Guest Editorial
Introduction to the Special Section on Contextual Object Analysis in Complex Scenes

IN RECENT years, with the vast development of deep learning techniques, a great deal of effort has been devoted in the computer vision and multimedia community toward the problems of visual object analysis, such as object representation, recognition, detection, identification, etc. Especially at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014, the computers have successfully outperformed humans with a lower error rate at image recognition for the first time. However, most of existing algorithms focus more on analyzing objects in a relatively simple and restricted situation, which may perform poorly in natural environments.

More recently, with the widespread availability of digital cameras and a growing need for artificial intelligence applications, e.g., visual surveillance and autonomous driving, the ability of analyzing visual objects in complex scenes becomes critically important in the intelligent systems. It has drawn increasing research attention from multiple communities. Compared with object analysis in simple scenes, it can perceive the visual contexts effectively for complex scene understanding. Despite the promising progress in this direction, many fundamental problems are still not well solved so far, e.g., how to detect moving objects in urban traffic scenes or adverse weather conditions, how to count all the certain objects in crowded scenes, how to re-identify pedestrians across camera views, and how to recognize or describe the objects in a fine-grained manner.

As Guest Editors of this Special Section on Contextual Object Analysis in Complex Scenes, we are happy to present 17 accepted articles that represent the most recent research developments in this field, including novel techniques, frameworks, models, datasets, and solutions. We would like to thank the authors of the 17 accepted articles for contributing wonderful work to this section. We would also like to thank the Editor-in-Chief (EIC), Feng Wu, for making this Special Section possible. The 17 accepted articles can be roughly categorized into four groups: 1) object detection; 2) object recognition and captioning; 3) object re-identification; and 4) crowd object counting.

A. Object Detection

This group includes six articles aiming at developing effective object detection methods in complex scenes.

The article "Smoke vehicle detection based on spatiotemporal bag-of-features and professional convolutional neural network" by H. Tao and X. Lu proposes automatic smoke vehicle detection methods based on manually crafted spatiotemporal bag-of-features (S-BoF) and deeply learned spatiotemporal features. The S-BoF contains multiple groups of features: color moments on three orthogonal planes, completed robust local binary pattern on three orthogonal planes, and histogram of oriented gradient on three orthogonal planes, which are fed into the SVM for detecting smoke vehicles based on a score-level fusion. The deeply learned spatiotemporal features are extracted from three convolutional neural networks (CNNs) trained on the color, texture, and gradient feature channels, separately. The authors also collect five new datasets from the traffic surveillance scenes that clearly indicate the effectiveness of the proposed methods.

The work "Salient features for moving object detection in adverse weather conditions during night time" by A. Singha and M. K. Bhowmik investigates the foreground segmentation of moving objects in adverse atmospheric conditions. It proposes an improved background model that utilizes both thermal pixel intensity features and spatial video salient features. The spatial video salient features are represented as an Akin-based per-pixel Boolean string over a local region block. The background model is controlled via the automatic adaptation of parameters, e.g., decision threshold and learning parameters, and the updating of background samples. Experiments on a newly collected weather-degraded video dataset and a public dataset demonstrate the effectiveness of the improved background model.

The article "S^3D: Scalable pedestrian detection via score scale surface discrimination" by X. Wang *et al.* investigates the pedestrian detection at various scales in visual surveillance scenes. It presents a score scale surface discrimination method to distinguish the pedestrians at various scales according to their locations on the discriminant surface. Experiments on four pedestrian detection datasets show that the S^3D is robust to scale changes and outperforms the state-of-the-arts.

The work "Three-dimensional point cloud object detection using scene appearance consistency among multi-view projection directions" by D. Sugimura *et al.* studies the three-dimensional (3D) object detection in point clouds. It treats the 3D object detection problem as the determination of optimal correspondence among image sets and uses the principal component analysis to estimate effective image-projection directions for object point clouds. Experiments on public

Digital Object Identifier 10.1109/TCSVT.2020.3017118

datasets demonstrate the effectiveness of the simultaneous correspondence of image sets. It can obtain reliable candidates that belong to the target object region.

