, which not only leads to greater expressive power but also stronger generalization capability.

Experiments are conducted both on the widely used anomaly detection dataset ShanghaiTech and our newly proposed dataset NJUST-Anomaly, which mainly contains irregular behaviors happened in the campus. Our dataset expands the existing datasets by giving more abnormal actions attracting public attention in social security, which happen in more complex scenes and dynamic backgrounds. Experimental results on both datasets demonstrate the superiority of our method over the-state-of-the-art methods.

The source code and NJUST-Anomaly dataset will be made public at https://github.com/datangzhengqing/MP-GCN.

Attention feature matching for weakly-supervised video relocalization

Authors:

Haoyu Tang • Xi'an Jiaotong University, Xian, China

Jihua Zhu • Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

Zan Gao • Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

Tao Zhuo • National University of Singapore, Singapore

Zhiyong Cheng • Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446317

Localizing the desired video clip for a given query in an untrimmed video has been a hot research topic for multimedia understanding. Recently, a new task named video relocalization, in which the query is a video clip, has been raised. Some methods have been developed for this task, however, these methods often require dense annotations of the temporal boundaries inside long videos for training. A more practical solution is the weakly-supervised approach, which only needs the matching information between the query and video.

Motivated by that, we propose a weakly-supervised video relocalization approach based on an attention-based feature matching method. Specifically, it recognizes the video clip by finding the clip whose frames are the most relevant to the query clip frames based on the matching results of the frame embeddings. In addition, an attention module is introduced to identify the frames containing rich semantic correlations in the query video. Extensive experiments on the ActivityNet dataset demonstrate that our method can outperform several weakly-supervised methods consistently and even achieve competing performance to supervised baselines.

Pulse localization networks with infrared camera

Authors:

Bohong Yang • Fudan University, P.R.China

Kai Meng • Fudan University, P.R.China

Hong Lu • Fudan University, P.R.China

Xinyao Nie • Fudan University, P.R.China

Guanhao Huang • Fudan University, P.R.China

Jingjing Luo • Fudan University, P.R.China

Xing Zhu • Jihua Laboratory, P.R.China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446318

Pulse localization is the basic task of the pulse diagnosis with robot. More accurate location can reduce the misdiagnosis caused by different types of pulse. Traditional works usually use a collection surface with a certain area for contact detection, and move the collection surface to collect changes of power for pulse localization. These methods often require the subjects place their wrist in a given position. In this paper, we propose a novel pulse localization method which uses the infrared camera as the input sensor, and locates the pulse on wrist with the neural network. This method can not only reduce the contact between the machine and the subject, reduce the discomfort of the process, but also reduce the preparation time for the test, which can improve the detection efficiency. The experiments show that our proposed method can locate the pulse with high accuracy. And we have applied this method to pulse diagnosis robot for pulse data collection.

Structure-preserving extremely low light image enhancement with fractional order differential mask guidance

Authors:

Yijun Liu • University of Electronic Science and Technology of China

Zhengning Wang • University of Electronic Science and Technology of China

Ruixu Geng • University of Electronic Science and Technology of China

Hao Zeng • University of Electronic Science and Technology of China

Yi Zeng • University of Electronic Science and Technology of China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446319

Low visibility and high-level noise are two challenges for low-light image enhancement. In this paper, by introducing fractional order differential, we propose an end-to-end conditional generative adversarial network(GAN) to solve those two problems. For the problem of low visibility, we set up a global discriminator to improve the overall reconstruction quality and restore brightness information. For the high-level noise problem, we introduce fractional order differentiation into both the generator and the discriminator. Compared with conventional end-to-end methods, fractional order can better distinguish noise and high-frequency details, thereby achieving superior noise reduction effects while maintaining details. Finally, experimental results show that the proposed model obtains superior visual effects in low-light image enhancement. By introducing fractional order differential, we anticipate that our framework will enable high quality and detailed image recovery not only in the field of low-light enhancement but also in other fields that require details.

Change detection from SAR images based on deformable residual convolutional neural networks

Authors:

Junjie Wang • Ocean University of China

Feng Gao • Ocean University of China

Junyu Dong • Ocean University of China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446320

Convolutional neural networks (CNN) have made great progress for synthetic aperture radar (SAR) images change detection. However, sampling locations of traditional convolutional kernels are fixed and cannot be changed according to the actual structure of the SAR images. Besides, objects may appear with different sizes in natural scenes, which requires the network to have stronger multi-scale representation ability. In this paper, a novel <u>D</u>eformable <u>R</u>esidual Convolutional Neural <u>N</u>etwork (DRNet) is designed for SAR images change detection. First, the proposed DRNet introduces the deformable convolutional sampling locations, and the shape of convolutional kernel can be adaptively adjusted according to the actual structure of ground objects. To create the deformable sampling locations, 2-D offsets are calculated for each pixel according to the spatial information of the input images. Then the sampling location of pixels can adaptively reflect the spatial structure of the input images. Moreover, we proposed a novel pooling module replacing the vanilla pooling to utilize multi-scale information effectively, by constructing hierarchical residual-like connections within one pooling layer, which improve the multi-scale representation ability at a granular level. Experimental results on three real SAR datasets demonstrate the effectiveness of the proposed DR-Net.

Local structure alignment guided domain adaptation with few source samples

Authors:

Yuying Cai • China University of Petroleum, Huangdao District, Qingdao, China

Jinfeng Li • China University of Petroleum, Huangdao District, Qingdao, China

Baodi Liu • China University of Petroleum, Huangdao District, Qingdao, China

Weifeng Liu • China University of Petroleum, Huangdao District, Qingdao, China

Kai Zhang • China University of Petroleum, Huangdao District, Qingdao, China

Changsheng Xu • Chinese Academy of Sciences, Beijing, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446327

Domain adaptation has received lots of attention for its high efficiency in dealing with cross-domain learning tasks. Most existing domain adaptation methods adopt the strategies relying on large amounts of source label information, which limits their applications in the real world where only a few label samples are available. We exploit the local geometric connections to tackle this problem and propose a Local Structure Alignment (LSA) guided domain adaptation method in this paper. LSA leverages the Nyström method to describe the distribution difference from the geometric perspective and then perform the distribution alignment between domains. Specifically, LSA constructs a domain-invariant Hessian matrix to locally connect the data of the two domains through minimizing the Nyström approximation error. And then it integrates the domain-invariant Hessian matrix with the semi-supervised learning and finally builds an adaptive semi-supervised model. Extensive experimental results validate that the proposed LSA outperforms the traditional domain adaptation methods especially when only sparse source label information is available.

Semantic feature augmentation for fine-grained visual categorization with few-sample training

Authors:

Xiang Guan • University of Electronic Science and Technology of China, Chengdu, China

Yang Yang • University of Electronic Science and Technology of China, Chengdu, China and UESTC, Guangdong, China

Zheng Wang • University of Electronic Science and Technology of China, Chengdu, China

Jingjing Li • University of Electronic Science and Technology of China, Chengdu, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446264

Small data challenges have emerged in many learning problems, since the success of deep neural networks often relies on the availability of a huge number of labeled data that is expensive to collect. We explore a highly challenging task, few-sample training, which uses a small number of labeled images of each category and corresponding textual descriptions to train a model for fine-grained visual categorization. In order to tackle overfitting caused by small data, in this paper, we propose two novel feature augmentation approaches, Semantic Gate Feature Augmentation (SGFA) and Semantic Boundary Feature Augmentation (SBFA). Instead of generating a new image instance, we propose to directly synthesize instance features by leveraging semantic information, and its main novelties are: (1) The SGFA method is proposed to reduce the overfitting of small data by adding random noise to different regions of the image's feature maps through a gating mechanism. (2) The SBFA approach is proposed to optimize the decision boundary of the classifier. Technically, the decision boundary of the image feature is estimated through the assistance of semantic information, and then feature augmentation is performed by sampling in this region. Experiments in fine-grained visual categorization benchmark demonstrate that our proposed approach can significantly improve the categorization performance.

Unsupervised learning of co-occurrences for face images retrieval

Authors:

Thomas Petit • Univ Lyon, Bry-sur-Marne, France

Pierre Letessier • Institut National de l'Audiovisuel, Bry-sur-Marne, France

Stefan Duffner • Univ Lyon, Villeurbanne, France

Christophe Garcia • Univ Lyon, Villeurbanne, France

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446265

Despite a huge leap in performance of face recognition systems in recent years, some cases remain challenging for them while being trivial for humans. This is because a human brain is exploiting much more information than the face appearance to identify a person. In this work, we aim at capturing the social context of unlabeled observed faces in order to improve face retrieval. In particular, we propose a framework that substantially improves face retrieval by exploiting the faces occurring simultaneously in a query's context to infer a multi-dimensional social context descriptor. Combining this compact structural descriptor with the individual visual face features in a common feature vector considerably increases the correct face retrieval rate and allows to disambiguate a large proportion of query results of different persons that are barely distinguishable visually.

To evaluate our framework, we also introduce a new large dataset of faces of French TV personalities organised in TV shows in order to capture the co-occurrence relations between people. On this dataset, our framework is able to improve the mean Average Precision over a set of internal queries from 67.93% (using only facial features extracted with a state-of-the-art pre-trained model) to 78.16% (using both facial features and faces co-occurrences), and from 67.88% to 77.36% over a set of external queries.

Hierarchical clustering via mutual learning for unsupervised person re-identification

Authors:

Xu Xu • Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

Liyan Zhang • Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

Zhaomeng Huang • Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

Guodong Du • Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446268

Person re-identification (re-ID) aims to establish identity correspondence across different cameras. State-of-the-art re-ID approaches are mainly clustering-based Unsupervised Domain Adaptation (UDA) methods, which attempt to transfer the model trained on the source domain to target domain, by alternatively generating pseudo labels by clustering target-domain instances and training the network with generated pseudo labels to perform feature learning. However, these approaches suffer from the problem of inevitable label noise caused by the clustering procedure that dramatically impact the model training and feature learning of the target domain. To address this issue, we propose an unsupervised Hierarchical Clustering via Mutual Learning (HCML) framework, which can jointly optimize the dual training network and the clustering procedure to learn more discriminative features from the target domain. Specifically, the proposed HCML framework can effectively update the hard pseudo labels generated by clustering process and soft pseudo label generated by the training network both in on-line manner. We jointly adopt the repelled loss, triplet loss, soft identity loss and soft triplet loss to optimize the model. The experimental results on Market-to-Duke, Duke-to-Market, Market-to-MSMT and Duke-to-MSMT

unsupervised domain adaptation tasks have demonstrated the superiority of our proposed HCML framework compared with other state-of-the-art methods.

Self-supervised adversarial learning for cross-modal retrieval

Authors:

Yangchao Wang • University of Electronic Science and Technology of China

Shiyuan He • University of Electronic Science and Technology of China

Xing Xu • University of Electronic Science and Technology of China

Yang Yang • University of Electronic Science and Technology of China

Jingjing Li • University of Electronic Science and Technology of China

Heng Tao Shen • University of Electronic Science and Technology of China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446269

Cross-modal retrieval aims at enabling flexible retrieval across different modalities. The core of cross-modal retrieval is to learn projections for different modalities and make instances in the learned common subspace comparable to each other. Self-supervised learning automatically creates a supervision signal by transformation of input data and learns semantic features by training to predict the artificial labels. In this paper, we proposed a novel method named Self-Supervised Adversarial Learning (SSAL) for Cross-Modal Retrieval, which deploys self-supervised learning and adversarial learning to seek an effective common subspace. A feature projector tries to generate modality-invariant representations in the common subspace that can confuse an adversarial discriminator consists of two classifiers. One of the classifiers aims to predict rotation angle from image representations, while the other classifier tries to discriminate between different modalities from the learned embeddings. By confusing the self-supervised adversarial model, feature projector filters out the abundant high-level visual semantics and learns image embeddings that are better aligned with text modality in the common subspace. Through the joint exploitation of the above, an effective common subspace is learned, in which representations of different modlities are aligned better and common information of different modalities is well preserved. Comprehensive experimental results on three widely-used benchmark datasets show that the proposed method is superior in cross-modal retrieval and significantly outperforms the existing cross-modal retrieval methods.

Multi-level expression guided attention network for referring expression comprehension

Authors:

Liang Peng • University of Electronic Science and Tenchnology of China

Yang Yang • University of Electronic Science and Tenchnology of China

Xing Xu • University of Electronic Science and Tenchnology of China

Jingjing Li • University of Electronic Science and Tenchnology of China

Xiaofeng Zhu • University of Electronic Science and Tenchnology of China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446270

Referring expression comprehension is a task of identifying a text-related object or region in a given image by a natural language expression. In this task, it is essential to understand the expression sentence in multi-aspect and adapt it to region representations for generating the discriminative information. Unfortunately, previous approaches usually focus on the important words or phrases in the expression using self-attention mechanisms, which causes that they may fail to distinguish the target region from others, especially the similar regions. To address this problem, we propose a novel model, termed Multi-level Expression Guided Attention network (MEGA-Net). It contains a multi-level visual attention schema guided by the expression representations in different levels, i.e., sentence-level, word-level and phrase-level, which allows generating the discriminative region features and helps to locate the related regions accurately. In addition, to distinguish the similar regions, we design a two-stage structure, where we first select top-K candidate regions according to their matching scores in the first stage, then we apply an object comparison attention mechanism to learn the difference between the candidates for matching the target region. We evaluate the proposed approach on three popular benchmark datasets and the experimental results demonstrate that our model performs against state-of-the-art methods.

Adaptive feature aggregation network for nuclei segmentation

Authors:

Ruizhe Geng • Peking University, Shenzhen, China

Zhongyi Huang • Peking University, Shenzhen, China

Jie Chen • Peng Cheng Laboratory, Shenzhen, China and Peking University, Shenzhen, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446271

Nuclei instance segmentation is essential for cell morphometrics and analysis, playing a crucial role in digital pathology. The problem of variability in nuclei characteristics among diverse cell types makes this task more challenging. Recently, proposal-based segmentation methods with feature pyramid network (FPN) has shown good performance because FPN integrates multi-scale features with strong semantics. However, FPN has information loss of the highest-level feature map and sub-optimal feature fusion strategies. This paper proposes a proposal-based adaptive feature aggregation methods (AANet) to make full use of multi-scale features. Specifically, AANet consists of two components: Context Augmentation Module (CAM) and Feature Adaptive Selection Module (ASM). In feature fusion, CAM focus on exploring extensive contextual information and capturing discriminative semantics to reduce the information loss of feature map at the highest pyramid level. The enhanced features are then sent to ASM to get a combined feature representation adaptively over all feature levels for each RoI. The experiments show our model's effectiveness on two publicly available datasets: the Kaggle 2018 Data Science Bowl dataset and the Multi-Organ nuclei segmentation dataset.

Classification of multimedia SNS posts about tourist sites based on their focus toward predicting eco-friendly users

Authors:

Naoto Kashiwagi • Kyushu University, Fukuoka Japan

Tokinori Suzuki • Kyushu University, Fukuoka Japan

Jounghun Lee • Kyushu University, Fukuoka Japan

Daisuke Ikeda • Kyushu University, Fukuoka Japan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446272

Overtourism has had a negative impact on various things at tourist sites. One of the most serious problems is environmental issues, such as littering, caused by too many visitors to tourist sites. It is important to change people's mindset to be more environmentally aware in order to improve such situation. In particular, if we can find people with comparatively high awareness about environmental issues for overtourism, we will be able to work effectively to promote eco-friendly behavior for people. However, grasping a person's awareness is inherently difficult. For this challenge, we introduce a new task, called Detecting Focus of Posts about Tourism, which is given users' posts of pictures and comment on SNSs about tourist sites, to classify them into types of their focuses based on such awareness. Once we classify such posts, we can see its result showing tendencies of users awareness and so we can discern awareness of the users for environmental issues at tourist sites. Specifically, we define four labels on focus of SNS posts about tourist sites. Based on these labels, we create an evaluation dataset. We present experimental results of the classification task with a CNN classifier for pictures or an LSTM classifier for comments, which will be baselines for the task.

Learning intra-inter semantic aggregation for video object detection

Authors:

Jun Liang • Xiamen University, Xiamen, China

Haosheng Chen • Xiamen University, Xiamen, China

Kaiwen Du • Xiamen University, Xiamen, China

Yan Yan • Xiamen University, Xiamen, China

Hanzi Wang • Xiamen University, Xiamen, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446273

Video object detection is a challenging task due to the appearance deterioration problems in video frames. Thus, object features extracted from different frames of a video are usually deteriorated in varying degrees. Currently, some state-of-the-art methods enhance the deteriorated object features in a reference frame by aggregating the undeteriorated object features extracted from other frames, simply based on their learned appearance relation among object features. In this paper, we propose a novel intra-inter semantic aggregation method (ISA) to learn more effective intra and inter relations for semantically aggregating object features. Specifically, in the proposed ISA, we first introduce an intra semantic aggregation module (Intra-SAM) to enhance the deteriorated spatial features based on the learned intra relation among the features at different positions of an individual object. Then, we present an inter semantic aggregation module (Inter-SAM) to enhance the deteriorated object features in the temporal domain based on the learned inter relation among object features. As a result, by leveraging Intra-SAM and Inter-SAM, the proposed ISA can generate discriminative features from the novel perspective of intra-inter semantic aggregation for robust video object detection. We conduct extensive experiments on the ImageNet VID dataset to evaluate ISA. The proposed ISA obtains 84.5% mAP and 85.2% mAP with ResNet-101 and ResNeXt-101, and it achieves superior performance compared with several state-of-the-art video object detectors.

Robust visual tracking via scale-aware localization and peak response strength

Authors:

Ying Wang • Xiamen University, Xiamen, China

Luo Xiong • Xiamen University, Xiamen, China

Kaiwen Du • Xiamen University, Xiamen, China

Yan Yan • Xiamen University, Xiamen, China

Hanzi Wang • Xiamen University, Xiamen, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446274

Existing regression-based deep trackers usually localize a target based on a response map, where the highest peak response corresponds to the predicted target location. Nevertheless, when the background distractors appear or the target scale changes frequently, the response map is prone to produce multiple sub-peak responses to interfere with model prediction. In this paper, we propose a robust online tracking method via Scale-Aware localization and Peak Response strength (SAPR), which can learn a discriminative model predictor to estimate a target state accurately. Specifically, to cope with large scale variations, we propose a Scale-Aware Localization (SAL) module to provide multi-scale response maps based on the scale pyramid scheme. Furthermore, to focus on the target response, we propose a simple yet effective Peak Response Strength (PRS) module to fuse the multi-scale response maps and the response maps generated by a correlation filter. According to the response map with the maximum classification score, the model predictor iteratively updates its filter weights for accurate target state estimation. Experimental results on three benchmark datasets, including OTB100, VOT2018 and LaSOT, demonstrate that the proposed SAPR accurately estimates the target state, achieving the favorable performance against several state-of-the-art trackers.

Hungry networks: 3D mesh reconstruction of a dish and a plate from a single dish image for estimating food volume

Authors:

Shu Naritomi • The University of Electro-Communications, Tokyo, Japan

Keiji Yanai

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446275

Dietary calorie management has been an important topic in recent years, and various methods and applications on image-based food calorie estimation have been published in the multimedia community. Most of the existing methods of estimating food calorie amounts use 2D-based image recognition. On the other hand, in this paper, we would like to make inferences based on 3D volume for more accurate estimation. We performed 3D reconstruction of a dish (food and plate) and a plate (without foods), from a single image. We succeeded in restoring the 3D shape with high accuracy while maintaining the consistency between a plate part of an estimated 3D dish and an estimated 3D plate. To achieve this, the following contributions were made in this paper. (1) Proposal of "Hungry Networks," a new network that generates two kinds of 3D volumes from a single image. (2) Introduction of plate consistency loss that matches the shapes of the plate parts of the two reconstructed models. (3) Creating a new dataset of 3D food models that are 3D scanned of actual foods and plates. We also conducted an experiment to infer the volume of only the food region from the difference of the two reconstructed volumes. As a result, it was shown that the introduced new loss function not only matches the 3D shape of the plate, but also contributes to obtaining the volume with higher accuracy. Although there are some existing studies that consider 3D shapes of foods, this is the first study to generate a 3D mesh volume from a single dish image.

Scene graph generation via multi-relation classification and cross-modal attention coordinator

Authors:

Xiaoyi Zhang • University of Electronic Science and Technology of China

Zheng Wang • University of Electronic Science and Technology of China

Xing Xu • University of Electronic Science and Technology of China

Jiwei Wei • University of Electronic Science and Technology of China

Yang Yang • University of Electronic Science and Technology of China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446276

Scene graph generation intends to build graph-based representation from images, where nodes and edges respectively represent objects and relationships between them. However, scene graph generation today is heavily limited by imbalanced class prediction. Specifically, most of existing work achieves satisfying performance on simple and frequent relation classes (e.g. on), yet leaving poor performance with fine-grained and infrequent ones (e.g. walk on, stand on). To tackle this problem, in this paper, we redesign the framework as two branches, representation learning branch and classifier learning branch, for a more balanced scene graph generator. Furthermore, for representation learning branch, we propose Cross-modal Attention Coordinator (CAC) to gather consistent features from multi-modal using dynamic attention. For classifier learning branch, we first transfer relation classes' knowledge from large scale corpus, then we leverage Multi-Relationship classifier via Graph Attention neTworks (MR-GAT) to bridge the gap between frequent relations and infrequent ones. The comprehensive experimental results on VG200, a challenge dataset, indicate the competitiveness and the significant superiority of our proposed approach.

