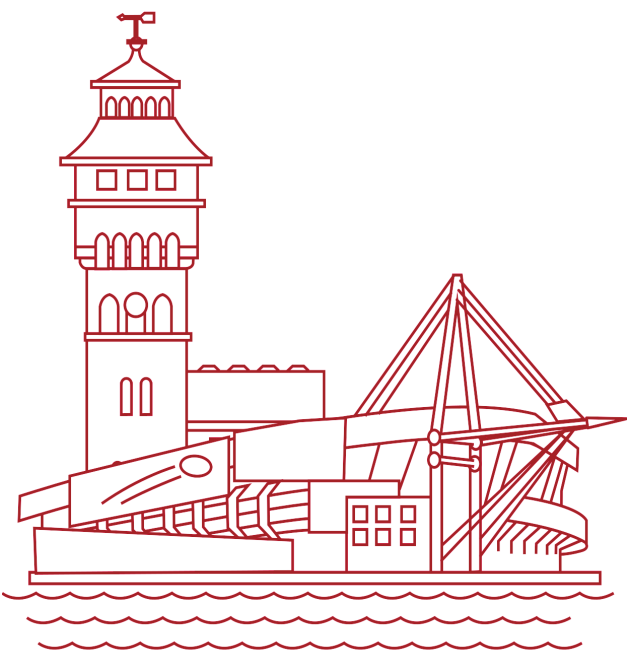




Thirtieth British Machine Vision Conference



Platinum Sponsors

facebook
Reality Labs

amazon

 **Microsoft**

 **REALSENSE™**
TECHNOLOGY



Roke

Part of the
Chemring Group

Gold Sponsors



Silver Sponsors



Special Support



CONTENTS

1	WELCOME	1
2	ORGANISERS	3
3	PROGRAMME	11
	Keynote Speakers	11
	4D Vision in the Wild	11
	Automatic Understanding of the Visual World	12
	Dissecting Neural Nets	13
	Tutorials	15
	Computational Face Analysis	15
	Robust Visual Search and Matching	16
	Location Guide	18
	Programme at a Glance	19
	Detailed Programme	20
	Monday 9 September	20
	Tuesday 10 September	20
	Wednesday 11 September	23
	Thursday 12 September	26
4	ABSTRACTS	29
	3D Computer Vision	31
	Deep Learning for Vision	45
	Document Processing	65
	Face and Gesture	67
	Statistics and Machine Learning	78
	Motion, Flow and Tracking	81
	Action and Event Recognition	87
	Biologically Inspired Vision	93
	Illumination and Reflectance	95
	Image Processing Techniques	96
	Objects and Textures	107
	Segmentation and Grouping	129
	Video Analysis	137

WELCOME

It is our great pleasure to welcome you to Cardiff for the British Machine Vision Conference (BMVC)! This is the 30th BMVC since its inception in 1990, and it is second time in Cardiff.

Cardiff, the capital of Wales, is the youngest capital city of Europe, but its history goes back to the days of the Roman Empire. The remains of the Roman stone fortifications can be seen within the walls of Cardiff Castle. Cardiff's reputation as one of Britain's six cities of elegance arises from its remarkable civic centre, regarded as being a world-ranking example of civic architecture.

The conference is hosted at Cardiff University, an ambitious and innovative Russell Group university dating back to 1883. Its world-leading research was ranked 5th amongst UK universities in the 2014 Research Excellence Framework for quality and 2nd for impact.

BMVC is one of the top events in the Computer Vision conference calendar, and is a truly international event, with a majority of papers submitted from outside the UK. This year, BMVC attracted a total of 1008 full paper submissions, which is the highest number in the history of BMVC. Of the 1008 submissions, 815 were considered valid. Of these, a total of 231 papers were accepted (38 as oral presentations and 193 as poster presentations). This amounts to a 28% overall acceptance rate. A further 32 outstanding papers, that were accepted as posters, have been invited to give a 3 minute spotlight presentation with accompanying video. There were 858 reviewers and 96 area chairs involved in the review process, generously donating their time.

