Guest Editorial: Introduction to the Special Section on Fine-Grained Visual Categorization

Jingdong Wang, Zhuowen Tu, *Fellow, IEEE*, Jianlong Fu, Nicu Sebe, and Serge Belongie

1 Introduction

With the techniques for standard supervised image classification becoming increasingly practical, fine-grained visual categorization (FGVC) has recently attracted a lot of attention and has emerged as an important task in computer vision. The task of FGVC aims to classify an image into subordinate categories. Examples of FGVC include but are not limited to recognizing, e.g., animal species, sub-groups of plants, and car makes and models. Fine-grained categorization is different from general-purpose image categorization tasks, such as the ImageNet Challenge of 1K general categories. FGVC pays much attention to subtle details that are not easily captured using the off-the-shelf image classifiers. FGVC is a promising direction in visual perception and image understanding beyond generic labels. In addition, the absence of sufficient training data with the presence of a large number of fine-grained categories, e.g., about 10K species for birds and over 250K species for flowers, makes the problem of FGVC particularly challenging yet worthwhile.

Some existing research works follow the state-of-the-art general-purpose image classification approaches, which directly apply deep neural networks to FGVC tasks. However, the fine-grained categorization problem is not easily solved merely by training modern deep convolutional neural networks. In the past, results for fine-grained image categorization have been mostly obtained by using classifiers with strong supervision, where detailed labels such as body parts, attributes, and viewpoints were manually annotated and used in training. Many questions generally arise when the fine-grained categorization task is adopted in more general and broad applications: How do we alleviate the burden of obtaining fine-grained manual annotations? How can top-down information and domain knowledge be included to assist FGVC prediction? How can we make the best use of web data and online resources like Mechanical Turk to improve training FGVC models?

- *Jingdong Wang and Jianlong Fu are with the Microsoft Research Asia, Beijing 100080, China. E-mail: {jingdw, jianf}@microsoft.com.*
- *Zhuowen Tu is with the Department of Cognitive Science, University of California at San Diego, San Diego, CA 92093 USA. E-mail: ztu@ucsd.edu.*
- *Nicu Sebe is with the Department of Information Engineering and Computer Science (DISI), University of Trento, 38123 Trento, Italy. E-mail: niculae.sebe@unitn.it.*
- *Serge Belongie is with the Department of Computer Science, Cornell University, Ithaca, NY 14850 USA. E-mail: sjb344@cornell.edu.*

Digital Object Identifier no. 10.1109/TPAMI.2021.3065094

This special section on fine-grained visual categorization has attracted many research works on fine-grain related topics. We thank all authors for submitting their papers to the special section and all reviewers who have provided professional, insightful, and timely reviews, leading to the high quality of accepted papers. We also thank *TPAMI* EIC Sven Dickinson and the Associate EICs for recognizing the widespread interest in this field, which warrants this special section. The accepted papers are divided into four groups based on their different focuses:

- Fine-grained image recognition,
- Fine-grained human analysis,
- Fine-grained video action recognition, and
- Fine-grained vision-language reasoning.

In the following section, we will review the accepted papers in each of the groups.

2 Fine-Grained Visual Categorization

2.1 Fine-Grained Image Recognition

The paper "Hierarchical Deep Click Feature Prediction for Fine-Grained Image Recognition" by Jun Yu, Min Tan, Hongyuan Zhang, Yong Rui, and Dacheng Tao explores how to effectively utilize the user-click-frequency feature of an image for fine-grained image recognition. In particular, they design a Hierarchical Deep Word Embedding (HDWE) model to capture hierarchical semantics. A feature selection module is introduced for the prediction of click features by integrating a sparse constraint and improved RELU operator. Experiments on dog and bird breed datasets show the effectiveness of the proposed approach and its one-shot learning ability and scalability to unseen categories.

Junwei Han, Xiwen Yao, Gong Cheng, Xiaoyu Feng, and Dong Xu propose a unified fine-grained visual categorization framework which improves traditional part-based image representations. The title of their paper is "P-CNN: Part-Based Convolutional Neural Networks for Fine-Grained Visual Categorization." The framework consists of three closely-realted modules: (1) A Squeeze-and-Excitation (SE) block that learns to recalibrate channel-wise feature responses by emphasizing informative channels. (2) A Part Localization Network (PLN) trained in an unsupervised fashion to locate distinctive object parts. (3) A Part Classification Network (PCN) that classifies each individual part into image-level categories and combines part-level local features and object-level global features for the final

classification. Experiments on three widely-used FGVC datasets demonstrate the superiority of the P-CNN model.