A novel system architecture and an automatic monitoring method for remote production

Authors:

Yasuhiro Mochida • Nippon Telegraph and Telephone Corporation Kanagawa, Japan

Daisuke Shirai • Nippon Telegraph and Telephone Corporation Kanagawa, Japan

Takahiro Yamaguchi • Nippon Telegraph and Telephone Corporation Kanagawa, Japan

Seiki Kuwabara • Nippon Telegraph and Telephone Corporation Kanagawa, Japan

Hideki Nishizawa • Nippon Telegraph and Telephone Corporation Kanagawa, Japan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446277

Remote production is an emerging concept concerning the outside-broadcasting workflow enabled by Internet Protocol (IP)-based production systems, and it is expected to be much more efficient than the conventional workflow. However, long-distance transmission of uncompressed video signals and time synchronization of distributed IP-video devices are challenging. A system architecture for remote production using optical transponders (capable of long-distance and large-capacity optical communication) is proposed. A field experiment confirmed that uncompressed video signals can be transmitted successfully by this architecture. The status monitoring of uncompressed video transmission in remote production is also challenging. To address the challenge, a method for automatically monitoring the status of IP-video devices is also proposed. The monitoring system was implemented by using whitebox transponders, and it was confirmed that the system can automatically register IP-video devices, generate an IP-video flow model, and detect traffic anomalies.

Graph convolution network with node feature optimization using cross attention for few-shot learning

Authors:

Ying Liu • Xi'an University of Posts and Telecommunications, Xi'an, China

Yanbo Lei • Xi'an University of Posts and Telecommunications, Xi'an, China

Sheikh Faisal Rashid • University of Engineering & Technology (UET), Peshawar, Pakistan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446278

Graph convolution network (GCN) is an important method recently developed for few-shot learning. The adjacency matrix in GCN models is constructed based on graph node features to represent the graph node relationships, according to which, the graph network achieves message-passing inference. Therefore, the representation ability of graph node features is an important factor affecting the learning performance of GCN. This paper proposes an improved GCN model with node feature optimization using cross attention, named GCN-NFO. Leveraging on cross attention mechanism to associate the image features of support set and query set, the proposed model extracts more representative and discriminative salient region features as initialization features of graph nodes through information aggregation. Since graph network can represent the relationship between samples, the optimized graph node features transmit information through the graph network, thus implicitly enhances the similarity of intra-class samples and the dissimilarity of inter-class samples, thus enhancing the learning capability of GCN. Intensive experimental results on image classification task using different image datasets prove that GCN-NFO is an effective few-shot learning algorithm which significantly improves the classification accuracy, compared with other existing models.

A multi-scale language embedding network for proposal-free referring expression comprehension

Authors:

Taijin Zhao • University of Electronic Science and Technology of China, Chengdu, China

Hongliang Li • University of Electronic Science and Technology of China, Chengdu, China

Heqian Qiu • University of Electronic Science and Technology of China, Chengdu, China

Qingbo Wu • University of Electronic Science and Technology of China, Chengdu, China

King Ngi Ngan • University of Electronic Science and Technology of China, Chengdu, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446279

Referring expression comprehension (REC) is a task that aims to find the location of an object specified by a language expression. Current solutions for REC can be classified into proposal-based methods and proposal-free methods. Proposal-free methods are popular recently because of its flexibility and lightness. Nevertheless, existing proposal-free works give little consideration to visual context. As REC is a context sensitive task, it is hard for current proposal-free methods to comprehend expressions that describe objects by the relative position with surrounding things. In this paper, we propose a multi-scale language embedding network for REC. Our method adopts the proposal-free structure, which directly feeds fused visual-language features into a detection head to predict the bounding box of the target. In the fusion process, we propose a grid fusion module and a grid-context fusion module to compute the similarity between language features and visual features in different size regions. Meanwhile, we extra add fully interacted vision-language information and position information to strength the feature fusion. This novel fusion strategy can help to utilize context flexibly therefore the network can deal with varied expressions, especially

expressions that describe objects by things around. Our proposed method outperforms the state-of-the-art methods on Refcoco, Refcoco+ and Refcocog datasets.

Similar scene retrieval in soccer videos with weak annotations by multimodal use of bidirectional LSTM

Authors:

Tomoki Haruyama • Hokkaido University, Sapporo, Hokkaido, Japan

Sho Takahashi • Hokkaido University, Sapporo, Hokkaido, Japan

Takahiro Ogawa • Hokkaido University, Sapporo, Hokkaido, Japan

Miki Haseyama • Hokkaido University, Sapporo, Hokkaido, Japan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446280

This paper presents a novel method to retrieve similar scenes in soccer videos with weak annotations via multimodal use of bidirectional long short-term memory (BiLSTM). The significant increase in the number of different types of soccer videos with the development of technology brings valid assets for effective coaching, but it also increases the work of players and training staff. We tackle this problem with a nontraditional combination of pre-trained models for feature extraction and BiLSTMs for feature transformation. By using the pre-trained models, no training data is required for feature extraction. Then effective feature transformation for similarity calculation is performed by applying BiLSTM trained with weak annotations. This transformation allows for highly accurate capture of soccer video context from less annotation work. In this paper, we achieve an accurate retrieval of similar scenes by multimodal use of this BiLSTM-based transformer trainable with less human effort. The effectiveness of our method was verified by comparative experiments with state-of-the-art using actual soccer video dataset.

Patch assembly for real-time instance segmentation

Authors:

Yutao Xu • Tongji University, Shanghai, P. R. China

Hanli Wang • Tongji University, Shanghai, P. R. China

Jian Zhu • Tongji University, Shanghai, P. R. China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446281

The paradigm of sliding window is proven effective for the task of visual instance segmentation in many popular research works. However, it still suffers from the bottleneck of inference time. To accelerate existing instance segmentation approaches which are dense sliding window based, this work introduces a novel approach, called patch assembly, which can be integrated into bounding box detectors for segmentation without extra up-sampling computations. A well-designed detector named PAMask is proposed to verify the effectiveness of the proposed approach. Benefitting from the simple structure as well as a fusion of multiple representations, PAMask has the ability to run in real time while achieving competitive performances. Besides, another effective technique called Center-NMS is designed to reduce the number of boxes for intersection of union calculation, which can be fully parallelized on device and contributes 0.6% mAP improvement both in detection and segmentation for free.

Full-resolution encoder-decoder networks with multi-scale feature fusion for human pose estimation

Authors:

Jie Ou • University of Electronic Science and Technology of China, Chengdu, China

Mingjian Chen • University of Electronic Science and Technology of China, Chengdu, China

Hong Wu • University of Electronic Science and Technology of China, Chengdu, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446282

To achieve more accurate 2D human pose estimation, we extend the successful encoder-decoder network, simple baseline network (SBN), in three ways. To reduce the quantization errors caused by the large output stride size, two more decoder modules are appended to the end of the simple baseline network to get full output resolution. Then, the global context blocks (GCBs) are added to the encoder and decoder modules to enhance them with global context features. Furthermore, we propose a novel spatial-attention-based multi-scale feature collection and distribution module (SA-MFCD) to fuse and distribute multi-scale features to boost the pose estimation. Experimental results on the MS COCO dataset indicate that our network can remarkably improve the accuracy of human pose estimation over SBN, our network using ResNet34 as the backbone network can even achieve the same accuracy as SBN with ResNet152, and our networks can achieve superior results with big backbone networks.

Graph-based variational auto-encoder for generalized zero-shot learning

Authors:

Jiwei Wei • University of Electronic Science and Technology of China, China

Yang Yang • University of Electronic Science and Technology of China, China

Xing Xu • University of Electronic Science and Technology of China, China

Yanli Ji • University of Electronic Science and Technology of China, China

Xiaofeng Zhu • University of Electronic Science and Technology of China, China

Heng Tao Shen • University of Electronic Science and Technology of China, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446283

Zero-shot learning has been a highlighted research topic in both vision and language areas. Recently, generative methods have emerged as a new trend of zero-shot learning, which synthesizes unseen categories samples via generative models. However, the lack of fine-grained information in the synthesized samples makes it difficult to improve classification accuracy. It is also time-consuming and inefficient to synthesize samples and using them to train classifiers. To address such issues, we propose a novel Graph-based Variational Auto-Encoder for zero-shot learning. Specifically, we adopt knowledge graph to model the explicit inter-class relationships, and design a full graph convolution auto-encoder framework to generate the classifier from the distribution of the class-level semantic features on individual nodes. The encoder learns the latent representations of individual nodes, and the decoder generates the classifiers from latent representations of individual nodes. In contrast to synthesize samples, our proposed method directly generates classifiers from the distribution of the class-level semantic features for both seen and unseen categories, which is more straightforward, accurate and computationally efficient. We conduct extensive experiments and evaluate our method on the widely used large-scale ImageNet-21K dataset. Experimental results validate the efficacy of the proposed approach.

A multi-scale human action recognition method based on Laplacian pyramid depth motion images

Authors:

Chang Li • Hohai University, Jiangsu, Nanjing, China

Qian Huang • Hohai University, Jiangsu, Nanjing, China

Xing Li • Hohai University, Jiangsu, Nanjing, China

Qianhan Wu • Hohai University, Jiangsu, Nanjing, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446284

Human action recognition is an active research area in computer vision. Aiming at the lack of spatial muti-scale information for human action recognition, we present a novel framework to recognize human actions from depth video sequences using multi-scale Laplacian pyramid depth motion images (LP-DMI). Each depth frame is projected onto three orthogonal Cartesian planes. Under three views, we generate depth motion images (DMI) and construct Laplacian pyramids as structured multi-scale feature maps which enhances multi-scale dynamic information of motions and reduces redundant static information in human bodies. We further extract the multi-granularity descriptor called LP-DMI-HOG to provide more discriminative features. Finally, we utilize extreme learning machine (ELM) for action classification. Through extensive experiments on the public MSRAction3D datasets, we prove that our method outperforms state-of-the-art benchmarks.

Fixed-size video summarization over streaming data via non-monotone submodular maximization

Authors:

Ganfeng Lu • Nanjing University of Aeronautics & Astronautics, China

Jiping Zheng • Nanjing University of Aeronautics & Astronautics, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446285

Video summarization which potentially fast browses a large amount of emerging video data as well as saves storage cost has attracted tremendous attentions in machine learning and information retrieval. Among existing efforts, determinantal point processes (DPPs) designed for selecting a subset of video frames to represent the whole video have shown great success in video summarization. However, existing methods have shown poor performance to generate fixed-size output summaries for video data, especially when video frames arrive in streaming manner. In this paper, we provide an efficient approach k-seqLS which summarizes streaming video data with a fixed-size k in vein of DPPs. Our k-seqLS approach can fully exploit the sequential nature of video frames by setting a time window and the frames outside the window have no influence on current video frame. Since the log-style of the DPP probability for each subset of frames is a non-monotone submodular function, local search as well as greedy techniques with cardinality constraints are adopted to make k-seqLS fixed-sized, efficient and with theoretical guarantee. Our experiments show that our proposed k-seqLS exhibits higher performance while maintaining practical running time.

Overlap classification mechanism for skeletal bone age assessment

Authors:

Pengyi Hao • Zhejiang University of Technology, Hangzhou, China

Xuhang Xie • East China Normal University, Shanghai, China

Tianxing Han • Zhejiang University of Technology, Hangzhou, China

Cong Bai • Zhejiang University of Technology, Hangzhou, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446286

The bone development is a continuous process, however, discrete labels are usually used to represent bone ages. This inevitably causes a semantic gap between actual situation and label representation scope. In this paper, we present a novel method named as overlap classification network to narrow the semantic gap in bone age assessment. In the proposed network, discrete bone age labels (such as 0-228 month) are considered as a sequence that is used to generate a series of subsequences. Then the proposed network makes use of the overlapping information between adjacent subsequences and output several bone age ranges at the same time for one case. The overlapping part of these age ranges is considered as the final predicted bone age. The proposed method without any preprocessing can achieve a much smaller mean absolute error compared with state-of-the-art methods on a public dataset.

Fixation guided network for salient object detection

Authors:

Zhe Cui • University of Chinese Academy of Sciences

Li Su • University of Chinese Academy of Sciences

Weigang Zhang • Harbin Institute of Technology, Weihai

Qingming Huang • University of Chinese Academy of Sciences

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446288

Convolutional neural network (CNN) based salient object detection (SOD) has achieved great development in recent years. However, in some challenging cases, i.e. small-scale salient object, low contrast salient object and cluttered background, existing salient object detect methods are still not satisfying. In order to accurately detect salient objects, SOD networks need to fix the position of most salient part. Fixation prediction (FP) focuses on the most visual attractive regions, so we think it could assist in locating salient objects. As far as we know, there are few methods jointly consider SOD and FP tasks. In this paper, we propose a fixation guided salient object detection network (FGNet) to leverage the correlation between SOD and FP. FGNet consists of two branches to deal with fixation prediction and salient object detection respectively. Further, an effective feature cooperation module (FCM) is proposed to fuse complementary information between the two branches. Extensive experiments on four popular datasets and comparisons with twelve state-of-the-art methods show that the proposed FGNet well captures the main context of images and locates salient objects more accurately.

Fixation Guided Network for Salient Object Detection

Zhe Cui
cuizhe18@mails.ucas.ac.cn
University of Chinese Academy of Sciences

Li Su
suli@ucas.ac.cn
University of Chinese Academy of Sciences

Weigang Zhang
wgzhang@hit.edu.cn
Harbin Institute of Technology, Weihai

Qingming Huang
qmhuang@ucas.ac.cn
University of Chinese Academy of Sciences
Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences

ABSTRACT

Convolutional neural network (CNN) based salient object detection (SOD) has achieved great development in recent years. However, in some challenging cases, i.e. small-scale salient object, low contrast salient object and cluttered background, existing salient object detect methods are still not satisfying. In order to accurately detect salient objects, SOD networks need to fix the position of most salient part. Fixation prediction (FP) focuses on the most visual attractive regions, so we think it could assist in locating salient objects. As far as we know, there are few methods jointly consider SOD and FP tasks. In this paper, we propose a fixation guided salient object detection network (FGNet) to leverage the correlation between SOD and FP. FGNet consists of two branches to deal with fixation prediction and salient object detection respectively. Further, an effective feature cooperation module (FCM) is proposed to fuse complementary information between the two branches. Extensive experiments on four popular datasets and comparisons with twelve state-of-the-art methods show that the proposed FGNet well captures the main context of images and locates salient objects more accurately.

CCS CONCEPTS

• **Computing methodologies** → **Interest point and salient region detections**.

KEYWORDS

salient object detection, fixation prediction, convolutional neural network, computer vision

ACM Reference Format:
Zhe Cui, Li Su, Weigang Zhang, and Qingming Huang. 2021. Fixation Guided Network for Salient Object Detection. In *ACM Multimedia Asia (MMAsia '20), March 7–9, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3444685.3446288

MMAsia '20, March 7–9, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8308-0/21/03…$15.00
https://doi.org/10.1145/3444685.3446288

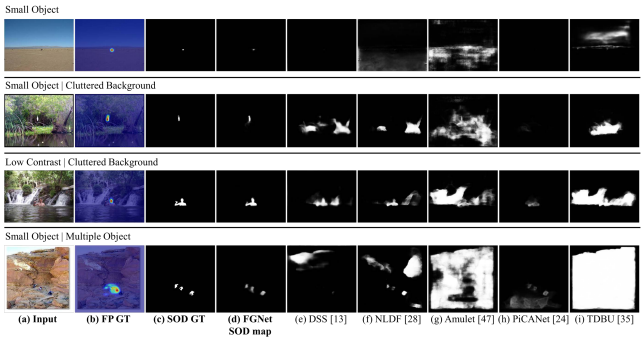


Figure 1: Examples to show some problems in current methods. (a) Input. (b) FP ground-truth. (c) SOD ground-truth. (d) SOD maps generated by the proposed FGNet. (e)-(i) SOD maps generated by DSS, NLDF, Amulet, PiCANet and TDBU respectively.

1 INTRODUCTION

Visual attention mechanism in human visual system means that people can focus the most attractive regions when looking at an image. Through mimicking the visual attention mechanism, salient object detection aims to segment the most visual distinctive objects in an image. As a fundamental task in computer vision, salient object detection is widely applied to many other visual tasks, such as image retrieval [11], semantic segmentation [38, 39], and image captioning [9].

During the past decades, a number of traditional methods [3, 43] were proposed to deal with SOD, which only use low-level cues and hand-crafted features to detect and segment salient objects. With the success of deep learning in computer vision, a number of CNN-based methods have been applied to salient object detection in recent years and have significantly outperformed traditional methods. Early CNN-based methods [19, 20, 31, 50] extract deep features and predict saliency scores for each image regions one by one, which is time-consuming. Currently, the most popular salient object detection methods [5, 10, 13, 22, 24, 28, 33–37, 40, 41, 46–48] are based on the fully convolutional networks (FCN) [27]. These FCN-based salient object detection methods focus on exploring different feature fusion strategies and have achieved satisfactory performance.

Although significant progresses have been achieved, some salient objects can hardly be precisely discovered in challenging cases as shown in Fig. 1. This is because the visual contrast between salient

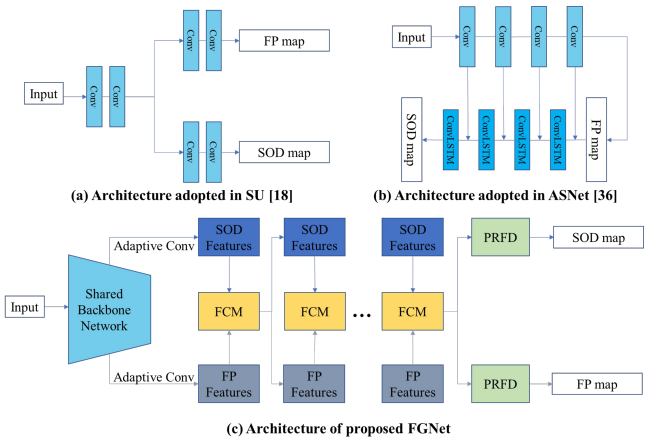


Figure 2: Architecture comparison between the proposed FGNet and previous models that combine SOD and FP. (a) Architecture adopted in SU only share features in low layers (b) Architecture adopted in ASNet use FP to help SOD unidirectional. (c). Architecture of the proposed FGNet can mutually exchange information between SOD and FP with cascade FCMs.

objects and background is not obvious. Under these circumstances, human visual attention is critical to discovering salient objects. In order to detect and segment salient objects accurately, we must to locate attention-grabbing regions firstly. FP aims to predict fixation points where humans look during scene free viewing, while SOD takes a further step to segment the whole extent of salient objects. The corresponding pixels with higher value in both FP and SOD maps are more likely to be attended. In addition, the analyses in [2, 4] indicate the intrinsic correlation between FP and SOD.

However, most saliency researches treat SOD and FP as two individual tasks and only few works [18, 36] try to explore the relationship between them. Kruthiventi et al. [18] implemented both fixation prediction and salient object detection via a two-branch unified network SU which share features only in low layers as shown in Fig. 2 (a). Wang et al. [36] presented an attentive saliency network ASNet to progressively refine saliency map from fixation map through a top-down pathway by aggregating multi-level features as shown in Fig. 2 (b).

Based on the above observation, we focus on leveraging the complementarity between salient object detection and fixation prediction information. We propose a fixation guided salient object detection network (FGNet) to couple salient object detection and fixation prediction in a unified network as shown in Fig. 2 (c). Compared with SU [18] and ASNet [36], the proposed FGNet utilize advantages of both fixation prediction and salient object detection by building a bidirectional information flow architecture. Instead of only using fixation features to guide salient object detection, we optimize both features by mutually flowing information between SOD and FP.

In summary, the contributions are as follows:

- We propose FGNet to explicitly combine complementary information between salient object detection and fixation prediction. Compared with exiting methods only using FP as an auxiliary task to unidirectionally improve the representation ability of SOD features, we propose cascade feature cooperation modules to exchange information between SOD and FP bidirectional.
- The proposed FGNet promotes SOD and FP at the same time by allowing them to mutually pass information to each other, yielding more accurate SOD maps.
- Experimental results on four popular datasets show that the proposed FGNet substantially improves the SOD performance compared with 12 state-of-the-art methods.

2 RELATED WORK

FP is an active research topic in computer vision area for a long time. Itti et al. [16] proposed a bottom-up attention models based on psychological theories and started the field of visual attention. Traditional fixation prediction models [45] were commonly built on bottom-up structure using stimulus-driven biologically features and certain heuristics. Recent deep learning based fixation prediction models [15, 23] leverage CNN to extract representative features for fixation prediction by aggregating features from multi-stream, multi-scale and multi-level, etc.

Compared with fixation prediction, the history of SOD is short and the original works of SOD can trace back to the work of Liu et al. [26] and Achanta et al. [1]. Traditional salient object detection methods rely on hand-crafted features to predict saliency scores, such as center prior [17], contrast prior [6, 30], and spectral information [14]. These approaches can usually extract hand-crafted features with little time cost. However, the hand-crafted features are low-level and hardly capture high-level semantic knowledge of the salient objects, which will decrease detection accuracy.

With the popular of the CNN which has powerful feature extraction capability, SOD began to leverage CNN to extract features and got great progress. Zhao et al. [50] presented a multi-context deep learning framework to extract local and global context simultaneously, which were then fed into CNN for saliency classification. Wang et al. [31] used local estimation and global contrast information to produce saliency maps. These methods fed each processing unit into classifiers for saliency score prediction and have obvious disadvantages: time-consuming, unable to use overall spatial information, and all pixels in each processing unit share the same saliency score.