We have put together an interesting programme and are delighted to welcome Adrian Hilton, Cordelia Schmid, and Antonio Torralba to the conference as keynote speakers, as well as Michel Valstar, John Collomosse, and Ondrej Chum as tutorial speakers.

BMVC has always had strong links with industry, and again we are very grateful to our industrial sponsors for supporting the event. Platinum Sponsors this year include: Facebook Reality Labs, Scape, Amazon, Microsoft, Snap, Roke, Apple, Huawei, and Intel. Gold Sponsors: Dyson, IET The Institution of Engineering and Technol-

ogy, CGG Compagnie Générale de Géophysique. Silver Sponsors: Greyparrot, Hewlett-Packard, TandemLaunch. Special support is offered by Springer.

Last but not least, we wish to thank all members of the organizing committee, the programme chairs, area chairs, reviewers, emergency reviewers, authors, and the CMT and TPMS teams for the immense amount of hard work and professionalism that has gone into making BMVC 2019 a first rate conference. We hope you find BMVC 2019 in Cardiff both an enjoyable and inspiring experience.

Kirill Sidorov and Yulia Hicks
BMVC 2019 General Chairs

ORGANISERS

BMVC 2019 is organised by members of the Visual Computing research group (School of Computer Science and Informatics) and the Sensors, Signals and Imaging research group (School of Engineering), Cardiff University.

General Chairs



Kirill Sidorov

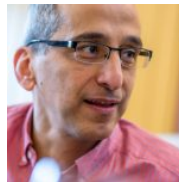


Yulia Hicks

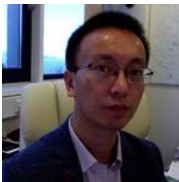
Programme Chairs



David Marshall
(Cardiff University)



Majid Mirmehdi
(University of Bristol)



Xianghua Xie
(Swansea University)

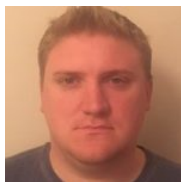


Bernard Tiddeman
(Aberystwyth University)