Several options for normalizing second-order features are studied in the paper "Power Normalizations in Fine-Grained Image, Few-Shot Image and Graph Classification" by Piotr Koniusz and Hongguang Zhang. They also establish theoretical foundations for second-order image feature designing from a probabilistic viewpoint. They first investigate the role of two power normalizations, namely MaxExp and Gamma, and then propose surrogate functions SigmE and AsinhE for end-to-end training to handle the so-called negative evidence. They further investigate spectral power normalizations and demonstrate their close connection to the time-reversed heat diffusion process on graph Laplacians, which explains that spectral MaxExp and Gamma reduce diffusion between features. They propose a fast spectral MaxExp which rivals the matrix square root approximation via Newton-Schulz iterations. Evaluations on standard image, graph classification and few-shot learning problems including fine-grained datasets demonstrate the effectiveness of the proposed methods.

2.2 Fine-Grained Human Analysis

A fine-grained human-centric tracklet segmentation problem is proposed and addressed in the paper "Fine-Grained Human-Centric Tracklet Segmentation with Single Frame Supervision" by Si Liu, Guanghui Ren, Yao Sun, Jinqiao Wang, Changhu Wang, Bo Li, and Shuicheng Yan. It targets to parse a human body in videos using the model that requires only one labeled frame per video in training stage. They propose a Temporal context Segmentation Network (TSN) to explore the pixel- and frame-level context in videos to use both unlabeled and labeled frames in an end-to-end fashion. They also release a new annotated dataset with four scenarios, including Indoor, Outdoor, iLIDS-Parsing, and Daily.

The paper "Pose-Guided Representation Learning for Person Re-Identification" by Jianing Li, Shiliang Zhang, Qi Tian, and Wen Gao aims to learn an efficient global representation in person re-identification task. In particular, they leverage the human pose and part cues to learn a Part-Guided Representation (PGR) consisting of Pose Invariant Feature (PIF) and Local Descriptive Feature (LDF) to deal with pose variations and misalignment errors, respectively. Extensive experiments on five benchmark datasets show the effectiveness of PGR over current state-of-the-art methods, considering both ReID accuracy and feature extraction efficiency.

2.3 Fine-Grained Video Action Recognition

Group activity recognition problem is a fine-grained and more challenging recognition task compared with traditional single-person action recognition and two persons' interaction recognition tasks. This problem is discussed and addressed in the paper "Coherence Constrained Graph LSTM for Group Activity Recognition" by Jinhui Tang, Xiangbo Shu, Rui Yan, and Liyan Zhang. They propose a Spatio-Temporal Context Coherence (STCC) constraint and a Global Context Coherence (GCC) constraint to capture the relevant motions and quantify their contributions to group activities. A Coherence Constrained Graph LSTM (CCG-LSTM) is proposed to model the relevant motions of individuals and further to recognize group activities. Experimental results on two widely-used datasets demonstrate the effectiveness of the proposed method.

Piotr Koniusz, Lei Wang, and Anoop Cherian present two tensor-based feature representations to capture higher-order relationships between visual features for the task of action recognition. Their paper title is "Tensor Representations for Action Recognition." This paper provides a novel view for solving action recognition tasks. In particular, sequences and the dynamics compatibility kernels are introduced to capture spatio-temporal evolution of body-joints for 3D skeleton based action sequences. Besides, this paper is the first to conduct a theoretical analysis of higher-order pooling with Tensor Power Normalization. Moreover, Tensor Power-Euclidean metric performs spectral detection of features falling into subspaces, which is used for generic/fine-grained action recognition. Extensive experiments on six regular/fine-grained datasets for skeleton- or video-based inputs show the effectiveness of the tensor-based feature representation.

2.4 Fine-Grained Vision-Language Reasoning

The paper "Fine-Grained Video Captioning via Graph-based Multi-Granularity Interaction Learning" by Yichao Yan, Ning Zhuang, Bingbing Ni, Jian Zhang, Minghao Xu, Qiang Zhang, Zheng Zhang, Shuo Cheng, Qi Tian, Yi Xu, Xiaokang Yang, and Wenjun Zhang propose a new fine-grained video captioning topic to generate continuous linguistic descriptions for multi-subject interactive videos. To tackle this new problem and facilitate the research community, they also introduce a new dataset called Fine-grained Sports Video Narrative dataset (SVN) and a novel performance evaluation metric named Fine-grained Captioning Evaluation (FCE). They also propose a framework named Graph-based Learning for Multi-Granularity Interaction Representation (GLMGIR) to generate action/event description of multiple spatio-temporal resolutions in a progressive way.