Inspired by the great success of FCN in semantic segmentation, salient object detection area has turned attention to FCN. FCN can extract multi-level features, where high-level features capture semantic knowledge and low-level features contain more detailed information. Researchers focus on designing different fusion strategies to aggregate multi-level features. Wang et al. [33] used salient priors to make the training of network easier, utilized cascaded FCNs to refine saliency map iteratively by correcting its previous errors, until the final prediction was generated in the last time step. Hou et al. [13] proposed a new method to fuse the low-level features and the high-level features by adding several short connections from deeper side-outputs to shallower ones based on the Holistically-Nested Edge Detector [42]. Chen et al. [5] proposed a reverse attention network to compensate the missing parts between the prediction and the ground-truth by erasing current generated

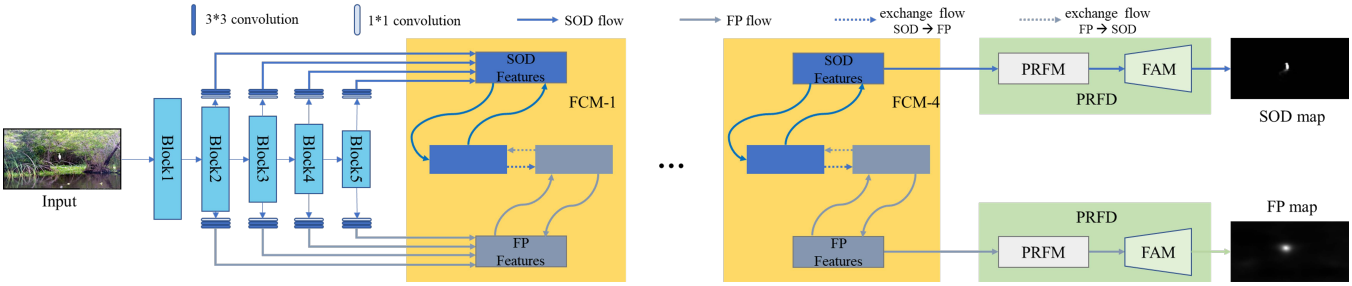


Figure 3: The overall architecture of FGNet. There are two branches (SOD branch and FP branch) based on the shared backbone network. FGNet consists of four cascade Feature Cooperation Modules (FCM) and two Pyramid Receptive Field Decoders (PRFD). FCM is proposed to exchange information between two branches. PRFD is used to generate SOD and FP maps by fusing refined features.

saliency maps. Most of aggregation-based methods fuse all features extracted from FCN to generate final saliency map. Wu et al. [40] discovered that fusing features in shallow layers brings little improvement in the final saliency map, but increases computation cost greatly.

3 FGNET

In this section, we first introduce the overall architecture of the proposed network. Then, we show details about how to fuse the features between two tasks and how to integrate the refined features to generate SOD and FP maps.

3.1 Overall Architecture

We choose ResNet-50 [12] as a common feature extractor and modify it to meet the SOD requirements. ResNet-50 consists of 49 convolutional layers with five convolutional blocks, following a global pooling layer and a fully connected layer. We only use five residual blocks to extract multi-scale features and denote them as $R = \{R_i | i = 1, 2, 3, 4, 5\}$. The block size is $\frac{W}{2^i} \times \frac{H}{2^i} \times C_i$, where H, W are the width and height of the input image, and C_i is the channel number of the i-th feature R_i. Since shallow layer contribute less to the final results but largely increase the computation cost as demonstrated in [40], so only $\{R_i | i = 2, 3, 4, 5\}$ are reserved. In addition, we add three convolutional layers on the top of each block before the branch of fixation prediction and salient object detection separately. On the one hand, it is for the sake of making the features extracted from the shared backbone network more adaptive to the two tasks. On the other hand, it is for changing the channel to 32 in order to reduce computation cost. Now we get two feature groups $S = \{S_i | i = 2, 3, 4, 5\}$ and $F = \{F_i | i = 2, 3, 4, 5\}$ for SOD and FP respectively.

FGNet is a two-branch architecture network: salient object detection branch and fixation prediction branch, as shown in Fig. 3. Salient object detection branch focuses on extracting fine-grained visual features to precisely segment salient objects. Fixation prediction branch aims to capture the global image information to locate salient objects. Since both SOD and FP are related to human attention, features from these two branches have inherent relevance and complementarity. We further propose a feature cooperation module (FCM) to facilitate information sharing and exchanging between the two branches. After the features have learned enough, we feed them into the pyramid receptive field decoder (PRFD) to generate SOD and FP maps.

3.2 Feature Cooperation Module

In order to make full use of the complementarity between SOD and FP features, the feature cooperation module (FCM) is proposed. FCM consists of two sub-modules: mutual feature exchange module (MFEM) and internal feature fusion module (IFFM) as shown in Fig. 4. MFEM builds a bidirectional information flow mechanism aiming to exchange information between features of two branches in each level. Considering that low-level features contain more background distractors, we only choose the equivalent and higher level features when aggregating complementary features to suppress distractors. In addition, we observe that there are some fixation points fall outside salient objects occasionally, so the FP features need to multiplied by corresponding SOD features before aggregated into SOD features. IFFM offers a top-down feature fusion mechanism within a single branch. IFFM further explicitly increases the weight of high-level information in each level features, to make sure each layer can capture enough location information by introducing high-level guiding flows between adjacent layers.

The detailed illustration of FCM is shown in Fig. 4. FCM takes both SOD features S and FP features F as input. After the process of MFEM and IFFM, FCM generates refined SOD features $S^{'}$ and FP features $F^{'}$. By stacking multiple FCMs (i.e., the output of one FCM is used as the input to the next FCM), both SOD and FP can learn more complementary information from each other. To achieve the best trade-off between performance and model size, we stack four FCMs.

3.3 Pyramid Receptive Field Decoder

By stacking multiple FCMs, complementary features in two branches have been conducted a full exchange of information. Another question deserves considering is how to fuse these multi-level refined features.

As demonstrated in [49], the receptive fields of CNN layers are much smaller compared with its theoretical value especially when CNNs go deeper. In order to solve this problem and allow each pixel in feature maps could capture different receptive fields of feature maps, we propose a pyramid receptive field module (PRFM) inspired by [25]. As shown in Fig. 5, PRFM comprises four parallel branches containing convolutional layers of different kernel size

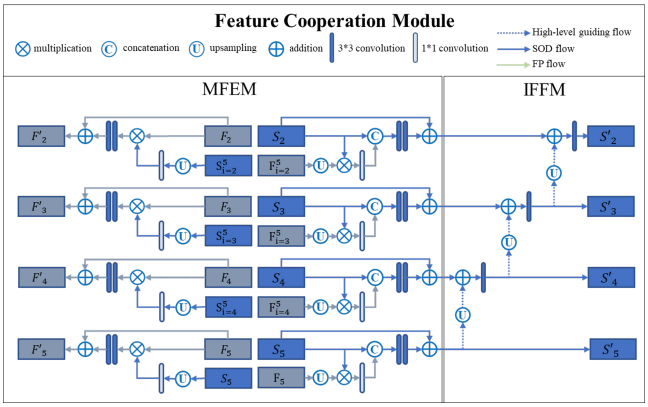


Figure 4: Detailed illustration of FCM. It comprises mutual feature exchange module (MFEM) and internal feature fusion module (IFFM). MFEM is developed to fuse complementary features from SOD and FP. IFFM is designed to further enhance the high-level location information.

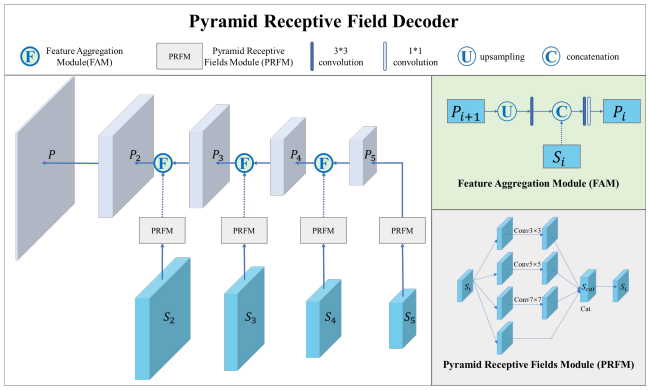


Figure 5: Detailed illustration of PRFD. It comprises pyramid receptive field module (PRFM) and feature aggregation module (FAM). PRFM is used to capture different receptive fields of feature maps. FAM is designed to seamlessly aggregate multi-scale features.

$\{1, 3, 5, 7\}$ to capture pyramid receptive fields for each pixel in feature maps. Then we utilize feature aggregation module (FAM) to progressively aggregate high-level features with low-level features in a top-down pathway. At each time step, the fused feature map P_i can be generated by combining S_i and feature maps $P_{(i+1)}$ from previous step using FAM. The feature maps P_2 in the last step has $[\frac{W}{4}, \frac{H}{4}]$ size and 32 channels, so additional convolutional layers and upsample layers are added to generate final SOD map P. The generation process of FP map Q is similar to that of SOD map.

After get SOD map P and FP map Q, the total train loss of FGNet L_{total} could be calculated by adding salient object detection loss $L_{CE}(P, GT_s)$ and fixation prediction loss $L_{CE}(Q, GT_f)$ as (1):

$$L_{total} = L_{CE}(P, GT_s) + L_{CE}(Q, GT_f) \tag{1}$$

where $GT = \{GT_s, GT_f\}$ are the ground-truth map for SOD and FP, $\theta = \{\theta_s, \theta_f\}$ are the parameters corresponding to maps $\{P, Q\}$, and L_{CE} is the standard pixel-wise cross entropy loss formulated as (2):

$$\begin{aligned} L_{CE}(P, GT_s|\theta_s) &= -\sum_{i=1}^{N} GT_s^i log(P^i) + (1 - GT_s^i) log(1 - P^i) \\ L_{CE}(Q, GT_f|\theta_f) &= -\sum_{i=1}^{N} GT_f^i log(Q^i) + (1 - GT_f^i) log(1 - Q^i) \end{aligned} \tag{2}$$

where N is the pixel number.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metric

To train and evaluate the proposed FGNet, five popular benchmark datasets are adopted, including ECSSD [43], HKU-IS [20], PASCAL-S [21], DUTS [32], DUT-OMRON [44]. **ECSSD** contains 1000 semantically meaningful and structure complex images. **HKU-IS** contains 4447 images with high quality annotations, which have multiple disconnected salient objects or objects touching image boundary. **PASCAL-S** contains 850 images selected from PASCAL VOC 2010. **DUTS** contains 10553 images for training and 5019 images for testing, and it is the largest SOD benchmark dataset. **DUT-OMRON** contains 5168 high-quality images which have annotations for both salient object detection and fixation prediction.

Table 1: Ablation analysis on DUTS-TEST dataset. Best scores in each row are highlighted in bold. FCM-4 is the adopted architecture

variant	maxF	meanF	MAE	S
single SOD branch	0.863	0.765	0.052	0.861
FCM-1	0.885	0.795	0.043	0.879
FCM-2	0.889	0.801	0.041	0.887
FCM-4	0.892	**0.815**	**0.038**	**0.891**
FCM-6	**0.895**	0.812	0.039	0.889
w/o PRFM	0.884	0.799	0.041	0.883

Considering that we focus on the task of SOD, so we just give visual results for FP to verify the effectiveness of the generated FP maps. And for SOD task, we adopt four widely used evaluation metrics including mean absolute error (MAE), max F-measure (max F), mean F-measure (mean F) and precision-recall curve in order to compare with other methods. Further, we also evaluate the proposed model based on the S-measure [7], E-measure [8] and weighted F-measure [29] to give more comprehensive evaluation.

F-measure is a balanced mean of average precision and average recall which can be calculated by (3), where β^2 is set to 0.3 to weigh precision more than recall as suggested in [1].

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{3}$$

4.2 Implementation Details

Firstly, we adopt DUT-OMRON dataset to pre-train FGNet because this dataset could provide both SOD and FP ground-truths. Then, FGNet is trained on DUTS-TRAIN dataset following most salient object detection works [24, 34, 40, 41, 48]. Since DUTS-TRAIN only contains SOD ground-truth, so the total train loss of equals to salient object detection loss $L_{total} = L_{CE}(P, GT_s)$ in this stage.

Table 2: Comparison with state-of-the-art methods. Max F-measure (maxF, larger is better), Mean F-measure (meanF, larger is better), MAE (smaller is better), S measure (larger is better) are used to measure the model performance. '-' denotes that the authors have not provided corresponding saliency maps. * means methods combine FP and SOD. The top three results are marked in red, blue, and green, respectively. FGNet achieves the state-of-the-art under all evaluation metrics on four popular datasets.

method	ECSSD				HKU-IS				PASCAL-S				DUTS-TEST			
	maxF	meanF	MAE	S	maxF	meanF	MAE	S	maxF	meanF	MAE	S	maxF	meanF	MAE	S
MDF	0.832	0.807	0.105	0.776	0.860	0.784	0.129	0.810	0.764	0.705	0.145	0.688	-	-	-	-
RFCN	0.890	0.834	0.107	0.852	0.893	0.835	0.089	0.859	0.829	0.747	0.132	0.794	0.784	0.711	0.090	0.791
DSS	0.908	0.865	0.062	0.883	0.898	0.854	0.051	0.879	0.824	0.763	0.103	0.797	0.813	0.712	0.065	0.826
NLDF	0.905	0.874	0.063	0.875	0.900	0.872	0.049	0.876	0.826	0.770	0.099	0.798	0.813	0.738	0.065	0.816
Amulet	0.915	0.870	0.059	0.894	0.894	0.838	0.053	0.882	0.832	0.764	0.097	0.815	0.779	0.672	0.085	0.803
BMPM	0.928	0.894	0.044	0.911	0.920	0.875	0.039	0.906	0.857	0.803	0.073	0.840	0.852	0.762	0.049	0.861
PAGR	0.927	0.894	0.061	0.889	0.918	0.886	0.048	0.887	0.851	0.803	0.092	0.813	0.854	0.783	0.056	0.838
DGRL	0.925	0.903	0.043	0.906	0.913	0.882	0.037	0.897	0.853	0.807	0.074	0.834	0.828	0.794	0.050	0.842
PiCANet-R	0.935	0.886	0.046	0.917	0.918	0.870	0.043	0.904	0.863	0.798	0.075	0.849	0.860	0.759	0.051	0.869
PiCANet	0.931	0.885	0.046	0.914	0.921	0.870	0.042	0.906	0.862	0.796	0.076	0.845	0.851	0.749	0.054	0.861
PAGE	0.931	0.906	0.042	0.912	0.918	0.882	0.037	0.903	0.852	0.810	0.077	0.835	0.838	0.777	0.052	0.854
TDBU	0.938	0.88	0.041	0.918	0.922	0.878	0.038	0.907	0.859	0.779	0.071	0.844	0.855	0.767	0.048	0.865
ASNet*	0.932	0.875	0.047	0.915	0.922	0.872	0.041	0.906	0.871	0.791	0.069	0.856	0.835	0.728	0.061	0.843
FGNet*	0.948	0.916	0.037	0.926	0.936	0.898	0.033	0.917	0.879	0.833	0.064	0.861	0.892	0.815	0.038	0.891

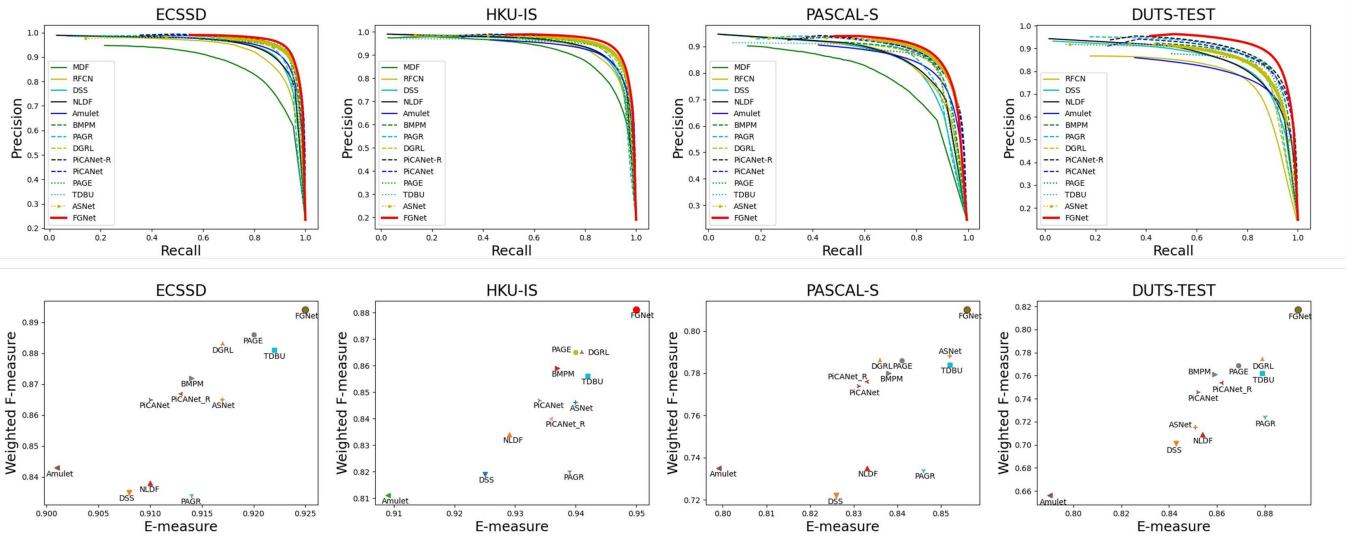


Figure 6: Performance comparison with state-of-the-art methods on four popular benchmark datasets. The first row shows precision-recall curves. The second row shows weighted F-measure and E-measure. It can be seen that the proposed method performs favorably against state-of-the-arts.

Remaining datasets DUTS-TEST, ECSSD, PASCAL-S, HKU-IS are used to evaluate the proposed model.

The parameters of the backbone are initialized by ResNet-50 pretrained on ImageNet. For all newly added convolution layers, their weights are initialized by normal distribution with standard deviation = 0.01 and mean value = 0. The proposed model is implemented in PyTorch. The whole network is trained by stochastic gradient descent (SGD). Momentum and weight decay are set to 0.9 and 5e-4. Epoch and batch size are set to 30 and 8 for both pretrain and train stages. In the pretrain stage, learning rate is initialized as 2e-3 and decreased by 10% at 20 epochs, and in the train stage, learning rate is initialized as 2e-4 and decreased by 10% at 20 epochs.

4.3 Ablation Study

In this subsection, We conduct a series of ablation experiments to confirm the effectiveness of our proposed FGNet. As shown in Tab. 1, we analyzing the contribution of each part on DUTS-TEST. Firstly, we evaluate the performance of the proposed model with only SOD branch to show the advantage of fusing two branches. Secondly, we explore how the number of FCMs influence the final detection results. Finally, we analyse how about the performance would be when replacing the PRFD with simple fusion strategies. It demonstrates that all modules can help locate and segment salient objects.

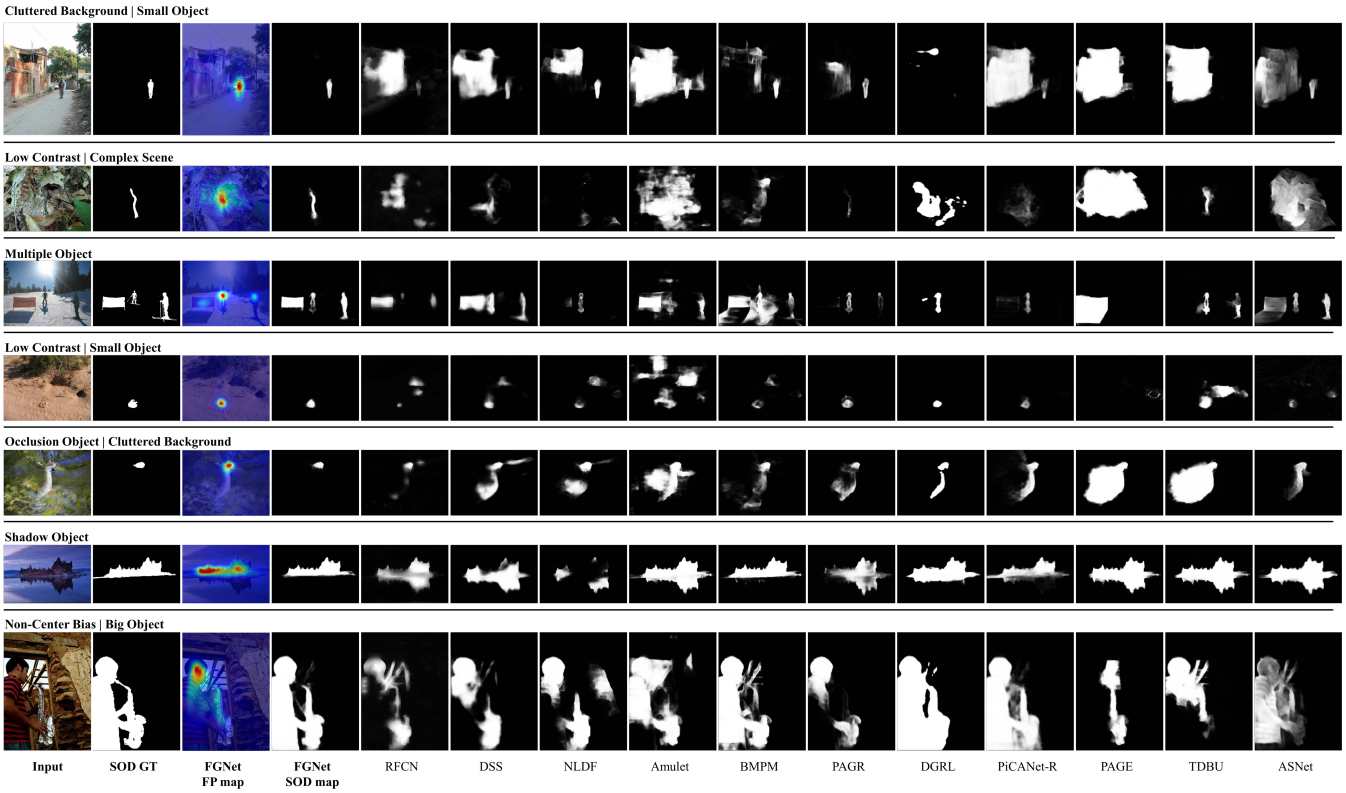


Figure 7: Visual comparison of FGNet results compared with other state-of-the-art methods. Each row represents one image and we highlight the main challenges. Each row represents the results of one method, where ASNet and the proposed FGNet combine FP and SOD. (GT: ground-truth)

4.4 Comparison with the State-of-the-arts

Quantitative Comparison: We compare the proposed method with 12 state-of-the-art salient object detection methods including MDF [20], RFCN [33], DSS [13], NLDF [28], Amulet [47], BMPM [46], PAGR [48], DGRL [34], PiCANet [24], PAGE [37], TDBU [35], ASNet [36]. Among these methods, ASNet and the proposed FGNet are methods that combine FP and SOD. All the saliency maps are provided by authors or generated by running their pre-trained models with parameter recommended in their papers. For fair comparison, we evaluate all the saliency maps with the same evaluation codes. Tab. 2 shows the max F-measure, mean F-measure, MAE, and S-measure. Fig. 6 shows the precision-recall curves in the first row, weighted F-measure and E-measure in the second row. FGNet achieves state-of-the-art performance on four datasets without any post-processing techniques.