ORGANISERS



Joseph Redfern
Publicity Chair



Tom Hartley
Local Arrangements
Chair



David Humphreys
Video Chair



Paul Rosin
Tutorial Chair



Ze Ji
Sponsorship Chair



Yukun Lai
Poster Chair



Padraig Corcoran
Workshop Chair



Daniel Gallichan
Workshop Chair



Jing Wu
Workshop Chair



Hantao Liu
Exhibition Chair

Area Chairs

Antonis A Argyros

CSD-UOC and ICS-FORTH

Nicola Bellotto

University of Lincoln

Matthew Blaschko

KU Leuven

Paul A Bromiley

University of Manchester

Andrea Cavallaro

Queen Mary University of London

John Collomosse

University of Surrey

Sergio Escalera

CVC and University of Barcelona

Jiashi Feng

NUS

Li Fuxin

Georgia Tech

Stratis Gavves

University of Amsterdam

Jean-Yves Guillemaut

University of Surrey

Edwin R Hancock

University of York

David Hogg

University of Leeds

Frederic Jurie

University of Caen

Tae-Kyun Kim

Imperial College London

Piotr Koniusz

Data61/CSIRO, ANU

Vincent Lepetit

TU Graz

Weiyao Lin

Shanghai Jiaotong University

Chen Change Loy

The Chinese University of Hong Kong

Renaud Marlet

Ecole des Ponts ParisTech

Richard Newcombe

Oculus Research

Xiao Bai

Beihang University

Rodrigo Benenson

Google

Edmond Boyer

Inria

Neill Campbell

University of Bath

Ondrej Chum

Czech Technical University in Prague

Timothy Cootes

University of Manchester

Paolo Favaro

Bern University, Switzerland

Basura Fernando

Australian National University

Jürgen Gall

University of Bonn

Theo Gevers

University of Amsterdam

Simon Hadfield

University of Surrey

Anders Heyden

LTH

Timothy Hospedales

Edinburgh University

Joni-Kristian Kamarainen

Tampere University

Hedvig Kjellström

KTH Royal Institute of Technology

Yu-Kun Lai

Cardiff University

Fuxin Li

Oregon State University

Jim Little

University of British Columbia, Canada

Feng Lu

Beihang University

Francesc Moreno

IRI

Minh Hoai Nguyen

Stony Brook University

Adrien Bartoli

Université Clermont Auvergne

Hakan Bilen

University of Edinburgh

Toby Breckon

Durham University

Gustavo Carneiro

University of Adelaide

Adrian Clark

University of Essex

Dima Damen

University of Bristol

Michael Felsberg

Linköping University

Ying Fu

Beijing Institute of Technology

Chuang Gan

MIT-Watson AI Lab

Shaogang Gong

Queen Mary University of London

Peter Hall

University of Bath

Adrian Hilton

University of Surrey

Jia-Bin Huang

Virginia Tech

Kwang In Kim

UNIST

Nikos Komodakis

ENPC, France

Laura Leal-Taixé

TUM

Dahua Lin

The Chinese University of Hong Kong

Yonghuai Liu

Edge Hill University

Dimitrios Makris

Kingston University

Vittorio Murino

Istituto Italiano di Tecnologia

Vladimir Pavlovic

Rutgers University

ORGANISERS

Marcello Pelillo

University of Venice

Mathieu Salzmann

EPFL

Konrad Schindler

ETH

Jan Paul Siebert

University of Glasgow

Cees Snoek

University of Amsterdam

Deqing Sun

Google

Jasper Uijlings

Google Research

Jan van Gemert

Delft University of Technology

Xianghua Xie

Swansea University

Hui Yu

University of Portsmouth

Wei-Shi Zheng

Sun Yat-sen University, China