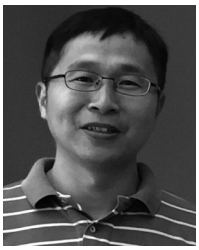
The task of natural language grounding in images requires to understand the fine-grained and compositional language space. The paper "Learning to Compose and Reason with Language Tree Structures for Visual Grounding" by Richang Hong, Hanwang Zhang, Daqing Liu, Xiaoyu Mo, and Xiangnan proposes a Recursive Grounding Tree (RVG-TREE) model to automatically compose a binary tree structure for parsing the language and then perform visual reasoning along the tree in a bottom-up fashion. Experimental Results on three benchmark datasets show the effectiveness of their method.

The paper "Plenty Is Plague: Fine-Grained Learning for Visual Question Answering" by Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Deyu Meng, Yue Gao, and Chunhua Shen proposes a fine-grained learning paradigm for the task of Visual Question and Answering (VQA). Instead of using all training data from scratch, they use an actor-critic based learning agent that schedules the most difficult question types in each training epoch. Experimental results show that this method is able to reduce the training time with small amount of training data while maintain

comparable performance. This fine-grained learning paradigm can be easily embedded to existing VQA models.

A novel image captioning task is proposed to generate longer, richer and more fine-grained sentences and paragraphs for an image in the paper "Context-Aware Visual Policy Network for Fine-Grained Image Captioning" by Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. To better capture visual context that is crucial for compositional reasoning such as object relationships, they propose a Context-Aware Visual Policy network (CAVP) for both sentence- and paragraph-level image captioning. At each time step, CAVP explicitly utilizes the previous visual attentions as context and decides whether it is used for the current word generation. Experimental results on two benchmark datasets demonstrate the effectiveness of CAVP in both quantitative and qualitative evaluations.

Jingdong Wang,
Zhuowen Tu,
Jianlong Fu,
Nicu Sebe, and
Serge Belongie
Guest Editors



Jingdong Wang received the BEng and MEng degrees from the Department of Automation, Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the PhD degree from the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2007. He is currently a senior principal research manager at the Visual Computing Group, Microsoft Research, Beijing, China. His research interests include neural network design, human pose estimation, semantic segmentation, large-scale indexing, and person re-identification. He is an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Circuits and Systems for Video Technology*, and is also an area chair of several leading computer vision and AI conferences, such as CVPR, ICCV, ECCV, ACM MM, IJCAI, and AAAI. He is an IAPR fellow and an ACM distinguished member.

Zhuowen Tu (Fellow, IEEE) received the PhD degree from Ohio State University, Columbus, Ohio, and the ME degree from Tsinghua University, China. He is currently a full professor of cognitive science and also affiliated with the Department of Computer Science and Engineering, University of California San Diego, San Diego, California. Before joining University of California San Diego, San Diego, California, in 2013 as an assistant professor, he was a faculty member at the University of California, Los Angeles, Los Angeles, California. Between 2011 and 2013, he took a leave to work at Microsoft Research Asia. He is a recipient of the David Marr Prize Award 2003 and a recipient of the David Marr Prize Honorable Mention Award 2015.

Jianlong Fu received the PhD degree from the Institute of Automation, Chinese Academy of Science, China, in 2015. He is currently a senior research manager with the Multimedia Search and Mining Group, Microsoft Research Asia (MSRA). His current research interests include computer vision, and multimedia content understanding. He has authored or coauthored more than 50 publications in journals and conferences, and one book chapter. He is an area chair of ACM Multimedia 2018, ICME 2019, and ICME 2020. He is a recipient of 2018 ACM Multimedia Best Paper Award.

Nicu Sebe is currently a professor at the University of Trento, Italy, leading the research in the areas of multimedia information retrieval and human behavior understanding. He was the general co-chair of ACM Multimedia 2013 and the program chair of the International Conference on Image and Video Retrieval in 2007 and 2010, ACM Multimedia 2007 and 2011, ECCV 2016, ICCV 2017, and ICPR 2020. He is a fellow of the International Association for Pattern Recognition.

Serge Belongie received the BS (with honor) degree in EE from the California Institute of Technology, Pasadena, California, in 1995 and the PhD degree in EECS from Berkeley, in 2000. He is the Andrew H. and Ann R. Tisch professor at Cornell Tech and the Computer Science Department, Cornell University, Ithaca, New York, and currently serves as an associate dean at Cornell Tech. While at Berkeley, his research was supported by the NSF Graduate Research Fellowship. From 2001-2013, he was a professor with the Department of Computer Science and Engineering, University of California, San Diego, San Diego, California. His research interests include computer vision, machine learning, crowdsourcing, and human-in-the-loop computing. He is also a co-founder of several companies including Digital Persona, Anchovi Labs (acquired by Dropbox) and Orpix. He is a recipient of the NSF CAREER Award, the Alfred P. Sloan Research Fellowship, and the Helmholtz Prize for fundamental contributions in Computer Vision.