Visual Comparison: As showed in Fig. 7, we can see that FGNet performs better when dealing with various challenging cases. For the first, second and fourth samples, it is difficult to detect salient objects only based on low-level visual features due to cluttered background, small salient objects and low contrast between foreground and background. For the fifth and sixth samples, other methods mistakenly detect hidden parts underwater and shadows as salient objects. Because they and the salient objects can be seen as a whole in terms of the low-level visual characteristics. However, benefiting from the complementary fixation information which mimics human visual mechanisms, FGNet can detect the most important object(s) accurately. It is worth mentioning that thanks to the cooperation of FP and SOD features, FGNet could not only locate salient region but also segment the whole salient object clearly.

5 CONCLUSION

In this paper, we propose a fixation guided salient object detection network FGNet to leverage the correlation between SOD and FP. Different from other methods which just take fixation map as an auxiliary information, the complementarity between SOD and FP is fully considered. Firstly, the SOD and FP features are extracted by the two-branch network. Then the FCMs are proposed to fuse the complementary information by building a mutual information exchange mechanism. Finally, the PRFDs are employed to integrate these refined features and generate SOD and FP maps. Extensive experiments demonstrate that FGNet outperforms most of the state-of-the-art approaches on four popular datasets.

ACKNOWLEDGMENTS

This work was supported in part by the National Key RD Program of China under Grant 2018AAA0102003 and 2018YFE0118400, in part by National Natural Science Foundation of China: 61931008, 61976069, and 61771457, and in part by the Fundamental Research Funds for Central Universities.

REFERENCES

[1] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. 2009. Frequency-tuned salient region detection. In *CVPR*. 1597–1604.

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2015. Salient Object Detection: A Benchmark. *IEEE Transactions on Image Processing* 24, 12 (2015), 5706–5722.

[3] Ali Borji and Laurent Itti. 2012. Exploiting local and global patch rarities for saliency detection. In *CVPR*. 478–485.

[4] Ali Borji, Dicky N.Sihite, and Laurent Itti. 2013. What stands out in a scene? A study of human explicit saliency judgment. *Vision Research* 91, 15 (2013), 62–77.

[5] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. 2018. Reverse Attention for Salient Object Detection. In *ECCV (9) (Lecture Notes in Computer Science)*, Vol. 11213. 236–252.

[6] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. 2011. Global contrast based salient region detection. In *CVPR*. 409–416.

[7] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *ICCV*. 4558–4567.

[8] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*. 698–704.

[9] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *CVPR*. 1473–1482.

[10] Mengyang Feng, Huchuan Lu, and Errui Ding. 2019. Attentive Feedback Network for Boundary-Aware Salient Object Detection. In *CVPR*. 1623–1632.

[11] Junfeng He, Jinyuan Feng, Xianglong Liu, Tao Cheng, Tai-Hsu Lin, Hyunjin Chung, and Shih-Fu Chang. 2012. Mobile product search with Bag of Hash Bits and boundary reranking. In *CVPR*. 3005–3012.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, , and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. 2019. Deeply Supervised Salient Object Detection with Short Connections. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 4 (2019), 815–828.

[14] Xiaodi Hou and Liqing Zhang. 2007. Saliency detection: A spectral residual approach. In *CVPR*. 1–8.

[15] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *CVPR*. 262–270.

[16] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 11 (1998), 1254–1259.

[17] Zhuolin Jiang and Larry S. Davis. 2013. Submodular salient region detection. In *CVPR*. 2043–2050.

[18] Srinivas S. S. Kruthiventi, Vennela Gudisa, Jaley H. Dholakiya, and R. Venkatesh Babu. 2016. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *CVPR*. 5781–5790.

[19] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. 2016. Deep Saliency with Encoded Low level Distance Map and High Level Features. In *CVPR*. 660–668.

[20] Guanbin Li and Yizhou Yu. 2015. Visual saliency based on multiscale deep features. In *CVPR*. 5455–5463.

[21] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. 2014. The Secrets of Salient Object Segmentation. In *CVPR*. 280–287.

[22] Jiangjiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. 2019. A Simple Pooling-Based Design for Real-Time Salient Object Detection. In *CVPR*. 3917–3926.

[23] Nian Liu, Junwei Han, Tianming Liu, and Xuelong Li. 2018. Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 29, 2 (2018), 392–404.

[24] Nian Liu, Junwei Han, and Ming-Hsuan Yang. 2018. PiCANet: Learning PixelWise Contextual Attention for Saliency Detection. In *CVPR*. 3089–3098.

[25] Songtao Liu, Di Huang, and Yunhong Wang. 2018. Receptive field block net for accurate and fast object detection. In *ECCV*.

[26] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. 2007. Learning to detect a salient object. In *CVPR*. 1–8.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*. 3431–3440.

[28] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. 2017. Non-Local Deep Features for Salient Object Detection. In *CVPR*. 6593–6601.

[29] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. 2014. How to evaluate foreground maps. In *CVPR*. 248–255.

[30] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. 2012. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*. 733–740.

[31] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2015. Deep Networks for Saliency Detection via Local Estimation and Global Search. In *CVPR*. 3183–3192.

[32] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. 2017. Learning to Detect Salient Objects with Image-Level Supervision. In *CVPR*. 3796–3805.

[33] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. 2016. Saliency Detection with Recurrent Fully Convolutional Networks. In *ECCV (4) (Lecture Notes in Computer Science)*, Vol. 9908. 825–841.

[34] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. 2018. Detect Globally, Refine Locally: A Novel Approach to Saliency Detection. In *CVPR*. 3127–3135.

[35] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. 2019. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*. 5968–5977.

[36] Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji, and Ruigang Yang. 2020. Inferring salient objects from human fixations. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 8 (2020), 1913–1927.

[37] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven C. H. Hoi, and Ali Borji. 2019. Salient Object Detection With Pyramid Attention and Salient Edges. In *CVPR*. 1448–1457.

[38] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. 2018. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*. 1354–1362.

[39] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*. 6488–6496.

[40] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *CVPR*. 3902–3911.

[41] Zhe Wu, Li Su, and Qingming Huang. 2019. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*. 7263–7272.

[42] Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *ICCV*. 1395–1403.

[43] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. 2013. Hierarchical Saliency Detection. In *CVPR*. 1155–1162.

[44] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2013. Saliency Detection via Graph-Based Manifold Ranking. In *CVPR*. 3166–3173.

[45] Jianming Zhang and Stan Sclaroff. 2013. Saliency Detection: A Boolean Map Approach. In *ICCV*. 153–160.

[46] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. 2018. A Bi-Directional Message Passing Model for Salient Object Detection. In *CVPR*. 1741–1750.

[47] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. 2017. Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection. In *ICCV*. 202–211.

[48] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. 2018. Progressive Attention Guided Recurrent Network for Salient Object Detection. In *CVPR*. 714–722.

[49] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *CVPR*. 6230–6239.

[50] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2015. Saliency detection by multi-context deep learning. In *CVPR*. 1265–1274.

Motion-transformer: self-supervised pre-training for skeleton-based action recognition

Authors:

Yi-Bin Cheng • Sun Yat-sen University, Guangzhou, Guangdong

Xipeng Chen • Sun Yat-sen University, Guangzhou, Guangdong

Dongyu Zhang • Sun Yat-sen University, Guangzhou, Guangdong

Liang Lin • Sun Yat-sen University, Guangzhou, Guangdong

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446289

With the development of deep learning, skeleton-based action recognition has achieved great progress in recent years. However, most of the current works focus on extracting more informative spatial representations of the human body, but haven't made full use of the temporal dependencies already contained in the sequence of human action. To this end, we propose a novel transformer-based model called Motion-Transformer to sufficiently capture the temporal dependencies via self-supervised pre-training on the sequence of human action. Besides, we propose to predict the motion flow of human skeletons for better learning the temporal dependencies in sequence. The pre-trained model is then fine-tuned on the task of action recognition. Experimental results on the large scale NTU RGB+D dataset shows our model is effective in modeling temporal relation, and the flow prediction pre-training is beneficial to expose the inherent dependencies in time dimensional. With this pre-training and fine-tuning paradigm, our final model outperforms previous state-of-the-art methods.

Interactive re-ranking for cross-modal retrieval based on object-wise question answering

Authors:

Rintaro Yanagi • Hokkaido University, Sapporo, Hokkaido, Japan

Ren Togo • Hokkaido University, Sapporo, Hokkaido, Japan

Takahiro Ogawa • Hokkaido University, Sapporo, Hokkaido, Japan

Miki Haseyama • Hokkaido University, Sapporo, Hokkaido, Japan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446290

Cross-modal retrieval methods retrieve desired images from a query text by learning relationships between texts and images. This retrieval approach is one of the most effective ways in the easiness of query preparation. Recent cross-modal retrieval is convenient and accurate when users input a query text that can uniquely identify the desired image. Meanwhile, users frequently input ambiguous query texts, and these ambiguous queries make it difficult to obtain the desired images. To alleviate these difficulties, in this paper, we propose a novel interactive cross-modal retrieval method based on question answering (QA) with users. The proposed method analyses candidate images and asks users about information that can narrow retrieval candidates effectively. By only answering the questions generated by the proposed method, users can reach their desired images even from an ambiguous query text. Experimental results show the effectiveness of the proposed method.

A background-induced generative network with multi-level discriminator for text-to-image generation

Authors:

Ping Wang • Shandong Normal University, Shandong, China

Li Liu • Shandong Normal University, Shandong, China

Huaxiang Zhang • Shandong Normal University, Shandong, China

Tianshi Wang • Shandong Normal University, Shandong, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446291

Most existing text-to-image generation methods focus on synthesizing images using only text descriptions, but this cannot meet the requirement of generating desired objects with given backgrounds. In this paper, we propose a Background-induced Generative Network (BGNet) that combines attention mechanisms, background synthesis, and multi-level discriminator to generate realistic images with given backgrounds according to text descriptions. BGNet takes a multi-stage generation as the basic framework to generate fine-grained images and introduces a hybrid attention mechanism to capture the local semantic correlation between texts and images. To adjust the impact of the given backgrounds on the synthesized images, synthesis blocks are added at each stage of image generation, which appropriately combines the foreground objects generated by the text descriptions with the given background images. Besides, a multi-level discriminator and its corresponding loss function are proposed to optimize the synthesized images. The experimental results on the CUB bird dataset demonstrate the superiority of our method and its ability to generate realistic images with given backgrounds.

WFN-PSC: weighted-fusion network with poly-scale convolution for image dehazing

Authors:

Lexuan Sun • Hefei University of Technology (HFUT)

Xueliang Liu • Hefei University of Technology (HFUT)

Zhenzhen Hu • Hefei University of Technology (HFUT)

Richang Hong • Hefei University of Technology (HFUT)

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446292

Image dehazing is a fundamental task for the computer vision and multimedia and usually in the face of the challenge from two aspects, i) the uneven distribution of arbitrary haze and ii) the distortion of image pixels caused by the hazed image. In this paper, we propose an end-to-end trainable framework, named Weighted-Fusion Network with Poly-Scale Convolution (WFN-PSC), to address these dehazing issues. The proposed method is designed based on the Poly-Scale Convolution (PSConv). It can extract the image feature from different scales without upsampling and downsampled, which avoids the image distortion. Beyond this, we design the spatial and channel weighted-fusion modules to make the WFN-PSC model focus on the hard dehazing parts of image from two dimensions. Specifically, we design three Part Architectures followed by the channel weighted-fusion module. Each Part Architecture consists of three PSConv residual blocks and a spatial weighted-fusion module. The experiments on the benchmark demonstrate the dehazing effectiveness of the proposed method. Furthermore, considering that image dehazing is a low-level task in the computer vision, we evaluate the dehazed image on the object detection task and the results show that the proposed method can be a good pre-processing to assist the high-level computer vision task.

Video scene detection based on link prediction using graph convolution network

Authors:

Yingjiao Pei • Wuhan University, Wuhan, China

Zhongyuan Wang • Wuhan University, Wuhan, China

Heling Chen • Wuhan University, Wuhan, China

Baojin Huang • Wuhan University, Wuhan, China

Weiping Tu • Wuhan University, Wuhan, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446293

With the development of the Internet, multimedia data grows by an exponential level. The demand for video organization, summarization and retrieval has been increasing where scene detection plays an essential role. Existing shot clustering algorithms for scene detection usually treat temporal shot sequence as unconstrained data. The graph based scene detection methods can locate the scene boundaries by taking the temporal relation among shots into account, while most of them only rely on low-level features to determine whether the connected shot pairs are similar or not. The optimized algorithms considering temporal sequence of shots or combining multi-modal features will bring parameter trouble and computational burden. In this paper, we propose a novel temporal clustering method based on graph convolution network and the link transitivity of shot nodes, without involving complicated steps and prior parameter setting such as the number of clusters. In particular, the graph convolution network is used to predict the link possibility of node pairs that are close in temporal sequence. The shots are then clustered into scene segments by merging all possible links. Experimental results on BBC and OVSD datasets show that our approach is more robust and effective than the comparison methods in terms of F1-score.

Cross-cultural design of facial expressions for humanoids: is there cultural difference between Japan and Denmark?

Authors:

Ichi Kanaya • Nagasaki University, Nagasaki, Japan

Meina Tawaki • The University of Nagasaki, Nagayo, Nagasaki, Japan

Keiko Yamamoto • Kyoto Institute of Technology, Kyoto, Japan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446294

In this research, the authors succeeded in creating facial expressions made with the minimum necessary elements for recognizing a face. The elements are two eyes and a mouth made using precise circles, which are transformed to make facial expressions geometrically, through rotation and vertically scaling transformation. The facial expression patterns made by the geometric elements and transformations were composed employing three dimensions of visual information that had been suggested by many previous researches, slantedness of the mouth, openness of the face, and slantedness of the eyes. The authors found that this minimal facial expressions can be classified into 10 emotions: happy, angry, sad, disgust, fear, surprised, angry*, fear*, neutral (pleasant) indicating positive emotion, and neutral (unpleasant) indicating negative emotion. The authors also investigate and report cultural differences of impressions of facial expressions of above-mentioned simplified face.

Table detection and cell segmentation in online handwritten documents with graph attention networks

Authors:

Ying Liu • University of Chinese Academy of Sciences, Beijing, China and Institute of Automation Chinese Academy of Sciences, Beijing, China

Heng Zhang • Institute of Automation Chinese Academy of Sciences, Beijing, China

Xiao-Long Yun • University of Chinese Academy of Sciences, Beijing, China and Institute of Automation Chinese Academy of Sciences, Beijing, China

Jun-Yu Ye • Institute of Automation Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China

Cheng-Lin Liu • Institute of Automation Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446295

In this paper, we propose a multi-task learning approach for table detection and cell segmentation with densely connected graph attention networks in free form online documents. Each online document is regarded as a graph, where nodes represent strokes and edges represent the relationships between strokes. Then we propose a graph attention network model to classify nodes and edges simultaneously. According to node classification results, tables can be detected in each document. By combining node and edge classification resutls, cells in each table can be segmented. To improve information flow in the network and enable efficient reuse of features among layers, dense connectivity among layers is used. Our proposed model has been experimentally validated on an online handwritten document dataset IAMOnDo and achieved encouraging results.

Table Detection and Cell Segmentation in Online Handwritten Documents with Graph Attention Networks

Ying Liu[2,1]
liuying2019@ia.ac.cn

Heng Zhang[1]
heng.zhang@ia.ac.cn

Xiao-Long Yun[2,1]
xiaolong.yun@nlpr.ia.ac.cn

Jun-Yu Ye[1,2]
junyu.ye@nlpr.ia.ac.cn

Cheng-Lin Liu[1,2]
liucl@nlpr.ia.ac.cn

[1] National Laboratory of Pattern Recognitio, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China
[2] School of Artificial Intelligece, University of Chinese Academy of Sciences, Beijing 100049, China

ABSTRACT

In this paper, we propose a multi-task learning approach for table detection and cell segmentation with densely connected graph attention networks in free form online documents. Each online document is regarded as a graph, where nodes represent strokes and edges represent the relationships between strokes. Then we propose a graph attention network model to classify nodes and edges simultaneously. According to node classification results, tables can be detected in each document. By combining node and edge classification resutls, cells in each table can be segmented. To improve information flow in the network and enable efficient reuse of features among layers, dense connectivity among layers is used. Our proposed model has been experimentally validated on an online handwritten document dataset IAMOnDo and achieved encouraging results.

KEYWORDS

table detection, cell segmentation, online handwritten document, graph neural networks

ACM Reference Format:
Ying Liu[2,1], Heng Zhang[1], Xiao-Long Yun[2,1], Jun-Yu Ye[1,2], and Cheng-Lin Liu[1,2]. 2021. Table Detection and Cell Segmentation in Online Handwritten Documents with Graph Attention Networks. In *ACM Multimedia Asia (MMAsia '20), March 7–9, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3444685.3446295

1 INTRODUCTION

A document normally contains textblocks, lists, formulas, graphs, tables and so on. Specially, most ink documents contain various types of tables. As a key element in documents, tables can express more information in fewer words, and enable readers to search, compare and understand the content rapidly. Thus, table analysis with detection and cell segmentation is a essential work in ink document analysis.

MMAsia '20, March 7–9, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8308-0/21/03…$15.00
https://doi.org/10.1145/3444685.3446295

Table detection task aims to extract tables from free-form documents and cells are segmanted in each table for futher analysis. In previous works, researchers tend to use heuristic rules to detect tables or divide documents into several parts for table extraction. Few works focus on cell setmention after table detection in ink documents. Jain et al. [2] proposed an approach which used the Hough transform to detect table lines(at least five lines). Because this approach focuses on table lines, it can not extract tables either without lines nor with less than five lines. Zhang et al. [3] modeled each ink document as a matrix and then extract tables on the matrix. In recent years, with the popular use of deep learning, CNNs (Convolutional Neural Networks) are proved effective in image processing. By converting ink documents into offline images, a lot of deep learning based methods can transform the table detection problem as semantic segmentation or object detection tasks [4,5]. But when online documents are transformed into offline images, a lot of temporal information is lost. The lost information is critical to extract components from ink documents.

In this paper, we propose a multi-task learning approach to solve the table detection and cell segmentation problems at the same time. We formulate the ink document as a graph, the nodes are strokes and the edges are the temporal and spatial relationships between strokes. Thus, table detection and cell segmentation problems can be regarded as node and edge classification in the graph. Unary geometric features extracted from strokes and binary geometric features extracted from pairs of strokes are used as node and edge features, which are fed into the graph attention network model for node and edge classification in the graph with a multi-task learning mode. The strokes are classified into three class i.e. table line, table text and non-table. The stroke pairs are classified into same cell, different cell or other (at least one stroke is non-table or table-line). According to the stroke classification results and some post-processing steps, tables can be extracted from the document graph. In the detected table, cells can be segmented according to the stroke pair classification results.

The main contributions of this work are summarized as follows:

1. Our table detection approach doesn't rely on table lines, which means tables, with complete, incomplete, or without bounding and separating lines, can be processed with a uniform framework.

2. Our model uses the idea of multi-task learning to do node and edge classification at the same time, which means that our method can segment cells while detecting the table region.

3. We use strict evaluation indicators: only when all the table and non-table strokes in the document are recognized correctly, the table detection is considered correct. Only when strokes in each segmented cell exactly match with the ground truth, the resulting cell is regarded as a correct segmentation example. That means our table detection and cell segmentation results are more refined, and we have achieved encouraging results on the IAMOnDo [21] dataset.

The rest of the paper is structured as follows. We first provide the overview of related works in Section II. Then we describe the proposed approach with all the deatils in Section III. Experimental settings and results are presented in Section IV and conclusion in Section V.

2 RELATED WORKS

2.1 Table Analysis

The task of table analysis has been studied for many years. A large number of methods on table analysis have been proposed. In earlier years, researchers tried to solve these problems by heuristic rule-based methods. Jain et al. [2] presented an approach by detecting table lines. They restricted that each table must have five or more than five lines, and four boundaries lines together with a separating line must be contained. Then, they used Hough transform to detect lines, and cell segmentation was determined by the horizontal and vertical lines. This approach cannot detect tables without lines or identify cell boundaries in a non-regular table. To overcome the restricting of table lines, Zhang et al. [3] proposed a method to transform the document into a matrix which was constructed by grouping strokes that are proximate or collinear. Then, by judging the horizontal and vertical distances of the matrix rows and columns, grouped strokes were classified into table rows or columns. Similarly, Shilman et al.[7] combined strokes with similar sizes and orientations to form words, lines, and blocks (paragraphs) in a bottom-up order. Blanchard et al. [6] extracted paragraphs, lines, and words based on an extension of a probabilistic approach. These approaches deal with the written information of strokes indirectly, and so complicated and redundant with large amount of calculations. Delaye et al.[1] applied tree conditional random fields to classify strokes in documents, and then clustered strokes into sub-blocks. This method grouped strokes into serval classes directly using geometric features, but the method of segmenting document with spatially cluster lack enough generalization ability and the computation amount of conditional random fields is large.