Thomas Pock

Graz University of Technology

Dimitris Samaras

Stony Brook University

Nicu Sebe

University of Trento

Leonid Sigal

University of British Columbia

Yi-Zhe Song

University of Surrey

Bernard Tiddeman

Aberystwyth University

Michel Valstar

University of Nottingham

Nuno Vasconcelos

UCSD, USA

Ming-Hsuan Yang

University of California at Merced

Jianguo Zhang

University of Dundee

Huiyu Zhou

University of Leicester

Elisa Ricci

U. Perugia

Yoichi Sato

University of Tokyo

Boxin Shi

Peking University

William Smith

University of York

Yusuke Sugano

The University of Tokyo

Emanuele Trucco

University of Dundee

Joost van de Weijer

Computer Vision Center

Richard Wilson

University of York

Kwang Moo Yi

University of Victoria

Zhao Zhang

Hefei University of Technology

Reyer Zwiggelaar

Aberystwyth University

Technical Committee

Anan Liu

Abdelaziz Djelouah

Adalberto Claudio Quiros

Adrian Barbu

Aggeliki Tsoli

Akisato Kimura

Alexander Andreopoulos

Alvaro Samagaio

Anastasios Roussos

Andrea Apicella

Andrea Tagliasacchi

Andrew Davison

Anil Armagan

Antonio Anjos

Ardhendu Behera

Armin Mustafa

Atsushi Nakazawa

Azade Farshad

Baoyuan Wu

Behnam Gholami

Bernd Freisleben

Binod Bhattacharai

Bo Wang

Boulbaba Ben Amor

Bryan Plummer

Cagri Ozcinar

Carvalho Micael

Anoop Cheriai

Abdelrahman Abdelhamed

Adel Bibi

Adrian Davison

Ahmet Iscen

Alassane Seck

Alexandre Boulch

Amine Bourki

Ancong Wu

Andrea Giachetti

Andrea Torsello

Andrew French

Anil Baslamisli

Antreas Antoniou

Aria Ahmadi

Arslan Basharat

Avinash Kumar

Azam Hamidinekoo

Baptiste Angles

Behnaz Rezaei

Bertram Drost

Bishay Mina

Bodo Rosenhahn

Boyan Gao

Byeongjoo Ahn

Calden Wloka

Chandra Kambhampettu

Aakanksha Rana

Abdullah Abuolaim

Aditya Deshpande

Adrian Peter

Ahsan Iqbal

Alessandro Masullo

Alina Kuznetsova

Amir Rosenfeld

Anders Brun

Andrea Pilzer

Andrea Zunino

Andrew Gilbert

Ankan Bansal

Aravindh Mahendran

Arijit Biswas

Arun Ross

Axel Furlan

Bogdan Raducanu

Barath Daniel

Benjamin Busam

Bin Yang

Bishop Thomas

Bogdan Georgescu

Boyu Wang

Byeongsoo Kim

Can Chen

Changkui Lyu

Aaron Jackson

Abel Gonzalez-Garcia

Adria Recasens

Adrien Bousseau

Akihiro Sugimoto

Alessandro Torcinovich

Alvaro Parra

Ammar Belatreche

Anders Buch

Andrea Prati

Andreas Kuhn

Andrik Rampun

Antonio Agudo

Archana Sapkota

Arjan Kuijper

Atsushi Hashimoto

Ayan Bhunia

Baoguang Shi

Battista Biggio

Benjamin Kimia

Bingbing Zhuang

Bo Dong

Borna Ghotbi

Bruce Maxwell

C. Alejandro Parraga

Carolina Raposo

Changqing Zou

Chanhon Kim	Chao Ma	Charles Malleson	Chen Chen
Chen Gao	Chen Kong	Chen Liu	Chen Wang
Chen-Yu Lee	Chengjiang Long	Chenliang Xu	Chenshen Wu
Chetan Tonde	Chi Xu	Christian Rauch	Christian Simon
Christian Wolf	Christopher Thomas	Chu-Song Chen	Chuan Lu
Chunfeng Yuan	Cihang Xie	Claudia Lindner	Concetto Spampinato
Conghui Hu	Constantin Vertan	Corneliu Florea	Cosmin Ancuti
Da Chen	Da Li	Daeyun Shin	Daisuke Miyazaki
Damien Muselet	Dan Casas	Danica Kragic	Daniel Aliaga