The article "Aggregating attentional dilated features for salient object detection" by L. Zhu *et al.* presents a novel deep learning model to aggregate the attentional dilated features for salient object detection, which explores the complementary information between the global and local contexts in a convolutional neural network. It develops an attentional dense atrous (dilated) spatial pyramid pooling module to better utilize the local and global saliency cues and an aggregation network to integrate the refined features by formulating two consecutive chains of residual learning-based modules: one chain is from deep to shallow layers while another chain is from shallow to deep layers. Evaluations on seven widely used benchmarks validate the effectiveness of the attention module and the aggregation network.

The work "High-level semantic networks for multi-scale object detection" by J. Cao *et al.* studies the multi-scale object detection. It proposes a multi-branch and high-level semantic network by gradually splitting a base network into multiple different branches. Due to the difference of receptive fields, the different branches can detect objects at different scales. The authors also use skip-layer connections to add context to the branch of the relatively small receptive field and dilated convolutions to enlarge the resolutions of output feature maps. The authors demonstrate its effectiveness by evaluations on three pedestrian detection datasets, one face detection dataset, and two general object detection datasets.

B. Object Recognition and Captioning

This group includes three articles that develop effective methods for occluded object recognition, person attribute recognition, or novel object captioning.

The article "Occluded face recognition in the wild by identity-diversity inpainting" by S. Ge *et al.* proposes an identity-diversity inpainting method to recognize occluded faces. It integrates a CNN-based face recognizer into a generative adversarial network (GAN) to build an identity-diversity GAN (ID-GAN) that reconstructs visually plausible occlusions by face inpainting. It designs an identity-diversity loss with the supervision of a collection of identity-centered features to improve the discriminative capacity of the generated faces. The experimental results show that the proposed method can generate photorealistic results and improve the accuracy of occluded face recognition.

The work "Person attribute recognition by sequence contextual relation learning" by J. Wu *et al.* aims to identify the attribute labels from the pedestrian images. It proposes a sequence contextual relation learning (SCRL) method to capture the contextual relations from images and attributes. Specifically, it first encodes the person image and its affiliated attributes as an image patch sequence and an attribute sequence, respectively, and then learns the contextual relations by exploring intra-attention and inter-attention mechanisms on the sequences. The intra-attention mechanism captures spatial context from the image patch sequence and semantic correlation from the attribute sequence. The inter-attention mechanism captures the spatial-semantic relations across the two sequences. Extensive experiments on five datasets show that the SCRL can effectively improve the performance of person attribute recognition.

The article "Cascaded revision network for novel object captioning" by Q. Feng *et al.* aims to describe novel objects in images with a descriptive sentence. It presents a cascaded revision network (CRN) that aims to inject out-of-domain knowledge into the image-captioning model for captioning novel objects. Specifically, it utilizes the external knowledge from a pretrained object detection model to revise the primary caption by visual matching, followed by a semantic matching module to embed the out-of-domain object into the caption without breaking the grammar. Experiments on two public datasets validate that the CRN can accurately describe unseen objects in images.

C. Object Re-Identification

This group includes five articles aiming at developing effective methods for person re-identification. This task aims to match a specific individual across non-overlapping cameras, which is an important but challenging task in multi-camera video surveillance scenes.

The article "SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification" by K. Kansal *et al.* tackles the task of visible-infrared person re-identification, which is critically important for surveillance applications under poor illumination conditions. This work designs a non-adversarial and fast disentanglement method to disentangle the spectrum information while learning the identity discriminative features. To extract these features, it proposes a novel network with disentanglement loss that can distill identity features and dispel spectrum features. The entire network is trained in an end-to-end manner by minimizing spectrum information and maximizing invariant identity relevant information. Experiments on public datasets show the efficacy of the proposed method.