As the development of deep learning, especially the excellent performance of CNNs in image processing, a lot of researchers treated table detection as semantic segmentation or object detection. He et al. [9] proposed a multi-scale, multi-task fully convolutional neural network (FCN) for this task and used conditional random field (CRF) to make the network's output more accurate. Kavasidis et al. [10] proposed a saliency-based convolutional neural network for table and chart detection. In [11], Siddiqui et al. proposed the deep deformable CNN for table detection.

2.2 Graph Neural Network

The concept of graph neural network was first proposed by Gori et al. [12] and further clarified by Scarselli et al. [13]. These early studies learned the representation of the target node in an iterative manner by propagating neighboring information through a recurrent neural architecture until a stable fixed point was reached. Motivated by the great success of convolutional networks in the field of computer vision, Graph Convolutional Networks (GCN) [14] have recently emerged to redefine the concept of convolution for graph data. Bruna et al. [8] proposed the first important research on graph convolutional networks. They developed a variant of graph convolution based on the spectral graph theory. The attention mechanism [15] has almost become a standard configuration in sequence tasks to focus on the most important parts of objects. This mechanism has been proven to be useful in multiple tasks, such as machine translation and natural language understanding. Visi et al. [16] proposed graph attention networks, which can benefit a lot from the attention mechanism when it aggregates information and integrates the information from different degree nodes. Gong et al. [17] proposed that EGNN (Edge graph neural networks) makes full use of edge features. At each layer, a new formula is constructed to handle multi-dimensional edge features. DenseNet [18] was proposed to exploit dense connectivity among layers, and improved information flow in the network and enabled efficient reuse of features among layers. Inspired by DenseNet, Li et al. [19] adapted a similar idea to GCNs so as to exploit information flow from different GCN layers.

3 PROPOSED METHOD

3.1 Framework

Figure 1 shows the pipeline of our network. Given an ink handwritten document, we first construct a graph from strokes temporal/spatial relationship and use the unary/binary geometric features in [20] for the nodes and edges. Then, the graph, node and edge features are fed into three shared layers, the output of which are fed into two branches: node and edge classification. To overcome the problem of overfitting in graph neural network training, we use dense connectivity among layers. At the end, a softmax unit is used to predict node and edge labels.

3.2 Graph Construction

Our proposed method for table detection treat the ink handwritten document as a graph G(V,E), where node $i \in V$ represents a stroke and edge $(i, j) \in E$ represents temporal or spatial relationship of the pair of strokes (i, j). In particular, edges in the graph consisted of two categories.

Temporal edges For online ink handwritten documents, temporal relationship which captures the interaction between strokes is an important feature. To exploit temporal features, we roughly define that if any two strokes (i, j) are temporally adjacent, there is an edge between strokes (i, j).

Spatial edges In document structure and semantic analysis, contextual features of space also play a vital role. Thus, spatial edges are added to the graph by calculating the minimum distance between strokes. If the minimum distance of two strokes (i, j) is less than the threshold, the pair of strokes (i, j) is connected by the edge (i, j) in the graph. The threshold is chosen experientially. The overlapping edges among different types of construction will be removed so there are no repeated edges in the relational graph.

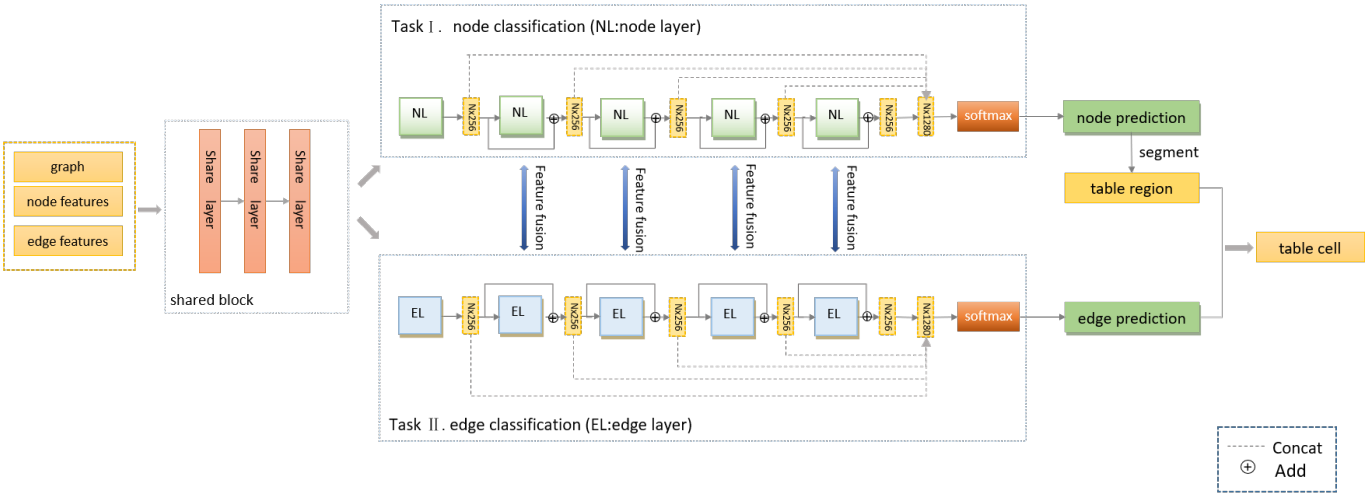


Figure 1: The framework of our multi-task graph attention network are densely connected. A handwritten document is first converted into a graph based on the temporal and spatial relationship of strokes. Graph together with node and edge features are fed into the network, passing through a shared block, then fed into two graph attention layers for node and edge classification. The nodes are classified into three classes, table line, table text and non-table. According to the node classification results and post processing, the document will be separated into table and non-table region. Then, strokes (sorted by writing order) in table region combine with edge prediction results, cells in table can be segmented.

Furthermore, we also add a self-loop edge of each node to the graph because the self feature of the node has been proved effectively to aggregate in the message passing procedure of graph neural networks [15].

3.3 Multi-task learning graph attention network with densely connection

In this work, we propose the multi-task learning graph attention network with densely connection. It is consisted of a shared layer block and two branches: node and edge classification.

To train the node and edge classification simultaneously, we use a basic multi-task learning framework where node and edge classification branches share a few layers as shown in Figure 1 and the loss of node and edge is added together:

$$Loss_{\text{sum}} = Loss_{\text{node}} + Loss_{\text{edge}} \tag{1}$$

Shared layer block we present a shared layer block which is consisted of three shared layers, and each layer is a graph attention layer. In each shared layer, only node features are updated. The output of each layer a set of node features H = $\{h_i|i \in V\}, h_i \in R^C$ and the original set of edge features F = $\{f_{i,j}|$ (i, j) $\in$ E $\}$,$f_{i,j} \in R^D$. Where V is the set of nodes, E is the set of edges , C and D are the numbers of features in each node and edge, respectively.

Node classification branch we present the node layer (NL) stacked to construct the node classification branch. The input is a set of node features H = $\{h_i|i \in V\}, h_i \in R^C$ and edge features F = $\{f_{i,j}|$ (i, j) $\in$ E $\}$,$f_{i,j} \in R^D$. The layer output is a new set of node features H′ = $\{h'_i|i \in V\}, h'_i \in R^{C'}$, where C' is the number of output features.

Two attention mechanisms are applied in the unit NL. The first is the self attention mechanism. A shared linear transformation is applied to each node, then perform self-attention on the nodes by a shared attentional mechanism $a : R^{C'} \times R^{C'} \rightarrow$ R to compute the node attention score:

$$s_{ij} = a(\mathbf{W}_h h_i, \mathbf{W}_h h_j) \tag{2}$$

where $\mathbf{W}_h \in R^{C' \times C}$ is a learnable parameter. The attention mechanism used in this work is the additive attention which is defined as :

$$a(\mathbf{W}_h h_i, \mathbf{W}_h h_j) = \delta(\mathbf{V}^T(\mathbf{W}_h h_i || \mathbf{W}_h h_j)) \tag{3}$$

where $\mathbf{V}^T$ is a learnable parameter and $\delta(\cdot)$ is a leaky ReLU activation function.

The second attention mechanism is used to aggerate the pairwise stroke features passing through edges in the graph. It is defined as:

$$s'_{ij} = \delta(v_f \delta(\mathbf{W}_f f_{ij} + b_f)) \tag{4}$$

where $\mathbf{W}_f \in R^{C' \times D}, b_f \in R^{C'}, v_f \in R^{C'}$ are learnable parameters.

Two kinds of attention coefficients are added together and normalized by the softmax function:

$$\alpha_{ij} = \frac{exp(s_{ij} + s'_{ij})}{\sum_{k \in N_i} exp(s_{ik} + s'_{ik})} \tag{5}$$

After obtaining the attention coefficients, the output node features are computed as weighted combination of its neighborhood node features with the attention scores:

$$h'_i = \delta(\sum_{j \in N_i} \alpha_{ij} \mathbf{W}_h h_j) \tag{6}$$

Then the standard multi-head attention technique is also employed:

$$h'_i = ||_{k=1}^{K} \delta(\sum_{j \in N_i} \alpha_{ij}^k \mathbf{W}_h^k h_j) \tag{7}$$

where || represents concatenation.

In the last layer, the concatenation operation is replaced by averaging operation and then the softmax function is applied:

$$h_i' = \delta(\frac{1}{k}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k \mathbf{W}_h^j h_j) \tag{8}$$

$$p_i = \text{softmax}(\mathbf{W}_o h_i') \tag{9}$$

where $\mathbf{W}_o \in R^{C'\times L}$ is the learnable weight matrix to transform features to outputs.

Edge classification branch To segment cells in the table, we present the edge layer (EL), which is stacked to construct the edge classification branch. After updating the node features, the edge features updated by combining the updated node features and themself features. The input of each edge layer is a set of node features H = $\{h_i|i \in V\}, h_i \in R^C$, and a set of edge features F = $\{f_{i,j}|$ (i, j) $\in E'\}$. The outputs of each edge layer is a new set of edge features: F′ = $\{f'_{i,j}|(i,j) \in E'\}$.

First, a shared linear transformation is applied to each edge feature :

$$e_{ij} = \delta(\mathbf{W}_{\text{edge}} f_{ij}) \tag{10}$$

where $\mathbf{W}_{\text{edge}}$ is a learnable parameter.

The classification of the two nodes connected by a edge influences the edge prediction, thus, node features are essential in the edge features updating process. So we add the node features into the edge update process by formulas below:

$$h_{ij} = \delta(\mathbf{W}_{\text{node}}(h_i||h_j)) \tag{11}$$

$$f'_{ij} = \delta(\mathbf{W}_{\text{concat}}(e_{ij}||h_{ij})) \tag{12}$$

where $\mathbf{W}_{\text{node}}, \mathbf{W}_{\text{concat}}$ are learnable parameters, || represents concatenation operation,δ is the leaky ReLU activation function.

To better use the node and edge features, the output features of NL and EL are fused during training procedure. As DenseNet [13] is proposed to exploit dense connectivity among layers, and improves information flow in the network and enables efficient reuse of features among layers, we adapt a similar idea for node and edge classification to exploit information flow from different layers. Finally, we also employ some standard tricks to stabilize and accelerate the training procedure, including residual connection and batch normalization. In this work, all the hyper-parameters of hidden layers are the same, and the residual connection is added except the input layer.

Table detection and cell segmentation According to stroke classification results, the document is naturally divided into two parts: table and non-table part. However, the table part may contain multiple tables, or includes non-table strokes that are predicted incorrectly. We combine the node(stroke) classification results, the edge(stroke pair) classification results, the time and spatial information contained in the strokes to segment the table part. Algorithm 1 summarizes the inference process.

After Algorithm 1, we can get several independent table blocks. Then the edge classification branch can classify stroke pairs into same cell, different cell or other (edges between non-table or non-table and table stroke pair). According to the edge classification results, we can group strokes into cells as this rule: If a pair of strokes is adjacent in time and the edge class between them is same cell, then they are grouped into the same cell, or they are grouped into different cells. However, there are some obvious errors in the cell segmentation results, which are shown in Figure 5(a), (b). For cell segmentation correction, we propose a post-processing rule: If the distance between a pair of two strokes time adjacent in the cell is greater than the threshold δ_1, the two strokes are regarded as coming from different cells and the original cell is devided into two parts from the stoke pair. The threshold δ_1 is chosen from the validation set.

Algorithm 1 Table block detection

Input: Table strokes outputed by node classification S = $\{s_i|i \in M\}$ (M is the table stroke index set in the document), and edge classification results.
Output: Detected table blocks B
Initialization: B=NULL and the minimum stroke index $prev$ = 0

for all s_i, s_j in S **do**
  if $(j - i)$ >1 **then**
    $b = \{s_k|prev < k < j \quad and \quad k \in M\}$
    B.append(b)
    prev = j
  end if
end for
According to edge classification results, each table block b in B can be segmented into several cells, if the cell number of b is less than th_1, strokes in b is regarded as non-table. And deleted b from B.
for all b in B: **do**
  for all time adjacent strokes (s_{k1}, s_{k2}) in b: **do**
    if the minimum distance of s_{k1} and s_{k2} in Y axis is greater than th_2 or in X axis is greater than th_3 **then**
      b is segmented into two part: b_1 and b_2
      if number of cells in b_1 > 4 **then**: B.append(b_1)
      end if
      if number of cells in b_2 > 4 **then**: B.append(b_2)
      end if
      Delete b in B
    end if
  end for
end for

4 EXPERIMENTS

We have evaluated our proposed method on the IAMOnDo dataset[21] which is an online handwritten document dataset with non-uniform contents. IAMOnDo dataset is consisted of 1000 documents produced by approximately 200 writers including a total of 329849 online strokes and divided into five disjoint subsets. We follow the original division of the dataset, set 0 and 1 are used for the training, set 2 is used to validate system parameters, and set 3 is the test set. Table 1 shows statistics of the dataset.

4.1 Data Augment

Since there are fewer documents containing tables in the dataset, we have done data enhancement operations on the dataset. The

Table 1: Statistics of the IAMOnDo dataset: number of documents and strokes

Set	Docs	Strokes	Table Strokes	Non-Table Strokes
Training	403	142512	79542	62970
Validation	200	68442	6555	61887
Evaluation	203	70543	6866	63677

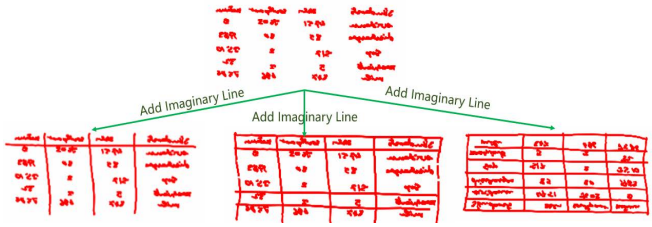


Figure 2: Table data augmented by adding imaginary table lines

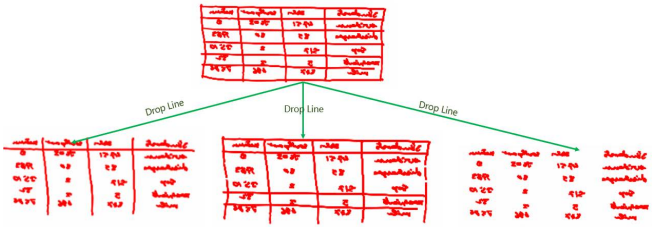


Figure 3: Table data augmented by droping table lines

following two methods are mainly used to enhance the diversity of table.

Imaginary Line Inspired by the idea of imaginary stroke[24], we randomly add a few virtual table line strokes to tables in the train set, and extract features from the added imaginary strokes and real writing strokes, as shown in Figure 2.

Drop Line Inspired by the idea of Dropout [23], we randomly remove table lines from each table, as shown in Figure 3.

After the data enhancement operation, the table document data is increased from 181 to 781.

4.2 Implementation Detail

We implement the multi-task learning graph neural network with the DGL library and its PyTorch backend and train with an NVIDIA GTX 1080Ti GPU. On average, it takes 10s in an epoch training and 10ms to process a test document.

Weight matrix parameters are randomly initialized with normal samples N $(0, \delta^2)$, where $\delta = r + c^2$ and r, c are the numbers of rows and columns in the matrix [3]. The node loss ($\text{Loss}_{\text{node}}$) and edge loss ($\text{Loss}_{\text{edge}}$) is standard cross entropy loss, which is defined as:

$$L(W) = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} logP_t[y_t^{(i)}] \tag{13}$$

Parameter optimization is performed by Adam [22] with batch size 64. We choose the initial learning rate η_0=0.005 and the learning rate decays according to the validation accuracy. The decay rate

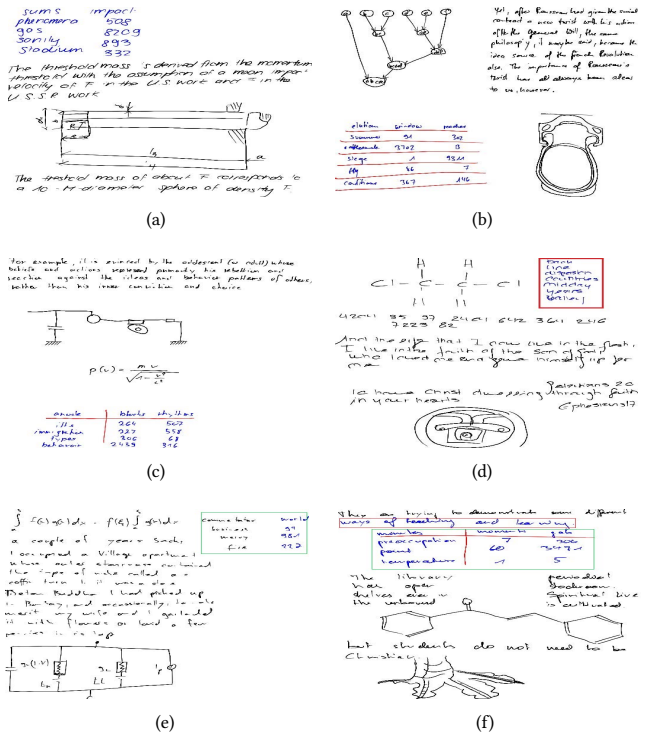


Figure 4: Examples of table detection results

λ = 0.1 and the number of patience round r = 15. We use early stopping based on the performance on the validation set and the best parameter is chosen by the best accuracy on the validation set. To mitigate overfitting, the dropout method [23] is applied in the input of each layer to regularize our model. We fix dropout rate at 0.1 for all dropout layers.

4.3 Results

We define that only if all table and the non-table strokes are completely correctly recognized, the table detection example can be regarded as a true positive (TP) else it is regarded as a false positive (FP). Those table regions that are not detected by our model are false negatives(FN). A correct cell segmentation example is defined as: strokes exactly match with the ground truth in each cell.

Table Detection Results We compare the table detection results of our method with methods in [1],[2] on IAMOnDo dataset. Results are shown in Table 2. Our model is called as MDGAT, and we compute the detection accuracy from three aspects: without data augment, without dense connect and with them all. Visualization results are shown in Figure 4, (a), (b), (c) are correctly detection results, (d), (e), (f) are incorrectly detection results.

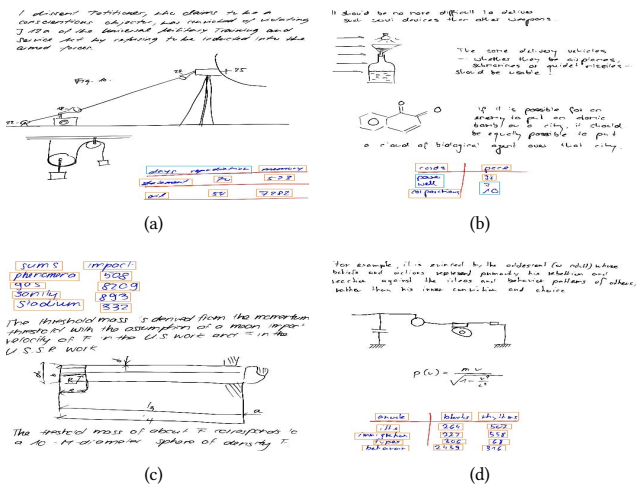
Cell Segmentation Results We compare the cell segmentation results of our method with method in [2] on IAMOnDo dataset. The cell segmentation results are listed in Table 3: edge means cells are segmented only by edge classification results. *DT* represents table regions are detected by the model, *GT* represents ground truth table regions , *p1* is the simple post-processing program, which is listed in Section 3.3. Besides, cell segmentation visualization results

Table 2: Table detection results

Method	Precision	Recall	F
Delaye et al.[1]	-	64.1%	-
Jain et al.[2]	68.32%	67.64%	67.99%
MDGAT without data augment	78.64%	77.02%	77.82%
MDGAT without dense connect	79.00%	77.56%	78.27%
MDGAT	80.20%	79.41%	79.80%

Table 3: Statistics of the cell segmentation:

Method	Precision	Recall	F
edge in DT	78.89%	74.80%	76.79%
Jain[2] in DT	70.75%	67.08%	68.87%
edge in DT + p1	94.45%	89.55%	91.93%
Jain[2] in DT + p1	88.40%	83.81%	86.04%
edge in GT + p1	95.56%	-	-
Jain[2] in GT + p1	89.55%	-	-



(a) (b) (c) (d)

Figure 5: Examples of cell segmentation results

are shown in Figure 5: orange boxes represent cells are segmented correctly, while blue boxes represent cells were segmented incorrectly.