Daniel Asmar	Daniel Prusa	Daniel Rebain	Daniel Zoran
Danping Zou	David Bermudez	David George	David Hunter
David Masip	David Suter	Diego Thomas	Dimiccoli Mariella
Dimitrios Kosmopoulos	Dimitrios Sakkos	Ding Liu	Ditzel Carsten
Dmitrij Csetverikov	Dmytro Mishkin	Donald Dansereau	Dong Lao
Dong Li	Dong Zhang	Donghoon Lee	Donghyun Yoo
Donglai Wei	Dániel Baráth	Eddy Ilg	Edmond S. L. Ho
Edward Kim	Ehrhardt Sebastien	Ehsan Adeli	Eirikur Agustsson
Eldar Insafutdinov	Elena Balashova	Emre Akbas	Engel Nico
Enliang Zheng	Enric Meinhardt-Llopis	Enrique Sánchez-Lozano	Eraldo Ribeiro
Erhan Gundogdu	Eric Sommerlade	Eric Wengrowski	Erich Kobler
Ernest Valveny	Etienne Grossmann	Eunwoo Kim	Eyasu Zemene
Fabian Benítez-Quiroz	Fabio Poiesi	Faisal Qureshi	Fan Jia
Fang Wang	Fanta Camara	Farnoosh Heidarivinchel	Fatemeh Karimi Nejadasl
Fatih Porikli	Fatma Güney	Feng Zhao	Filip Radenovic
Flavio Vidal	Francesco Isgro	Francisco Barranco	Francisco Flórez-Revuelta
Francisco Vasconcelos	Franck Vidal	Frank Michel	Fred Labrosse
Gokberk Cinbis	Gabriel Maicas	Gang Yu	Gary Tam
Gaurav Sharma	Gautam Singh	Geneviève Patterson	George Vogiatzis
Gerardo Aragon-Camarasa	Gianfranco Doretto	Giannis Pavlidis	Gim Hee Lee
Giorgos Tolias	Giovanni Farinella	Go Irie	Gregor Miller
Gregory Kramida	Guangming Zhu	Guangtao Nie	Guido Pusiol
G.-A. Bilodeau	Guillermo Gallego	G. Garcia-Hernando	Guler Riza
Gunhee Kim	Guodong Guo	Guofeng Zhang	Gurkirt Singh
H Wang	Hazim Ekenel	Hai Pham	Haijun Zhang
Haixia Wang	Hakki Karaimer	Han-Pang Chiu	Hang Zhang
Hanno Ackermann	Hansung Kim	Hanwang Zhang	Haotian Xu
He Zhang	Hector Basevi	Hedvig Kjellström	Heewon Kim
Heikki Huttunen	Helder Araujo	Helge Rhodin	Hengshuang Zhao
Herb Yang	Hedy Mendez-Vazquez	Hideo Saito	Hilde Kuehne
Hoang-An Le	Holger Caesar	Hongguang Zhang	Hongxing Wang
Hongyang Li	Hua Wang	Huaizhong Zhang	Huaizu Jiang
Huazhu Fu	Hubert P. H. Shum	Hueihan Jhuang	Hugues Talbot
Hui Fang	Hui Zhang	Huidong Liu	Huu Le
Hwann-Tzong Chen	Hyun Soo Park	Hyung Jin Chang	Hyunjung Shim
Ichiro Ide	Ignacio Rocco	Ilkay Ulusoy	Imari Sato
Ioannis Patras	Ismail Elezi	Ivan Oseledets	Iván Eichhardt
Jiahuan Zhou	Jacinto Nascimento	Jack Turner	Jacopo Cavazza
Jaesik Park	James Brown	Jan Prokaj	Javier Lorenzo-Navarro
Javier Ruiz-del-Solar	Javier Traver	Jean-Philippe Tarel	Jesus Bermudez-Cameo
Ji Zhang	Ji Zhu	Jia Xue	Jiabai Zeng
Jiajun Lu	Jian Dong	Jian Sun	Jian-Hao Luo
Jiangxin Dong	Jianhui Chen	Jianjia Wang	Jianwen Xie
Jiaqi Wang	Jiawei Zhang	Jie Qin	Jie Yang
Jifei Song	Jim Little	Jin Gao	Jin Sun
Jing Wu	Jingchun Cheng	Jingjing Deng	Jinglu Wang
Jingying Chen	Jingyong Su	Jinshan Pan	Jinshi Cui
Jinwei Ye	Joachim Dehais	Joachim Denzler	Joe Kileel
Joerg Stueckler	John Collomosse	John See	John Zelek
John chiverton	Jonathan Ventura	Jongwoo Lim	Jorge Batista
Joseph Tighe	Joss Whittle	Ju Hong Yoon	Julien Schroeter
Jun Tang	Jun Wan	Jun Zhou	Jun-Cheng Chen
Junchi Yan	Jungseock Joo	Junli Tao	Junliang Xing
Junseok Kwon	Junsik Kim	Junyu Gao	Justin Lazarow
Kai Chen	Kai Li	Kai Xu	Kaiwen Guo
Kan Chen	Kandan Ramakrishnan	Kaustav Kundu	Ke Ma
Ke Yu	Kean Chen	Keiji Yanai	Kenichi Kanatani