The work "Complementation-reinforced attention network for person re-identification" by C. Han *et al.* investigates the multi-head attention mechanism in person re-identification by proposing a complementation-reinforced attention network (CRAN). The CRAN can effectively reduce the redundancy among different attention heads by imposing complementing constraints among multiple attention heads. Specifically, it encourages each branch to attend to complementary attention regions and enforces orthogonality among the learned features of different regions. The authors also simplify the non-local block by sparsifying the query positions and explore the influence of query usage on network optimization. An adaptive fusion module is introduced to integrate the multi-branch features. The experimental results on three benchmark datasets demonstrate that the CRAN can outperform the state-of-the-arts.

The article "Bayesian inferred self-attentive aggregation for multi-shot person re-identification" by X. Liu *et al.* aims to alleviate the annotation inconsistency in current multi-shot person re-identification datasets. Specifically, it first formulates the multi-shot person re-identification as a multi-instance learning problem and then proposes a Bayesian inferred self-attentive aggregation model to investigate the importance of the instances and regions, thereby enhancing the discriminative ability of set-level representations for person re-identification. It also designs a collective aggregation function by adjusting the activation threshold, which makes the model robust to outliers. The authors validate the efficacy of the proposed method by conducting extensive experiments on four benchmark datasets.

The work "Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification" by Y. Huang *et al.* studies the problem of long-term person re-identification. To facilitate this task, this work first collects a large-scale re-identification dataset called "Celeb-reID" with more than 1000 person IDs and 34 186 captured images. Different from existing datasets, the same person in Celeb-reID is allowed to change clothes across different camera views, which is more challenging. This article presents a Re-IDentification Capsule network to deal with the appearance change in this task and applies soft embedding attention and feature sparse representation mechanisms to enhance the discrimination and generalization capacity of the network. Experiments show that the proposed method outperforms the state-of-the-arts by a large margin in the long-term re-identification scene.

The article "Attribute-identity embedding and self-supervised learning for scalable person re-identification" by H. Li *et al.* proposes a self-supervised learning algorithm for person re-identification, which can incrementally optimize the model by selecting unlabeled samples from the target domain based on attribute-identity embedding. Specifically, it first develops an attribute-identity joint prediction dictionary learning model for simultaneously learning a latent attribute space, a semantic attribute dictionary, and an identifier. Then, it proposes a prediction-training cycle self-supervised learning algorithm to optimize the model. Experiments show that the proposed method not only outperforms the state-of-the-art unsupervised methods but also some supervised methods.

D. Crowd Object Counting

This group includes three articles aiming at developing effective methods for estimating the number of certain objects in crowded scenes. It is a challenging task due to high appearance similarity, perspective changes, and severe congestion.

The article "PCC Net: Perspective crowd counting via spatial convolutional network" by J. Gao *et al.* proposes a perspective crowd counting network (PCC Net). The PCC Net mainly consists of three parts: 1) density map estimation to extract local features; 2) random high-level density classification to extract global features for predicting the coarse density labels of random patches in images; 3) fore-/background segmentation to encode mid-level features. It also designs a perspective module to encode the perspective changes in four directions. The proposed method can effectively encode the perspective changes and accurately estimate the density. Its efficacy is validated on four public crowd object counting datasets.

The work "ZoomCount: A zooming mechanism for crowd counting in static images" by U. Sajid *et al.* proposes a novel zoom-in and zoom-out mechanism for accurate crowd object counting in highly diverse images. Specifically, it mainly consists of three components: a crowd density classifier (CDC), a novel decision module (DM), and a count regressor module (CRM). It first divides the input image into patches and then feeds patches into CDC to classify the patches into four density categories: low, medium, high-density, and no-crowd. Then, it feeds the number of patches on the four categories into the DM to determine whether the image belongs to an extreme case (very low or high density) or normal case, followed by a corresponding zooming mechanism in CRM for crowd counting. The authors also create four benchmark datasets for evaluation. The experimental results demonstrate that the zooming mechanism is effective in the extreme crowed scenes.