5 CONCLUSION

In this paper, we present a multi-task learning approach for detecting table regions and segmenting cells with graph attention networks in free form online documents. Table detection and cell segmentation tasks are regarded as node and edge classification problems in a graph, which is constructed with strokes. Combining the results of node and edge classification, cells can be segmented while detecting the table region. Our proposed model has been experimentally validated on the online handwritten document dataset IAMOnDo [21] and achieved encouraging results. In the future, we may extend the approach to complicated table detection and work on the table structure recognition task.

6 ACKNOWLEDGMENTS

This work has been supported by the Major Project for New Generation of AI under Grant No. 2018AAA0100400, the National Natural Science Foundation of China (NSFC) grants 61936003 and 61721004.

REFERENCES

[1] A. Delaye, C. L. Liu. 2014. Multi-class Segmentation of Free-Form Online Documents with Tree Conditional Random Fields. International Journal on Document Analysis and Recognition, vol.17: 313-329.
[2] A. k. Jain, A.M. Namboodiri and J. Subrahmonia. 2001. Structure in On-line Documents. International Conference on Document Analysis and Recognition, pp. 844-848.
[3] X. W. Zhang, M. R. Lyu and G. Z. Dai. 2007. Extraction and Segmentation of Tables from Chinese Ink Documents Based on A Matrix Model. Pattern Recognition, vol. 40(7):1855-1867.
[4] O. Ronneberger, P. Fischer and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention, pp. 234-241.
[5] Y. T. Hu, J. B. Huang and A. G. Schwing. 2017. MaskRNN: Instance Level Video Object Segmentation. Advances in Neural Information Processing Systems, pp. 325-334.
[6] J. Blanchard, T. Artières. 2004. On-line Handwritten Documents Segmentation. International Workshop on Frontiers in Handwriting Recognition, pp. 148–153.
[7] M. Shilman, Z. Wei, S. Raghupathy, P. Simard and D. Jones. 2003. Discerning Structure from Freeform Handwritten Notes. International Conference on Document Analyses and Recognition,vol.1, pp. 60–65.
[8] J. Bruna, W. Zaremba and A. Szlam. 2014. Spectral Networks and Locally Connected Networks on Graphs. International Conference on Learning Representations.
[9] D. He, S. Cohen, B. Price, D. Kifer, and C. L. Giles .2017. Multiscale Multi-Task FCN for Semantic Page Segmentation and Table Detection. International Conference on Document Analysis and Recognition, vol.1:254-261.
[10] I. Kavasidis, S. Palazzo, C. Spampinato, C. Pino, D. Giordano, D. Giuffrida, and P. Messina. 2019. A Saliency-Based Convolutional Neural Network for Table and Chart Detection in Digitized Documents. International Conference on Image Analysis and Processing (2) 2019: 292-302.
[11] S. A. Siddiqui, M. I. Malik, S. Agne, A. Dengel, and S. Ahmed. 2018. Decnt: Deep Deformable CNN for Table Detection. IEEE Access, vol. 6: 74151-74161.
[12] M. Gori, G. Monfardini, and F. Scarselli. 2005. A New Model for Learning in Graph Domains. IEEE International Joint Conference on Neural Networks, pp. 729-734.
[13] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. 2009. The Graph Neural Network Model. IEEE Transactions on Neural Networks, vol. 20(1):61–80.
[14] T. N. Kipf, M. Welling. 2017. Semi-supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations (poster).
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is All You Need. Advances in Neural Information Processing Systems, pp.5998-6008.
[16] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y Bengio. 2018. Graph Attention Networks. International Conference on Learning Representations (poster).
[17] L. Gong, Q. Cheng. 2019. Exploiting Edge Features for Graph Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 9203-9211.
[18] G. Huang, Z. Liu.. 2017. Densely Connected Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261-2269.
[19] G. Li, M. Müller, A. Thabet and B. Ghanem. 2019. DeepGCNs: Can GCNs Go as Deep as CNNs? International Conference on Computer Vision, pp. 9266-9275.
[20] J. Y. Ye, Y. M. Zhang and C. L. Liu. 2016. Joint Training of Conditional Random Fields and Neural Networks for Stroke Classification in Online Handwritten Documents. International Conference on Pattern Recognition, pp. 3264-3269.
[21] E. Indermühle, M. Liwicki and H. Bunke. 2010. Iamondo-Database: An Online Handwritten Document Database with Non-uniform Contents. International Workshop on Document Analysis Systems, pp. 97-104.
[22] D. P.Kingma, J. Ba. 2015. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (poster).
[23] N. Srivastava, G. Hinton and A. Krizhevsky. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, vol. 15(1):1929-1958.
[24] M. Okamoto. 1998. Direction-change Features of Imaginary Strokes for On-Line Handwriting Character Recognition. International Conference on Pattern Recognition, vol. 2:1747-1751.

RICAPS: residual inception and cascaded capsule network for broadcast sports video classification

Authors:

Abdullah Aman Khan • University of Electronic Science and Technology of China, Chengdu, China

Saifullah Tumrani • University of Electronic Science and Technology of China, Chengdu, China

Chunlin Jiang • Sichuan Artificial Intelligence Research Institute, Yibin, China

Jie Shao • Sichuan Artificial Intelligence Research Institute, Yibin, China and University of Electronic Science and Technology of China, Chengdu, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446296

The field of broadcast sports video analysis requires attention from the research community. Identifying the semantic actions within a broadcast sports video aids better video analysis and highlight generation. One of the key challenges posed to sports video analysis is the availability of relevant datasets. In this paper, we introduce a new dataset SP-2 related to broadcast sports video (available at https://github.com/abdkhanstd/Sports2). SP-2 is a large dataset with several annotations such as sports category (class), playfield scenario, and game action. Along with the introduction of this dataset, we focus on accurately classifying the broadcast sports video category and propose a simple yet elegant method for the classification of broadcast sports video. Broadcast sports video classification plays an important role in sports video analysis as different sports games follow a different set of rules and situations. Our method exploits and explores the true potential of capsule network with dynamic routing, which was introduced recently. First, we extract features using a residual convolutional neural network and build temporal feature sequences. Further, a cascaded capsule network is trained using the extracted feature sequence. Residual inception cascaded capsule network (RICAPS) significantly improves the performance of broadcast sports video classification as deeper features are captured by the cascaded capsule network. We conduct extensive experiments on SP-2 dataset and compare the results with

previously proposed methods, and the results show that RICAPS outperforms the previously proposed methods.

Transfer non-stationary texture with complex appearance

Authors:

Cheng Peng • Beijing University of Technology, Beijing, Chaoyang, China

Na Qi • Beijing University of Technology, Beijing, Chaoyang, China

Qing Zhu • Beijing University of Technology, Beijing, Chaoyang, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446297

Texture transfer has been successfully applied in computer vision and computer graphics. Since non-stationary textures are usually complex and anisotropic, it is challenging to transfer these textures by simple supervised method. In this paper, we propose a general solution for non-stationary texture transfer, which can preserve the local structure and visual richness of textures. The inputs of our framework are source texture and semantic annotation pair. We record different semantics as different regions and obtain the color and distribution information from different regions, which is used to guide the the low-level texture transfer algorithm. Specifically, we exploit these local distributions to regularize the texture transfer objective function, which is minimized by iterative search and voting steps. In the search step, we search the nearest neighbor fields of source image to target image through Generalized PatchMatch (GPM) algorithm. In the voting step, we calculate histogram weights and coherence weights for different semantic regions to ensure color accuracy and texture continuity, and to further transfer the textures from the source to the target. By comparing with state-of-the-art algorithms, we demonstrate the effectiveness and superiority of our technique in various non-stationary textures.

Story segmentation for news broadcast based on primary caption

Authors:

Heling Chen • Wuhan University, Wuhan, China

Zhongyuan Wang • Wuhan University, Wuhan, China

Yingjiao Pei • Wuhan University, Wuhan, China

Baojin Huang • Wuhan University, Wuhan, China

Weiping Tu • Wuhan University, Wuhan, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446298

In the information explosion era, people only want to access the news information that they are interested in. News broadcast story segmentation is strongly needed, which is an essential basis for personalized delivery and short video. The existing advanced story boundary segmentation methods utilize semantic similarity of subtitles, thus entailing complex semantic computation. The title texts of news broadcast programs include headline (or primary) captions, dialogue captions and the channel logo, while the same story clips only render one primary caption in most news broadcast. Inspired by this fact, we propose a simple method for story segmentation based on the primary caption, which combines YOLOv3 based primary caption extraction and preliminary location of boundaries. In particular, we introduce mean hash to achieve the fast and reliable comparison for detected small-size primary caption blocks. We further incorporate scene recognition to exact the preliminary boundaries, because the primary captions always appear later than the story boundary. Experimental results on two Chinese news broadcast datasets show that our method enjoys high accuracy in terms of R, P and F1-measures.

Intermediate coordinate based pose non-perspective estimation from line correspondences

Authors:

Yujia Cao • Xi'an Jiaotong University, Xi'an, Shaanxi, China

Zhichao Cui • Xi'an Jiaotong University, Xi'an, Shaanxi, China

Yuehu Liu • Xi'an Jiaotong University, Xi'an, Shaanxi, China

Xiaojun Lv • China Academy of Railway Sciences, Beijing, China

Kaibei Peng • China Academy of Railway Sciences, Beijing, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446299

In this paper, a non-iterative solution to the non-perspective pose estimation from line correspondences was proposed. Specifically, the proposed method uses an intermediate camera frame and an intermediate world frame, which simplifies the expression of rotation matrix by reducing to the two freedoms from three in the rotation matrix R. Then formulate the pose estimation problem into an optimal problem. Our method solve the parameters of rotation matrix by building the fifteenth-order and fourth-order univariate polynomial. The proposed method can be applied into the pose estimation of the perspective camera. We utilize both the simulated data and real data to conduct the comparative experiments. The experimental results show that the proposed method is comparable or better than existing methods in the aspects of accuracy, stability and efficiency.

An autoregressive generation model for producing instant basketball defensive trajectory

Authors:

Huan-Hua Chang • National Cheng Kung University, Taiwan

Wen-Cheng Chen • National Cheng Kung University, Taiwan

Wan-Lun Tsai • National Cheng Kung University, Taiwan

Min-Chun Hu • National Tsing Hua University, Taiwan

Wei-Ta Chu • National Cheng Kung University, Taiwan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446300

Learning basketball tactic via virtual reality environment requires real-time feedback to improve the realism and interactivity. For example, the virtual defender should move immediately according to the player's movement. In this paper, we proposed an autoregressive generative model for basketball defensive trajectory generation. To learn the continuous Gaussian distribution of player position, we adopt a differentiable sampling process to sample the candidate location with a standard deviation loss, which can preserve the diversity of the trajectories. Furthermore, we design several additional loss functions based on the domain knowledge of basketball to make the generated trajectories match the real situation in basketball games. The experimental results show that the proposed method can achieve better performance than previous works in terms of different evaluation metrics.

Real-time arbitrary video style transfer

Authors:

Xingyu Liu • Shenzhen Research Institute of Nanjing University, Shenzhen, China and Nanjing University, Nanjing, China

Zongxing Ji • Shenzhen Research Institute of Nanjing University, Shenzhen, China and Nanjing University, Nanjing, China

Piao Huang • Nanjing University, Nanjing, China

Tongwei Ren • Shenzhen Research Institute of Nanjing University, Shenzhen, China and Nanjing University, Nanjing, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446301

Video style transfer aims to synthesize a stylized video that has similar content structure with a content video and is rendered in the style of a style image. The existing video style transfer methods cannot simultaneously realize high efficiency, arbitrary style and temporal consistency. In this paper, we propose the first real-time arbitrary video style transfer method with only one model. Specifically, we utilize a three-network architecture consisting of a prediction network, a stylization network and a loss network. Prediction network is used for extracting style parameters from a given style image; Stylization network is for generating the corresponding stylized video; Loss network is for training prediction network and stylization network, whose loss function includes content loss, style loss and temporal consistency loss. We conduct three experiments and a user study to test the effectiveness of our method. The experimental results show that our method outperforms the state-of-the-arts.

C3VQG: category consistent cyclic visual question generation

Authors:

Shagun Uppal • IIIT-Delhi, India

Anish Madan • IIIT-Delhi, India

Sarthak Bhagat • IIIT-Delhi, India

Yi Yu • NII, Japan

Rajiv Ratn Shah • IIIT-Delhi, India

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446302

Visual Question Generation (VQG) is the task of generating natural questions based on an image. Popular methods in the past have explored image-to-sequence architectures trained with maximum likelihood which have demonstrated meaningful generated questions given an image and its associated ground-truth answer. VQG becomes more challenging if the image contains rich contextual information describing its different semantic categories. In this paper, we try to exploit the different visual cues and concepts in an image to generate questions using a variational autoencoder (VAE) without ground-truth answers. Our approach solves two major shortcomings of existing VQG systems: (i) minimize the level of supervision and (ii) replace generic questions with category relevant generations. Most importantly, by eliminating expensive answer annotations, the required supervision is weakened. Using different categories enables us to exploit different concepts as the inference requires only the image and the category. Mutual information is maximized between the image, question, and answer category in the latent space of our VAE. A novel category consistent cyclic loss is proposed to enable the model to generate consistent predictions with respect to the answer category, reducing redundancies and irregularities. Additionally, we also impose supplementary constraints on the latent space of our generative model to provide structure based on categories and enhance generalization by encapsulating decorrelated features within each dimension. Through extensive experiments, the proposed model, C3VQG outperforms state-of-the-art VQG methods with weak supervision.

Cross-modal learning for saliency prediction in mobile environment

Authors:

Dakai Ren • Beijing University of Posts and Telecommunications, Beijing, China

Xiangming Wen • Beijing University of Posts and Telecommunications, Beijing, China

Xiaoya Liu • Xinyang Vocational and Technical College, Xinyang, China

Shuai Huang • Huazhong University of Science and Technology, Wuhan, China

Jiazhong Chen • Huazhong University of Science and Technology, Wuhan, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446304

The existing researches reveal that a significant impact is introduced by viewing conditions for visual perception when viewing media on mobile screens. This brings two issues in the area of visual saliency that we need to address: how the saliency models perform in mobile conditions, and how to consider the mobile conditions when designing a saliency model. To investigate the performance of saliency models in mobile environment, eye fixations in four typical mobile conditions are collected as the mobile ground truth in this work. To consider the mobile conditions when designing a saliency model, we combine viewing factors and visual stimuli as two modalities, and a cross-modal based deep learning architecture is proposed for visual attention prediction. Experimental results demonstrate the model with the consideration of mobile viewing factors often outperforms the models without such consideration.

Objective object segmentation visual quality evaluation based on pixel-level and region-level characteristics

Authors:

Ran Shi • Nanjing University of Science and Technology, Nanjing, China

Jian Xiong • Nanjing University of Posts and Telecommunications, Nanjing, China

Tong Qiao • Dianzi University, Hang Zhou, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446305

Objective object segmentation visual quality evaluation is an emergent member of the visual quality assessment family. It aims at developing an objective measure instead of a subjective survey to evaluate the object segmentation quality in agreement with human visual perception. It is an important benchmark to assess and compare performances of object segmentation methods in terms of the visual quality. In spite of its essential role, it still lacks of sufficient studying compared with other visual quality evaluation researches. In this paper, we propose a novel full-reference objective measure including a pixel-level sub-measure and a region-level sub-measure. For the pixel-level sub-measure, it assigns proper weights to not only false positive pixels and false negative pixels but also true positive pixels according to their certainty degrees. For the region-level sub-measure, it considers location distribution of the false negative errors and correlations among neighboring pixels. Thus, by combining these two sub-measures, our measure can evaluate similarity of area, shape and object completeness between one segmentation result and its ground truth in terms of human visual perception. In order to evaluate the performance of our proposed measure, we tested it on an object segmentation subjective visual quality assessment database. The experimental results demonstrate that our proposed measure with good robustness performs better in matching subjective assessments compared with other state-of-the-art objective measures.

Text-based visual question answering with knowledge base

Authors:

Fang Zhou • University of Science and Technology Beijing, Beijing Shi, China

Bei Yin • University of Science and Technology Beijing, Beijing Shi, China

Zanxia Jin • University of Science and Technology Beijing, Beijing Shi, China

Heran Wu • University of Science and Technology Beijing, Beijing Shi, China

Dongyan Zhang • University of Science and Technology Beijing, Beijing Shi, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446306

Text-based Visual Question Answering(VQA) usually needs to analyze and understand the text in a picture to give a correct answer for the given question. In this paper, a generic Text-based VQA with Knowledge Base (KB) is proposed, which performs text-based search on text information obtained by optical character recognition (OCR) in images, constructs task-oriented knowledge information and integrates it into the existing models. Due to the complexity of the image scene, the accuracy of OCR is not very high, and there are often cases where the words have individual character that is incorrect, resulting in inaccurate text information; here, some correct words can be found with help of KB, and the correct image text information can be added. Moreover, the knowledge information constructed with KB can better explain the image information, allowing the model to fully understand the image and find the appropriate text answer. The experimental results on the TextVQA dataset show that our method improves the accuracy, and the maximum increment is 39.2%.

Attention-constraint facial expression recognition

Author:

Qisheng Jiang • University of Science and Technology of China, Hefei, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446307

To make full use of existing inherent correlation between facial regions and expression, we propose an attention-constraint facial expression recognition method, where the prior correlation between facial regions and expression is integrated into attention weights for extracting better representation. The proposed method mainly consists of four components: feature extractor, local self attention-constraint learner (LSACL), global and local attention-constraint learner (GLACL) and facial expression classifier. Specifically, feature extractor is mainly used to extract features from overall facial image and its corresponding cropped facial regions. Then, the extracted local features from facial regions are fed into local self attention-constraint learner, where some prior rank constraints summarized from facial domain knowledge are embedded into self attention weights. Similarly, the rank correlation constraints between respective facial region and a specified expression are further embedded into global-to-local attention weights when the global feature and local features from local self attention-constraint learner are fed into global and local attention-constraint learner. Finally, the feature from global and local attention-constraint learner and original global feature are fused and passed to facial expression classifier for conducting facial expression recognition. Experiments on two benchmark datasets validate the effectiveness of the proposed method.

EvoGAN: an evolutionary GAN for face aging and rejuvenation

Authors:

Lianli Gao • University of Electronic Science and Technology of China, Chengdu, China

Jingqiu Zhang • University of Electronic Science and Technology of China, Chengdu, China

Jingkuan Song • University of Electronic Science and Technology of China, Chengdu, China

HengTao Shen • University of Electronic Science and Technology of China, Chengdu, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446323

In biology, evolution is the gradual change in the characteristics of a species over several generations. It has two properties: 1) The change is gradual, and 2) long-term changes are relied on short-term changes. Face aging/rejuvenation, which renders younger or elder facial images, follows the principles of evolution. Inspired by this, we propose an <u>Evo</u>lutionary <u>GAN</u>s (EvoGAN) for face aging/rejuvenation by making each age transformation smooth and decomposing a long-term transformation into several short-terms. Specifically, since short-term facial changes are gradual and relatively easy to render, we first divide the ages into several groups (i.e., chronologically from child, adult to elder). Then, for each pair of adjacent groups, we design two age transforms for face aging and rejuvenation, which are supposed to preserve personal identify information and predict age-specific characteristics. Compared with the mainstream for face aging/rejuvenation, i.e., conditional

GANs based methods utilizing one-hot age vector as an age transformation condition, our smooth EvoGAN abandons this condition and can better predict age-specific factors (e.g., the drastic shape and appearance change from an adult to a child). To evaluate our EvoGAN, we construct a challenging dataset FFHQ_Age. Extensive experiments conducted on the

dataset demonstrate that our model is able to generate significantly better results than the state-of-the-art methods qualitatively and quantitatively.

Destylization of text with decorative elements

Authors:

Yuting Ma • UCAS

Fan Tang • Jilin University

Weiming Dong • CAS

Changsheng Xu • CAS

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446324

Style text with decorative elements has a strong visual sense, and enriches our daily work, study and life. However, it introduces new challenges to text detection and recognition. In this study, we propose a text destylized framework, that can transform the stylized texts with decorative elements into a type that is easily distinguishable by a detection or recognition model. We arranged and integrate an existing stylistic text data set to train the destylized network. The new destylized data set contains English letters and Chinese characters. The proposed approach enables a framework to handle both Chinese characters and English letters without the need for additional networks. Experiments show that the method is superior to the state-of-the-art style-related models.

Destylization of Text with Decorative Elements

Yuting Ma
NLPR, Institute of Automation, CAS
School of Artificial Intelligence, UCAS
mayuting2018@ia.ac.cn

Fan Tang*
School of Artificial Intelligence, Jilin University
tangfan@jlu.edu.com

Weiming Dong*
NLPR, Institute of Automation, CAS
CASIA-LLVision Joint Lab
weiming.dong@ia.ac.cn

Changsheng Xu
NLPR, Institute of Automation, CAS
CASIA-LLVision Joint Lab
csxu@nlpr.ia.ac.cn

ABSTRACT

Style text with decorative elements has a strong visual sense, and enriches our daily work, study and life. However, it introduces new challenges to text detection and recognition. In this study, we propose a text destylized framework, that can transform the stylized texts with decorative elements into a type that is easily distinguishable by a detection or recognition model. We arranged and integrate an existing stylistic text data set to train the destylized network. The new destylized data set contains English letters and Chinese characters. The proposed approach enables a framework to handle both Chinese characters and English letters without the need for additional networks. Experiments show that the method is superior to the state-of-the-art style-related models.