ORGANISERS

Kenji Hara	Keshav Seshadri	Kevin Chen	Kinh Tieu
Kirill Gavriluk	Koichi Shinoda	Konstantinos Papoutsakis	Konstantinos Vougioukas
Kuang-Jui Hsu	Kwan-Yee Wong	Kwang Moo Yi	Kwok-Ping Chan
Kyle Wilson	Lakshmanan Nataraj	Lanaras Charis	Laura Sevilla-Lara
Le Hou	Le Wang	Lei He	Lei Tong
Lei Wang	Lei Zhou	Levente Hajder	Lezi Wang
Li Shen	Li Zhang	Liam Hiley	Liang Du
Liang Haoyi	Liang Zhang	Lilian Calvet	Liliana Lo Presti
Lin Chen	Lin Gao	Lin Ma	Lin Sun
Linguang Zhang	Lingxi Xie	Linxi Fan	Liuhao Ge
Liwei Wang	Liyan Chen	Long Chen	Long Mai
Longyin Wen	Lu Sheng	Lu Yu	Luca Cosmo
Lucas Deecke	Ludovic Magerand	Luis Herranz	Lukás Neumann
Mark Nixon	Minu George	Mahmoud Afifi	Mahmudul Hasan
Malleson Charles	Mandar Dixit	Maneesh Singh	Mang YE
Manjunath Narayana	Manolis Lourakis	Mantini Pranav	Manuel J. Marín-Jiménez
Marco Körner	Marco Paladini	Marco Pedersoli	Marco Piccirilli
Marco Volino	Markus Vincze	Martin Bach	Martin Fergie
Martin Kampel	Martin R. Oswald	Martin Weinmann	Masayuki Tanaka
Mason McGill	Mateusz Kozinski	Mathieu Aubry	Mathieu Bredif
Matt Leotta	Matthew Shere	Matthew Toews	Mattias Heinrich
Maxime Lhuillier	Mayank Vatsa	Megha Nawhal	Mehrtash Harandi
Mei Han	Mei-Chen Yeh	Meng Tang	Miaomiao Liu
Michael Breuß	Michael Edwards	Michael Gygli	Michael Hofmann
Michael Pound	Michael Wray	Michel Antunes	Michele Sasdelli
Michele Volpi	Mihai Puscas	Mikhail Sizintsev	Min H. Kim
Ming-Ming Cheng	Mingbo Zhao	Mingli Song	Minyoung Kim
Mohamed Daoudi	M. Soltaninejad	Mohammed Mahmoud	Mohsen Ghafoorian
Moin Nabi	Monica Hernandez	Morteza Ghahremani	Naemullah Khan
Nam Ik Cho	Nanne van Noord	Naoki Chiba	Nashid Alam
Nazli Ikizler-Cinbis	Nian Liu	Nicholas Rhinehart	Nick Barnes
Nick Michiels	Nicolau Leal Werneck	Nikhil Rasiwasia	Niloy Mitra
Nima Khademi Kalantari	Ognjen Arandjelovic	Olac Fuentes	Olaf Kaelher
Olivia Wiles	Omid Hosseini Jafari	Oscar Mendez	Oswald Lanz
Oussama Ennaffi	Pablo Márquez Neila	Pablo Zegers	Pan He
Pan Ji	Parmeshwar Khurd	Pascal Mettes	Pascal Monasse
Patrick Knöbelreiter	Patrick Peursum	Paul Bromiley	Paul Henderson
Pedro O. Pinheiro	Pedro Rodrigues	Peng Tang	Peter Barnum
Pfeuffer Andreas	Phuc Nguyen	Pinar Duygulu	Ping Wei
Piotr Ozimek	Pradip Mainali	Pramod Sharma	Praveer Singh
Qi Dong	Qi Ye	Qi Zou	Qian Zheng
Qian Yu	Qieyun Dai	Qifeng Chen	Qijun Zhao
Qing Wang	Qing Zhang	Radim Tylecek	Raghudeep Gadde
Rama Chellappa	Rama Nkovvuri	Ramanathan Subramanian	Rameswar Panda
Ran He	Ran Song	Raoul de Charette	Raviteja Vemulapalli
Remco Duits	Renjie Liao	Reyer Zwiggelaar	Riaan Zoetmulder
Richa Singh	Rizwan Chaudhry	Robert DeBortoli	Robert Fisher
Robert Tamburo	Roberto Cesar	Roberto Lopez-Sastre	Roberto Paredes
Roey Mechrez	Rohit Girdhar	Rohit Pandey	Ronald Clark
Rudrasis Chakraborty	Rui Huang	Ruigang Yang	Ruiping Wang
Ruiyu Li	Runhao Zeng	Runze Zhang	Ruohan Gao
Ryo Furukawa	Ryo Yonetani	Ryusuke Sagawa	Sabu Emmanuel
Saeid Motiian	Said Pertuz	Salman Khan	Sam Tsai
Samarth Brahmhatt	Samuel Albanie	Sarah Ostadabbas	Scott McCloskey
Scott Wehrwein	Scott Workman	Sebastiano Vascon	Selen Pehlivan
Senjian An	Seonwook Park	Ser-Nam Lim	Sergey Tulyakov
Seungryul Baek	Shadi Albarqouni	Shah Nawaz	Shalini De Mello
Shang-Hong Lai	Shao-Hua Sun	Shashank Jaiswal	Shell Hu
Sheng Huang	Sheng Li	Shiguang Shan	Shih-Yao Lin
Shiliang Zhang	Shin'ichi Satoh	Shiro Kumano	Shishir Shah
Shiyu Song	Shohei Nobuhara	Shu Liu	Shu Zhang
Shuang Yang	Shugao Ma	Shyamal Buch	Siddhartha Chandra
Silvia Pinteá	Simon Donne	Simone Bianco	Simone Gasparini
Sinan Kalkan	Simon Aslan	Siva Karthik Mustikovela	Sivapriya Kannappan
Song Bai	Soumyadip Sengupta	Spyros Gidaris	Srikanth Muralidharan
Srikar Appalaraju	Srinath Sridhar	Stan Li	Stavros Petridis