The article "Counting objects by blockwise classification" by L. Liu *et al.* presents a blockwise count-level classification method to address two issues in most existing crowd counting methods: inaccurately generated regression targets and serious sample imbalance. Specifically, it classifies the count levels of each block produced by nonlinearly quantizing the continuous counts, thus transforming the imbalance of sample patch counts to a class imbalance of count levels, followed by an information-entropy-inspired loss for optimization. The authors demonstrate its effectiveness by extensive evaluations on six public datasets and one newly collected sonar fish-counting dataset.

This Special Section summarizes the recent efforts in the field of contextual object analysis in complex scenes and presents interesting works that push the state-of-the-art in the field, opening new lines of research and areas of applications. But, we foresee many new contributions in developing more robust object analysis methods in complex scenes to meet real-world requirements. We believe this special Section will offer a timely collection of work to benefit the researchers in the research fields of computer vision, multimedia, and machine learning.

MENG WANG, *Professor*
School of Computer Science and
Information Engineering
Hefei University of Technology
Hefei 230009, China
e-mail: eric.mengwang@gmail.com

XIANGLONG LIU, *Associate Professor*
State Key Laboratory of
Software Development Environment
School of Computer Science and Engineering
Beihang University
Beijing 100083, China
e-mail: xlliu@buaa.edu.cn

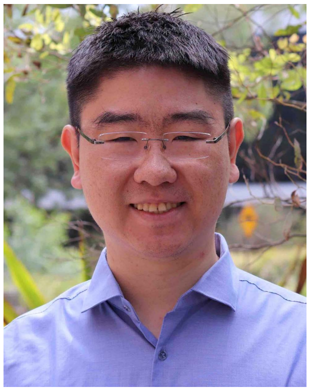
Xun Yang, *Research Fellow*
Department of Computer Science
School of Computing
National University of Singapore
Singapore 119077
e-mail: xunyang@nus.edu.sg

Liang Zheng, *Lecturer*
Research School of Computer Science
The Australian National University
Canberra, ACT 0200, Australia
e-mail: liang.zheng@anu.edu.au

Meng Wang received the B.E. and Ph.D. degrees from the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored more than 200 book chapters and journal articles and conference papers in these areas. He was a recipient of the ACM SIGMM Rising Star Award 2014. He is an Associate Editor of the IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), the IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT), the IEEE Transactions on Multimedia (IEEE TMM), and the IEEE Transactions on Neural Networks and Learning Systems (IEEE TNNLS).

Xianglong Liu received the B.S. and Ph.D. degrees in computer science from Beihang University, Beijing, China, in 2008 and 2014, respectively. From 2011 to 2012, he visited the Digital Video and Multimedia (DVMM) Laboratory, Columbia University, as a joint Ph.D. Student. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. He has published over 60 research articles at top venues, such as the IEEE Transactions on Image Processing, the Conference on Computer Vision and Pattern Recognition, the International Conference on Computer Vision, and the Association for the Advancement of Artificial Intelligence. His research interests include machine learning, computer vision, and multimedia information retrieval.

Xun Yang received the Ph.D. degree from the Hefei University of Technology, China, in 2017. He is currently a Post-Doctoral Research Fellow with the NExT++ Research Center, National University of Singapore. His current research interests include information retrieval, multimedia content analysis, and computer vision. He has served as a PC Member and an Invited Reviewer for top-tier conferences and prestigious journals, including ACM MM, IJCAI, AAAI, the *ACM Transactions on Multimedia Computing, Communications, and Applications*, the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Knowledge and Data Engineering, and the IEEE Transactions on Circuits and Systems for Video Technology.



Liang Zheng received the B.S. degree in life science and the Ph.D. degree in electronic engineering from Tsinghua University, China, in 2010 and 2015, respectively. He is currently a Lecturer and an ARC DECRA Fellow with the Research School of Computer Science, The Australian National University. His research interests include object matching and re-identification, deep learning, and data synthesis.