CCS CONCEPTS

• **Applied computing** → **Fine arts**; Computer-assisted instruction; • **Computing methodologies** → *Image representations.*

KEYWORDS

Text destylization; Style transfer; Decorative elements

ACM Reference Format:
Yuting Ma, Fan Tang, Weiming Dong, and Changsheng Xu. 2021. Destylization of Text with Decorative Elements. In *ACM Multimedia Asia (MMAsia '20), March 7–9, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3444685.3446324

1 INTRODUCTION

Text destylization restores text with various complex styles to a style-less state, that can be handled by text detection and recognition approaches. With the development of deep learning, the replication and transfer of styles have become increasingly popular. Ordinary people can "create" their desired art forms without professional training. Text is a prominent visual element in 2D design, and is almost everywhere. Compared with ordinary text, text with artistic rendering are required by people. Artistic effects, such as color, outline, shadow, reflection, luminescence, and texture, are additional stylistic features of text. Artistic writing is a type of art widely used in design and media. Through color, texture, shading and other text effects, combined with additional decorative elements, artistic text becomes visually pleasing and can convey more semantic information vividly. Visual effects are commonly applied to text in graphic design because of their special role in visual design. These effects are also applied in painting synthesis and photography post processing.

Artificial intelligence has developed rapidly in the past decade, and various social software and self-media applications have been created and popularized. As a result, even ordinary people can become micro social media developers. At the same time, the application of numerous text styles and easy production have produced many different styles of text, such as posters, advertising design, e-commerce platforms and visual creation tasks. However, although these text styles make our visual experience colorful, they bring new challenges to automatic text detection and recognition. Traditionally, text detection and recognition approaches are designed for ordinary text and do not address rotation, bending, mirroring and other transformations. Text detection and recognition tasks can be challenging in an open environment, and cannot meet the increasingly complex needs of users. In related engineering tasks, the study of text destylization can partly improve the accuracy and speed of depth detection and recognition models, which have a wide range of application scenarios and practical usage [33].

Text destylization is a new problem, which is the reverse application of text style transfer. It can be viewed as a problem of transforming from the stylized domain to the source domain, but it has been disregarded by scholars.Yang et al. [30] were the first to raise the issue by designing a subnetwork that removed the basic style. The destylization problem with decorative elements has not been studied, due to the neglect of many important attributes, such as decorative elements, orientation, and rule structure. To solve this problem, this study pays special attention to decorative elements in text designs and proposes a new text style transformation framework to remove such elements.

In summary, our contributions are threefold. First, we define a new problem of text destylization with decorative elements and propose a new framework to solve this problem. The framework

*Co-corresponding authors

MMAsia '20, March 7–9, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8308-0/21/03…$15.00
https://doi.org/10.1145/3444685.3446324

can effectively solve the impact of decorative elements on partial pixel occlusion of text. Second, our framework can destylize English letters and Chinese characters without training additional subnetworks. Third, through the permutation and combination of an existing text style data set, our data set contains both English letters and Chinese characters. Experiments show that our model can handle not only a single character, but also several simple words.

2 RELATED WORK

2.1 Neural Style Transfer

In computer vision, style transfer is usually studied as a generalized texture synthesis problem, that is, to extract texture from source images and transfer it to targets [7–11]. Gatys et al. [12] were the first to study the use of CNN in reproducing a famous painting style on natural images. The groundbreaking work of their team demonstrated the power of CNN in representing texture modeling. Their results showed that CNN can extract content information from an arbitrary photograph and style information from a famous artwork. On the basis of this discovery, Gatys et al [12] proposed the use of feature activation of CNN to recombine the content of a given photo and the style of famous artworks. The main purpose of CNN is to extract style features, such as texture. However, its structure does not contribute much to synthesis. To address this problem, Champandard et al. [2] improved the network structure of CNN by using a composite image that was enhanced with semantic information during the generation phase. The aim was to narrow the gap between the generation model and the pixel level classification neural network. Recently, many arbitrary style transfer methods were proposed [5, 6, 27, 35].

2.2 Image-to-Image Translation

The purpose of image-to-image translation is to learn an image generation function that maps the input image in the source domain to the target domain: examples include sketch to portrait [3], image colorization [32, 34], and rain removal [23, 31]. Hertzmann et al. [17] proposed a single image pair non-parametric framework. Isola et al. [18] developed a universal framework called Pix2Pix, in conjunction with GANs [13]. Pix2Pix combines an L1 loss and an adversarial loss with paired data samples from two domains. This approach is driven by paired data, which are sometimes difficult to obtain. To overcome this limitation, Zhu et al. [36] designed a CycleGAN that can learn to translate images without pairs of ground truth. Choi et al. [4] proposed StarGAN using a single model to handle multi-domain translation; it utilizes a one-hot vector to specify the target domain. However, the extension to a new domain is still expensive. To address this issue, Liu et al. [21] presented a few-shot, unsupervised image-to-image translation algorithm that works on a previously unseen target domain. Based on these image-to-image translation methods, our work focuses on the text destylization problem.

2.3 Text style transfer

The research on text style transfer developed relatively late. Before 2017, studies in the field of text style mainly focused on the strokes of text fonts. HelpHanding [22] is used to study the authoring of strokes from six degrees of freedom by employing graphic methods. EasyFont [20], whice is mainly about handwriting font style transfer can transfer the users' handwriting font style to the specified text so that the text looks like the users' own handwriting. GlyphGAN [15] uses DCGAN [24] as the base model for font generation. Given that the same style vector is used, the resulting fonts tend to have the same style. Yang et al. [29] were the first to raise the issue of text effect transfer. A matching and synthesis method based on the relative positions of image patches on hieroglyphics was proposed. This method is susceptible to the difference in hieroglyphics and requires a large amount of computation. Mc-GAN [1] combines font transmission and text effect transmission by adopting two continuous subnetworks and trains them end-to-end by using the synthesized font data and the collected text effect data. An unsupervised artistic word generation algorithm [25] has also been developed; the algorithm is different from the supervised method that required a pixel-level aligned original text image as the guide. The unsupervised method can deal with arbitrary style images without the corresponding original texts. However, due to the lack of relevant data sets, only a few studies have been conducted on text effect style migration. Yang proposed TET-GAN [30] that uses the learning method of GAN to build a data-driven model, which can learn the accurate mapping relationship between glyphics and character effects from a large amount of data. Although some research has been performed on text style transfer, only a few studies focused on text destylization. A new framework of character destylization is proposed in this study in accordance with actual project needs.

3 METHOD AND TRAINING

3.1 Destylized Networks

Following the network architecture of GANs [13], our text destylization model is a combination of a generator G, a discriminator D and a auxiliary classifier. Adopted from CycleGAN [36], our model has a generator network composed of two convolutional layers with a stride size of two for downsampling, six residual blocks [16], and two transposed convolutional layers with a stride size of two for upsampling. We use PatchGAN [18] as the discriminator. The auxiliary classifier consists of a convolutional layer and a three-layer MLP and is used to categorize text content to help the network improve its handling of details. The network structure is shown in Figure 1. Given D_x, which is a styled text image with extra decorative elements, generator G learns to generate a fake raw text image $F_x = G(D_x)$. Discriminator D needs to distinguish whether the input is real or generated. We use instance normalization [26] for the generator but no normalization is applied for the discriminator. The loss function is a combination of WGAN-GP [14], L1 loss and CrossEntropy loss, as follows:

$$L = \lambda_{adv} L_{adv} + \lambda_{L1} L_1 + \lambda_{L_{au}} L_{au} \tag{1}$$

where.

$$L_1 = \| F_x - \widetilde{F}_x \|_1 \tag{2}$$

$$\begin{aligned} L_{adv} = & E_{\widetilde{F}_x}[D(\widetilde{F}_x, D_x)] \\ & - E_{F_x}[D(F_x, D_x)] \\ & + \lambda_{gp} E_{\widehat{F}_x}[(\| \nabla D(\widehat{F}_x, D_x) \|_2 - 1)^2] \end{aligned} \tag{3}$$

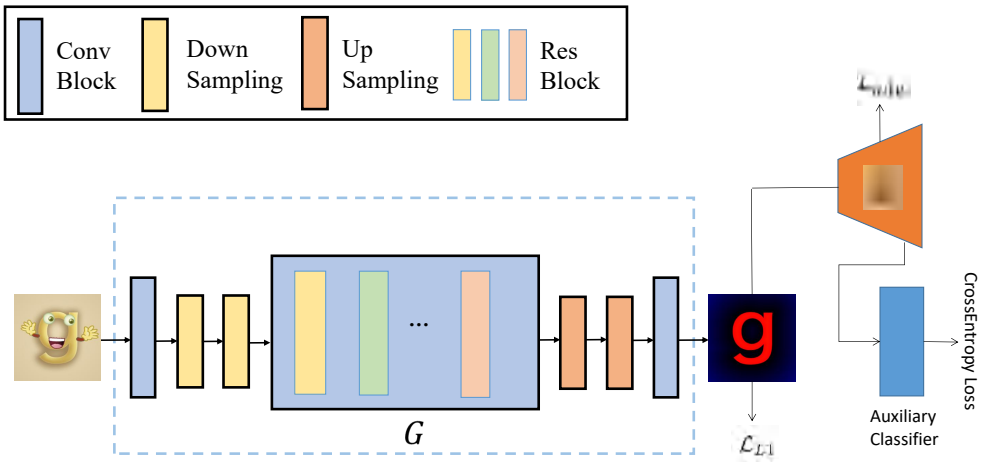


Figure 1: The network structure of our model.

$$L_{au} = -\sum_{i=1}^{C} y^{(i)} * log\, \widehat{y}^{(i)} \tag{4}$$

where. $\widetilde{F}_x$ is the ground truth, and $\widehat{F}_x$ is uniformly sampled along the straight lines between the sampling of F_x and $\widetilde{F}_x$. L_{au} is the cross-entropy loss. $\widehat{y}$ is the value that the output value of the auxiliary classifier is processed by Softmax. y is the true label, and C is the total number of categories.

3.2 Training

Our models are trained using the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We also use a fixed learning rate of 2e-4 throughout the entire training stage. For all experiments, the resolution of images increases from 64, 128, up to 256, and layers are gradually added to the front and rear of the generator. To enhance the effect of network learning and fully utilize GPU and CPU resources, we use different batch sizes for images with different resolutions. When the resolutions of the images are 64*64, 128*128 and 256*256, the corresponding batch size are 200, 90 and 40, respectively. A progressive growing strategy [19] is used in the destylized network to stabilize the training process. The networks are trained on NVIDIA Tesla M40 GPU. We rearrange and combine an existing text style data set [28, 30], which contains about 60,000 images of Chinese characters and English letters, for our network training and testing. For the training of the auxiliary classifier, we define different Chinese characters and English letters(case sensitive) as one category respectively. The class labels use the form of one-shot. For all experiments, we set $\lambda_{gp} = 10$, $\lambda_{L1} = 100$, and $\lambda_{adv} = \lambda_{L_{au}} = 1$.

4 EXPERIMENTS AND RESULTS

We selected five the state-of-the-art models of image-to-image translation and text style transfer for comparisons. The five models are NST [12], Doodle [2], CycleGAN [36], Pix2pix [18], and TET-GAN [30]. We first present the generated results of our model together with those of the compared models. Second, we compare our model with the five other models through subjective and objective evaluation. Lastly, we show the destylized results of our model for different combinations of text and decorations and several complex words.

4.1 Qualitative analysis

We compare the proposed text effect transfer network with five state-of-art transfer methods in Figure 2. At present, no individual or team has conducted extensive research on text destylization, and comparative methods for reference are few. In view of this situation, we select the most advanced methods of image-to-image translation and text style transfer to verify the superiority of our model in the destylization of decorative text.

Image-to-image translation and text destylization are currently the two most relevant research directions for text destylization. Neural Style Transfer (NST) [12] and Neural Doodle [2] are image style transfer methods. NST uses CNNs to transfer the style of an image to another. Doodles [2] uses neural-based patch fusion and has a context-sensitive manner in the algorithm. However, NST and Doodles do not learn the relationship between text style and font, resulting in a fuzzy and confused texture structure. CycleGAN [36] and Pix2Pix [18] are image-to-image translation methods based on GAN, and they are all re-trained on our dataset. The inputs of Pix2Pix [18] and CycleGAN [36] are revised to be the same as our input. CycleGAN only captures some of the texture features of the style and font. Thus, it fails in text destylization. Pix2Pix produces irregular textures and fonts. Benefiting from the progressive growing strategy [19] and the WGAN-GP [14], our model is more stable than Pix2Pix. Given that TET-GAN [30] involves the migration and removal of the basic style without considering decorative elements, the destylization of decorative text could fail because decorative elements block the font. From these comparisons, we could conclude that our model can not only reconstruct the details of text content, but also eliminate the influence of decorative elements.

In addition to these comparative experiments, we present the results of our model for more complex cases, as follows:

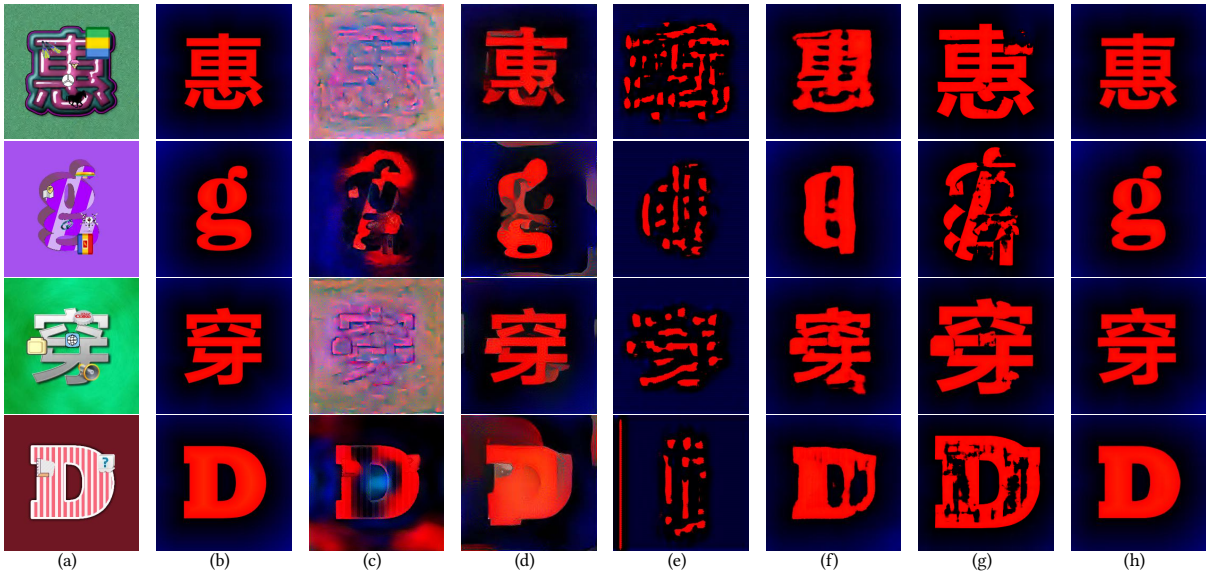


Figure 2: Comparison with current more advanced methods. The first column is the style image with decorative elements that the model inputs. The second column is the corresponding ground truth. The third column is the result of NST [12]. The fourth column is the result of the Doodle [2]. The fifth column is the result of CycleGAN [36]. The sixth column is the result of Pix2Pix [18]. The seventh column is the result of TET-GAN [30]. The last column is the result of our model.

- Same word, different fonts, same decoration. In this case, we use different fonts with the same decoration for the same word. We demonstrate that our model can remove decorations for different fonts, and can be unaffected by font changes. See Figure 3.
- Word combinations. In addition to dealing with individual Chinese characters and English letters, our model can also deal with simple vocabulary. See Figure 4.
- Same word, different decorations. We use different decorative elements to form different decorative styles for words with the same background, to demonstrate the capability of our model to remove decorative elements. As shown in Figure 5, both Chinese characters and English letters are included.

4.2 Subjective and Objective Evaluations

We conduct a user survey and objective index calculation to prove the superiority of our model over other models comprehensively and objectively.

User Study. From the results of the qualitative experiments, we randomly selected 100 groups of different text destylized renderings to make the questionnaire. We receive 40 responses, among which 37 responses are vaild. A total of 3,700 votes are obtained. Our model has 3276 votes, NST has 126 votes, Doodle has 48 votes, CycleGAN has 6 votes, Pix2Pix has 73 votes, and TET-GAN has 171 votes. On the basis of the feedback results, we compute the vote statistics and perform result comparisons, as shown in the Table 1.

Objective Indicators. We selected four metrics, namely, FID, mIoU, RMSE and PSNR. FID is the mean value between the ground truth and the generated value after extracting the feature vector, and evaluating the distance of the covariance. The closer the generated results are to the truth features, the smaller the square of the mean difference is, the smaller their covariance is, and the smaller the

Figure 3: Same word, different font, same decoration. The first and third lines have the same word, and four different font styles can be seen. The second and fourth rows are the destylized results of our model.

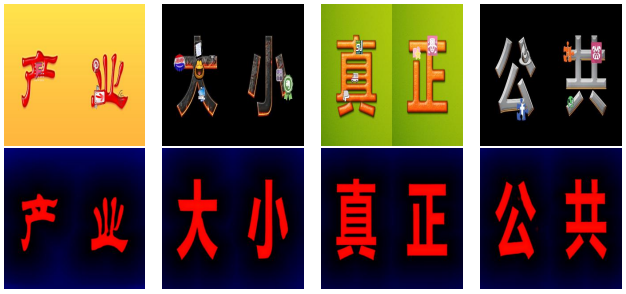


Figure 4: Word combinations. Destylized results of our model for complex words.

Figure 5: Same word, different decorations. The first and third lines respectively show two Chinese characters and two English letters have three different styles of decorative elements. The second and fourth rows are the destylized results of our model.

Table 1: Our user study results. The first and second column are respectively the number of votes and the winning rate of each method.

Method	Votes	Win Rate
NST	126	3.41%
Doodle	48	1.30%
CycleGAN	6	0.16%
Pix2pix	73	1.97%
TET-GAN	171	4.62%
Ours	**3276**	**88.54%**

Table 2: Model performance evaluation index.

Method	FID	mIoU	RMSE	PSNR
NST	285.8983	0.1416	0.0426	30.8222
Doodle	239.7642	0.2290	0.0408	30.6476
CycleGAN	250.3068	0.1197	0.0419	36.3651
Pix2pix	150.5688	0.5452	0.0422	38.2636
TET-GAN	83.5101	0.5677	0.0386	39.7569
Ours	**21.0336**	**0.8746**	**0.0336**	**42.4457**

Table 3: U-net refers to the network generator using U-NET. Resnet-4, Resnet-6 and Resnet-8 respectively refer to a ResNet network that uses four, six and eight residual blocks. Au refers to the addition of auxiliary classifiers to the network.

Method	FID	mIoU	RMSE	PSNR
U-Net	40.7418	0.6379	0.0371	40.4102
ResNet-4	108.7667	0.5169	0.0427	37.7571
ResNet-8	95.5848	0.6543	0.0401	37.2107
ResNet-6	38.8962	0.8659	0.0340	40.8679
ResNet-6+Au	**23.6686**	**0.8993**	**0.0299**	**42.7519**

sum of FID is. MIoU is the ratio of the intersection and union of two sets of ground truth and generated values. The larger the mIoU is, the more the intersection is and the closer the generated value is to the ground truth. RMSE measures the root-mean-square error between the generated value and ground truth. The smaller RMSE is, the closer the generated value is to the ground truth. PSNR is the peak signal-to-noise ratio of ground truth and the generated value. The larger the PSNR is, the better the generated results are. The performance of our model in comparison with that of the five other models is shown in Table 2.

As can be seen from the experimental results in the table, our model is superior to the five other models in four aspects in terms of the decorative text destylization task.

4.3 Ablation Study

The architecture of our model comprises a generator, a discriminator and an auxiliary classifier. In this section, we discuss the influence of each part of our model. We verify the effect of different model structures from three aspects: the influence of different generators, number of residual blocks, and branches of auxiliary classifier on the effect of the model.

U-Net and ResNet. We used U-Net and ResNet as model generators to perform experiments. The objective indicators of the experimental results between U-Net and ResNet-6 are compared in Table 3. Additional experimental results are provided in Supplementary Materials.

Auxiliary classifier. We verify the role of the auxiliary classifier by comparing the addition and non-addition of the auxiliary classifier to the model. Table 3 compares the objective indicators of the experimental results of ResNet-6 and ResNet-6+Au. From the results, we can see that the helper classifier can be added to help the model to learn additional details on the content. Other experimental results are provided in Supplementary Materials.

Number of residual blocks. Our model uses ResNet as a generator. We set the number of residual blocks as 4, 6 and 8 in our ablation study to verify the impact of the number of residual blocks on the network. Table 3 compares the objective indicators of the experimental results of ResNet-4, ResNet-8 and ResNet-6. Additional experimental results are provided in Supplementary Materials.

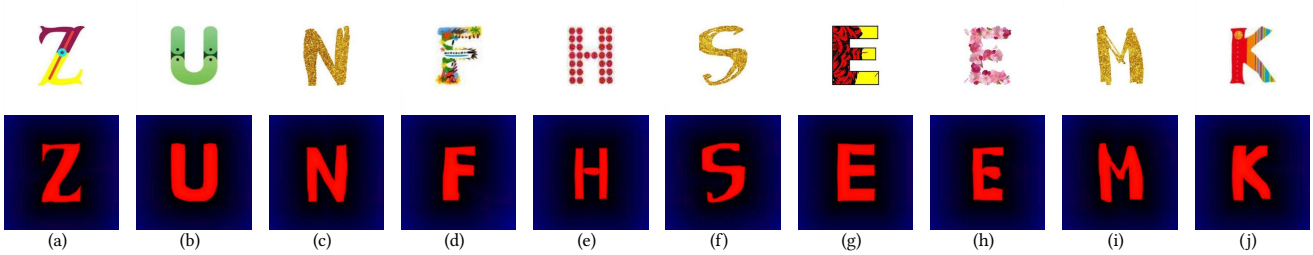


Figure 6: The first line is the style images in the wild. The second line is the processing results of our model.