Stephan Liwicki	Stephan Zheng	Stephane Herbin	Stephen Maybank
Stuart James	Sudipta Sinha	Sunghyun Cho	Suyog Jain
Søren Ingvor Olsen	Tae-Kyun Kim	Tak-Wai Hui	Takahiro Okabe
Takayoshi Yamashita	Takayuki Okatani	Takeshi Oishi	Tammy Riklin Raviv
Tanner Schmidt	Tanveer Syeda-Mahmood	Tao Wang	Tao Zhang
Tat-Jen Cham	Tat-Jun Chin	Tatiana Tommasi	Teofilo deCampos
Thibaut Durand	Thierry Bouwmans	Thomas Hartley	Thomas Mensink
Thomas Whelan	Tianmin Shu	Tianzhu Zhang	Tim Morris
Tingting Jiang	Tobias Fischer	Tobias Ritschel	Toby Collins
Toby Perrett	Tolga Birdal	Tom Cocksedge	Tom Haines
Tom Runia	Tomas Jakab	Tomasz Trzcinski	Tong He
Tony Tung	Toru Tamaki	Toshihiko Yamasaki	Toufiq Parag
Tryphon Lambrou	Tsung-Yu Lin	Tu Bui	Tushar Nagarajan
Tyng-Luh Liu	Umar Muhammad	Unaiza Ahsan	Vaclav Hlavac
Vakhitov Alexander	Varun Nagaraja	Vasileios Belagiannis	Vazquez-Corral Javier
Vicky Kalogeiton	Victor Fragoso	Victor Prisacariu	Victor Larsson
Viresh Ranjan	Vishal Patel	Volker Blanz	Véronique Prinet
Victor Ponce-López	Wangmeng Zuo	Wei Chen	Wei Hong
Wei Jiang	Wei Li	Wei Tang	Wei Yang
Wei Zeng	Wei-Chen Chiu	Wei-Chih Hung	Wei-Chih Tu
Wei-Chiu Ma	Wei-Shi Zheng	Weibo Liu	Weidi Xie
Weipeng Xu	Weisheng Dong	Weiwei Sun	Weiwei Lin
Wen Li	Wengang Zhou	Wenqi Ren	Wenwu Wang
Wenyan Yang	William Thong	Winston Hsu	Wolfgang Foerstner
Wonjun Hwang	Wonmin Byeon	Xide Xia	Xenophon Zabulis
Xialei Liu	Xianfang Sun	Xiang Wang	Xiang Yu
Xiangyu Xu	Xiangyu Zhang	Xiao Bai	Xiaochun Cao
Xiaodong Yang	Xiaohan Nie	Xiaohang Zhan	Xiaohe Wu
Xiaojuan Qi	Xiaolong Wang	Xiaoqun Zhang	Xiaoyang Wang
Xiaoyang Tan	Xiaoyi Jiang	Xiatian Zhu	Xilin Chen
Xin Li	Xin Yu	Xingang Pan	Xingchao Peng
Xinggang Wang	Xingrui Yang	Xinyi Zhou	Xinlei Chen
Xinqing Guo	Xintao Wang	Xinxin Zuo	Xirong Li
Xiu-Shen Wei	Xu Lan	Xu Zhang	Xucong Zhang
Xuehan Xiong	Xueting Li	Xun Xu	Yagiz Aksoy
Yahong Han	Yan Tong	Yanbei Chen	Yanchao Yang
Yang Long	Yang Wang	Yang Xiao	Yannis Panagakis
Yasrab Robail	Yasushi Makihara	Yasushi Yagi	Yasutomo Kawanishi
Yazan Abu Farha	Yi Fang	Yi Wu	Yi Zhu
Yi-Hsuan Tsai	Yibing Song	Yifan Xia	Yifan Yang
Yihong Wu	Yijun Li	Yijun Xiao	Yiming Wang
Yin Li	Yin Zheng	Yinda Zhang	Ying Nian Wu
Yingcong Chen	Yingying Zhu	Yining Li	Yinlin Hu
Yiru Shen	Yixin Zhu	Yongjie Li	Yongsheng Dong
Young Min Shin	Youngkyoon Jang	Yu-Chiang Frank Wang	Yu-Xiong Wang
Yuanlu Xu	Yuchao Dai	Yue Zhao	Yueming Wang
Yufei Wang	Yuhang Ming	Yulan Guo	Yulei Niu
Yuliang Zou	Yun-Chun Chen	Yung-Yu Chuang	Yunlu Chen
Yunqiang Li	Yusuf Tas	Zachary Daniels	Zechao Li
Zhang Zheng	Zhaopeng Cui	Zhe Wang	Zhen Lei
Zhen Liu	Zhen-Hua Feng	Zheng Wu	Zhenxing Niu
Zhibin Niu	Zhibo Yang	Zhihai He	Zhihua Liu
Zhili Chen	Zhiwei Deng	Zhixin Shu	Zhiyuan Fang
Zhu Siyu	Zhuolin Jiang	Zihan Zhou	Zijun Wei
Ziwei Liu	Ziyi Shen	Zoran Duric	Zoya Bylinskii
Zsolt Kira	Yaxing Wang		