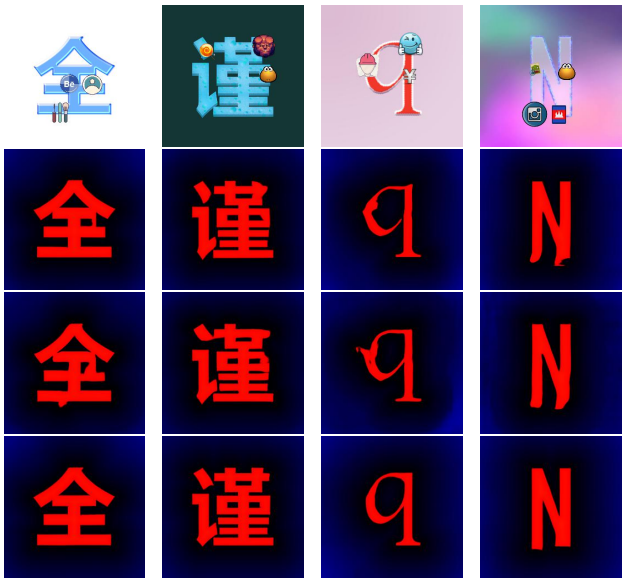


Figure 7: The first row is a text-style image. The second row is the result of the C64-C128 discriminator experiment. The third row is the result of the C64-C128-C256-C512-C512-C512 discriminator experiment. The fourth row is the result of the C64-C128-C256-C512 discriminator experiment.

Single-scale and multi-scale training. In the training process, multi-scale training is adopted to cut out patches of 64, 128, and 256 and send them into the model. A single-scale comparison experiment is conducted with patch 256 to prove the effectiveness of multi-scale training. The objective indicators of the different scale training methods are compared in Table 4. Additional experimental results are provided in Supplementary Materials.

Different Discrimators. In our model, the structure of the discriminator is C64-C128-C256-C512. In order to compare the effects of different discriminators on the model, ablation experiments of three different discrimators were carried out. The other two discriminators have the following structure: C64-C128 and C64-C128-C256-C512-C512-C512. All other discriminators follow the same basic architecture, with depth varied to modify the receptive field size. The experimental results are shown in Figure 7.

Table 4: The first and second lines respectively show the results of the single-scale training and the multi-scale training.

Method	FID	mIoU	RMSE	PSNR
single scale	95.7025	0.2441	0.0405	38.0133
multiple scale	**23.6686**	**0.8993**	**0.0299**	**42.7519**

Table 5: Discrimator1 is C64-C128. Discrimator2 is C64-C128-C256-C512-C512-C512.

Method	FID	mIoU	RMSE	PSNR
Discrimator1	55.0458	0.8244	0.0375	42.2382
Discrimator2	90.7623	0.7737	0.0396	40.7921
Ours	**23.6686**	**0.8993**	**0.0299**	**42.7519**

4.4 Unseen Styles

We also collected 1K artistic text of various text effects from the Internet. These styled text effects in the wild are used to verify the generalization of our model. The experimental results are shown in Figure 6. As can be seen from the results, our model performs well in the unseen style data.

5 CONCLUSION

In this study, we address the problem of destylization of decorative text and propose a novel framework for the text destylization. Our network combines residual block and PatchGAN. At the same time, we demonstrate the advantages of our model in the destylization of decorative text from three aspects, namely, comparative experiment, user study and performance evaluation. Finally, we show the results of the model for some more complex cases, and show more of the application of the model.

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China under no. 2020AAA0106200, and by National Natural Science Foundation of China under nos. 61832016, U20B2070 and 61672520.

REFERENCES

[1] S. Azadi, M. Fisher, V. Kim, Z. Wang, E. Shechtman, and T. Darrell. 2018. Multi-content GAN for Few-Shot Font Style Transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7564–7573.

[2] Alex J Champandard. 2016. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768* (2016).

[3] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. 2009. Sketch2photo: Internet image montage. *ACM Transactions on Graphics* 28, 5 (2009), 1–10.

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8789–8797.

[5] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. 2021. Arbitrary Video Style Transfer via Multi-Channel Correlation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*.

[6] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. 2020. Arbitrary Style Transfer via Multi-Adaptation Network. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA). Association for Computing Machinery, New York, NY, USA, 2719–2727.

[7] Lars Doyle, Forest Anderson, Ehren Choy, and David Mould. 2019. Automated pebble mosaic stylization of images. *Computational Visual Media* 5, 1 (2019), 33–44.

[8] Iddo Drori, Daniel Cohen-Or, and Hezy Yeshurun. 2003. Example-based style synthesis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. IEEE, II–143.

[9] Alexei A Efros and William T Freeman. 2001. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 341–346.

[10] Michael Elad and Peyman Milanfar. 2017. Style transfer via texture synthesis. *IEEE Transactions on Image Processing* 26, 5 (2017), 2338–2351.

[11] Oriel Frigo, Neus Sabater, Julie Delon, and Pierre Hellier. 2016. Split and match: Example-based adaptive patch sampling for unsupervised style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 553–561.

[12] L. A. Gatys, A. S. Ecker, and M. Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2414–2423.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*. 2672–2680.

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NIPS)*. 5767–5777.

[15] Hideaki Hayashi, Kohtaro Abe, and Seiichi Uchida. [n.d.]. GlyphGAN: Style-Consistent Font Generation Based on Generative Adversarial Networks. 186 ([n. d.]).

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[17] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. 2001. Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 327–340.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1125–1134.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation.

[20] Zhouhui Lian, Bo Zhao, Xudong Chen, and Jianguo Xiao. 2019. EasyFont: A Style Learning-Based System to Easily Build Your Large-Scale Handwriting Fonts. *ACM Transactions on Graphics* 38, 1 (2019), 6:1–6:18.

[21] Ming Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019. Few-Shot Unsupervised Image-to-Image Translation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 10550–10559.

[22] Jingwan Lu, Fisher Yu, Adam Finkelstein, and Stephen DiVerdi. 2012. HelpingHand: Example-Based Stroke Stylization. *ACM Transactions on Graphics* 31, 4, Article 46 (July 2012), 10 pages.

[23] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. 2018. Attentive generative adversarial network for raindrop removal from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2482–2491.

[24] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*.

[25] Shuai, Yang, Jiaying, Liu, Wenhan, Zongming, and Guo. 2018. Context-Aware Text-Based Binary Image Stylization and Synthesis. *IEEE Transactions on Image Processing* 28, 2 (Feb. 2018), 952–964.

[26] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).

[27] H. Wang, Y. Li, Y. Wang, H. Hu, and M. H. Yang. 2020. Collaborative Distillation for Ultra-Resolution Universal Style Transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1857–1866.

[28] Wenjing Wang, Jiaying Liu, Shuai Yang, and Zongming Guo. 2019. Typography with Decor: Intelligent text style transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5889–5897.

[29] Shuai Yang, Jiaying Liu, Zhouhui Lian, and Zongming Guo. 2017. Awesome Typography: Statistics-Based Text Effects Transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2886–2895.

[30] Shuai Yang, Jiaying Liu, Wenjing Wang, and Zongming Guo. 2019. TET-GAN: Text Effects Transfer via Stylization and Destylization. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*. 1238–1245.

[31] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. 2017. Deep joint rain detection and removal from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1357–1366.

[32] Richard Zhang, Phillip Isola, and Alexei A. Efros. 2016. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*. Springer, 649–666.

[33] Rui Zhang, Mingkun Yang, Xiang Bai, Baoguang Shi, and Minghui Liao. 2019. ICDAR 2019 Robust Reading Challenge on Reading Chinese Text on Signboard. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*.

[34] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. 2017. Real-Time User-Guided Image Colorization with Learned Deep Priors. *ACM Transactions on Graphics* 36, 4, Article 119 (July 2017), 11 pages.

[35] Y. Zhang, C. Fang, Y. Wang, Z. Wang, Z. Lin, Y. Fu, and J. Yang. 2019. Multimodal Style Transfer via Graph Cuts. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 5942–5950.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*. 2223–2232.

Multi-focus noisy image fusion based on gradient regularized convolutional sparse representatione

Authors:

Xuanjing Shen • Jilin University, Changchun, Jilin, China

Yunqi Zhang • Jilin University, Changchun, Jilin, China

Haipeng Chen • Jilin University, Changchun, Jilin, China

Di Gai • Jilin University, Changchun, Jilin, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446325

The method proposes a multi-focus noisy image fusion algorithm combining gradient regularized convolutional sparse representatione and spatial frequency. Firstly, the source image is decomposed into a base layer and a detail layer through two-scale image decomposition. The detail layer uses the Alternating Direction Method of Multipliers (ADMM) to solve the convolutional sparse coefficients with gradient penalties to complete the fusion of detail layer coefficients. Then, The base layer uses the spatial frequency to judge the focus area, the spatial frequency and the "choose-max" strategy are applied to achieved the multi-focus fusion result of base layer. Finally, the fused image is calculated as a superposition of the base layer and the detail layer. Experimental results show that compared with other algorithms, this algorithm provides excellent subjective visual perception and objective evaluation metrics.

Determining image age with rank-consistent ordinal classification and object-centered ensemble

Authors:

Shota Ashida • Kyoto University, Kyoto, Japan

Adam Jatowt • Kyoto University, Kyoto, Japan

Antoine Doucet • University of La Rochelle, La Rochelle, France

Masatoshi Yoshikawa • Kyoto University, Kyoto, Japan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446326

A significant number of old photographs including ones that are posted online do not contain the information of the date at which they were taken, or this information needs to be verified. Many of such pictures are either scanned analog photographs or photographs taken using a digital camera with incorrect settings. Estimating the date of such pictures is useful for enhancing data quality and its consistency, improving information retrieval and for other related applications. In this study, we propose a novel approach for automatic estimation of the shooting dates of photographs based on a rank-consistent ordinal classification method for neural networks. We also introduce an ensemble approach that involves object segmentation. We conclude that assuring the rank consistency in the ordinal classification as well as combining models trained on segmented objects improve the results of the age determination task.

A treatment engine by multimodal EMR data

Authors:

Zhaomeng Huang • Nanjing University of Aeronautics & Astronautics, Nanjing, Jiangsu, China

Liyan Zhang • Nanjing University of Aeronautics & Astronautics, Nanjing, Jiangsu, China

Xu Xu • Nanjing University of Aeronautics & Astronautics, Nanjing, Jiangsu, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446254

In recent years, with the development of electronic medical record (EMR) systems, it has become possible to mine patient clinical data to improve medical care quality. After the treatment engine learns knowledge from the EMR data, it can automatically recommend the next stage of prescriptions and provide treatment guidelines for doctors and patients. However, this task is always challenged by the multi-modality of EMR data. To more effectively predict the next stage of treatment prescription by using multimodal information and the connection between the modalities, we propose a cross-modal shared-specific feature complementary generation and attention fusion algorithm. In the feature extraction stage, specific information and shared information are obtained through a shared-specific feature extraction network. To obtain the correlation between the modalities, we propose a sorting network. We use the attention fusion network in the multimodal feature fusion stage to give different multimodal features at different stages with different weights to obtain a more prepared patient representation. Considering the redundant information of specific modal information and shared modal information, we introduce a complementary feature learning strategy, including modality adaptation for shared features, project adversarial learning for specific features, and reconstruction enhancement. The experimental results on the real EMR data set MIMIC-III prove its superiority and each part's effectiveness.

Storyboard relational model for group activity recognition

Authors:

Boning Li • Nanjing University of Science and Technology

Xiangbo Shu • Nanjing University of Science and Technology

Rui Yan • Nanjing University of Science and Technology

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446255

This work concerns how to effectively recognize the group activity performed by multiple persons collectively. As known, Storyboards (i.e., medium shot, close shot) jointly describe the whole storyline of a movie in a compact way. Likewise, the actors in small subgroups (similar to Storyboards) of a group activity scene contribute a lot to such group activity and develop more compact relationships among them within subgroups. Inspired by this, we propose a Storyboard Relational Model (SRM) to address the problem of Group Activity Recognition by splitting and reintegrating the group activity based on the small yet compact Storyboards. SRM mainly consists of a Pose-Guided Pruning (PGP) module and a Dual Graph Convolutional Networks (Dual-GCN) module. Specifically, PGP is designed to refine a series of Storyboards from the group activity scene by leveraging the attention ranges of individuals. Dual-GCN models the compact relationships among actors in a Storyboard. Experimental results on two widely-used datasets illustrate the effectiveness of the proposed SRM compared with the state-of-the-art methods.

Distilling knowledge in causal inference for unbiased visual question answering

Authors:

Yonghua Pan • Nanjing University of Science & Technology, Nanjing, China

Zechao Li • Nanjing University of Science & Technology, Nanjing, China

Liyan Zhang • Nanjing University Of Aeronautics & Astronautics, Nanjing, China

Jinhui Tang • Nanjing University of Science & Technology, Nanjing, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446256

Current Visual Question Answering (VQA) models mainly explore the statistical correlations between answers and questions, which fail to capture the relationship between the visual information and answers. The performance dramatically decreases when the distribution of handled data is different from the training data. Towards this end, this paper proposes a novel unbiased VQA model by exploring the Casual Inference with Knowledge Distillation (CIKD) to reduce the influence of bias. Specifically, the causal graph is first constructed to explore the counterfactual causality and infer the casual target based on the causal effect, which well reduces the bias from questions and obtain answers without training. Then knowledge distillation is leveraged to transfer the knowledge of the inferred casual target to the conventional VQA model. It makes the proposed method enable to handle both the biased data and standard data. To address the problem of the bad bias from the knowledge distillation, the ensemble learning is introduced based on the hypothetical bias reason. Experiments are conducted to show the performance of the proposed method. The significant improvements over the state-of-the-art methods on the VQA-CP v2 dataset well validate the contributions of this work.

Incremental multi-view object detection from a moving camera

Authors:

Takashi Konno • AIST, Japan

Ayako Amma • Toyota Motor Corporation, Japan

Asako Kanezaki • AIST, Japan

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446257

Object detection in a single image is a challenging problem due to clutters, occlusions, and a large variety of viewing locations. This task can benefit from integrating multi-frame information captured by a moving camera. In this paper, we propose a method to increment object detection scores extracted from multiple frames captured from different viewpoints. For each frame, we run an efficient end-to-end object detector that outputs object bounding boxes, each of which is associated with the scores of categories and poses. The scores of detected objects are then stored in grid locations in 3D space. After observing multiple frames, the object scores stored in each grid location are integrated based on the best object pose hypothesis. This strategy requires the consistency of object categories and poses among multiple frames, and thus it significantly suppresses miss detections. The performance of the proposed method is evaluated on our newly created multi-class object dataset captured in robot simulation and real environments, as well as on a public benchmark dataset.

An automated method with anchor-free detection and U-shaped segmentation for nuclei instance segmentation

Authors:

Xuan Feng • Beijing University of Technology, Beijing, China

Lijuan Duan • Beijing University of Technology, Beijing, China

Jie Chen • Peng Cheng Laboratory, Shenzhen, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446258

Nuclei segmentation plays an important role in cancer diagnosis. Automated methods for digital pathology become popular due to the developments of deep learning and neural networks. However, this task still faces challenges. Most of current techniques cannot be applied directly because of the clustered state and the large number of nuclei in images. Moreover, anchor-based methods for object detection lead a huge amount of calculation, which is even worse on pathological images with a large target density. To address these issues, we propose a novel network with an anchor-free detection and a U-shaped segmentation. An altered feature enhancement module is attached to improve the performance in dense target detection. Meanwhile, the U-Shaped structure in segmentation block ensures the aggregation of features in different dimensions generated from the backbone network. We evaluate our work on a Multi-Organ Nuclei Segmentation dataset from MICCAI 2018 challenge. In comparisons with others, our proposed method achieves state-of-the-art performance.

Improving face recognition in surveillance video with judicious selection and fusion of representative frames

Authors:

Zhaozhen Ding • ZTE Corporation, Nanjing, China

Qingfang Zheng • ZTE Corporation, Nanjing, China

Chunhua Hou • ZTE Corporation, Nanjing, China

Guang Shen • ZTE Corporation, Nanjing, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446259

Face recognition in unconstrained surveillance videos is challenging due to the different acquisition settings and face variations. We propose to utilize the complementary correlation between multi-frames to improve face recognition performance. We design an algorithm to build a representative frame set from the video sequence, selecting faces with high quality and large appearance diversity. We also devise a refined Deep Residual Equivariant Mapping (DREAM) block to improve the discriminative power of the extracted deep features. Extensive experiments on two relevant face recognition benchmarks, YouTube Face and IJB-A, show the effectiveness of the proposed method. Our work is also lightweight, and can be easily embedded into existing CNN based face recognition systems.

Two-stage structure aware image inpainting based on generative adversarial networks

Authors:

Jin Wang • Beijing University of Technology

Xi Zhang • Beijing University of Technology

Chen Wang • Beijing University of Technology

Qing Zhu • Beijing University of Technology

Baocai Yin • Beijing University of Technology

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446260

In recent years, the image inpainting technology based on deep learning has made remarkable progress, which can better complete the complex image inpainting task compared with traditional methods. However, most of the existing methods can not generate reasonable structure and fine texture details at the same time. To solve this problem, in this paper we propose a two-stage image inpainting method with structure awareness based on Generative Adversarial Networks, which divides the inpainting process into two sub tasks, namely, image structure generation and image content generation. In the former stage, the network generates the structural information of the missing area; while in the latter stage, the network uses this structural information as a prior, and combines the existing texture and color information to complete the image. Extensive experiments are conducted to evaluate the performance of our proposed method on Places2, CelebA and Paris Streetview datasets. The experimental results show the superior performance of the proposed method compared with other state-of-the-art methods qualitatively and quantitatively.

Low-quality watermarked face inpainting with discriminative residual learning

Authors:

Zheng He • Wuhan University, Wuhan, China

Xueli Wei • Wuhan University, Wuhan, China

Kangli Zeng • Wuhan University, Wuhan, China

Zhen Han • Wuhan University, Wuhan, China

Qin Zou • Wuhan University, Wuhan, China

Zhongyuan Wang • Wuhan University, Wuhan, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

 • doi > 10.1145/3444685.3446261

Most existing image inpainting methods assume that the location of the repair area (watermark) is known, but this assumption does not always hold. In addition, the actual watermarked face is in a compressed low-quality form, which is very disadvantageous to the repair due to compression distortion effects. To address these issues, this paper proposes a low-quality watermarked face inpainting method based on joint residual learning with cooperative discriminant network. We first employ residual learning based global inpainting and facial features based local inpainting to render clean and clear faces under unknown watermark positions. Because the repair process may distort the genuine face, we further propose a discriminative constraint network to maintain the fidelity of repaired faces. Experimentally, the average PSNR of inpainted face images is increased by 4.16dB, and the average SSIM is increased by 0.08. TPR is improved by 16.96% when FPR is 10% in face verification.

A multimedia solution to motivate childhood cancer patients to keep up with cancer treatment

Authors:

Carmen Wang Er Chai • Swinburne University of Technology Sarawak Campus, Kuching Sarawak Malaysia

Bee Theng Lau • Swinburne University of Technology Sarawak Campus, Kuching Sarawak Malaysia

Abdullah Al Mahmud • Swinburne University of Technology, Melbourne Victoria, Australia

Mark Kit Tsun Tee • Swinburne University of Technology Sarawak Campus, Kuching Sarawak Malaysia

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446262

Childhood cancer is a deadly illness that requires the young patient to adhere to cancer treatment for survival. Sadly, the high treatment side-effect burden can make it difficult for patients to keep up with their treatment. However, childhood cancer patients can manage these treatment side effects through daily self-care to make the process more bearable. This paper outlines the design and development process of a multimedia-based solution to motivate these young patients to adhere to cancer treatment and manage their treatment side effects. Due to the high appeal of multimedia-based interventions and the proficiency of young children in using mobile devices, the intervention of this study takes the form of a virtual pet serious game developed for mobile. The intervention which is developed based on the Protection Motivation Theory, includes multiple game modules with the purpose of improving the coping appraisal of childhood cancer patients on using cancer treatment to fight cancer, and taking daily self-care to combat treatment side-effects. The prototype testing results show that the intervention is well received by the voluntary play testers. Future work of this study includes the evaluation of the intervention developed with childhood cancer patients to determine its effectiveness.

Global and local feature alignment for video object detection

Authors:

Haihui Ye • Xiamen Universit, Xiamen, China

Qiang Qi • Xiamen Universit, Xiamen, China

Ying Wang • Xiamen Universit, Xiamen, China

Yang Lu • Xiamen Universit, Xiamen, China

Hanzi Wang • Xiamen Universit, Xiamen, China

Publication:

Proceeding

MMAsia '20 Proceedings of the 2nd ACM International Conference on Multimedia in Asia

ISBN: 9781450383080

©2021 • doi > 10.1145/3444685.3446263

Extending image-based object detectors into video domain suffers from immense inadaptability due to the deteriorated frames caused by motion blur, partial occlusion or strange poses. Therefore, the generated features of deteriorated frames encounter the poor quality of misalignment, which degrades the overall performance of video object detectors. How to capture valuable information locally or globally is of importance to feature alignment but remains quite challenging. In this paper, we propose a Global and Local Feature Alignment (abbreviated as GLFA) module for video object detection, which can distill both global and local information to excavate the deep relationship between features for feature alignment. Specifically, GLFA can model the spatial-temporal dependencies over frames based on propagating global information and capture the interactive correspondences within the same frame based on aggregating valuable local information. Moreover, we further introduce a Self-Adaptive Calibration (SAC) module to strengthen the semantic representation of features and distill valuable local information in a dual local-alignment manner. Experimental results on the ImageNet VID dataset show that the proposed method achieves high performance as well as a good trade-off between real-time speed and competitive accuracy.