Keynote Speakers

4D Vision in the Wild

Over the past decade 3D Computer Vision has advanced from reconstruction of static scenes under controlled conditions towards full 4D spatio-temporal reconstruction and structured modelling of complex dynamic scenes. This talk will review recent advances in this field towards general 4D reconstruction and understanding of unconstrained scenes highlighting real-world challenges. High-level semantic understanding and reconstruction of dynamic scenes leveraging deep-learning will be presented. 4D dynamic understanding of scenes and people is an enabling technology for applications ranging from healthcare and security, through to immersive entertainment and autonomous robotic systems that can work safely alongside people at home or work. Examples of collaborative research to enable immersive audio-visual entertainment and monitoring of people for healthcare at home will be presented.



Prof Adrian Hilton
University of Surrey

Prof Adrian Hilton, BSc (hons), DPhil, CEng, FIET, is Professor of Computer Vision and Director of the Centre for Vision, Speech and Signal Processing at the University of Surrey, UK.

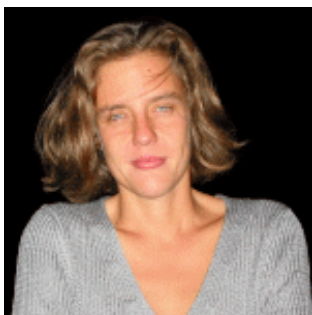
The focus of his research is Perceptual AI enabling machines to understand and interact with the world through seeing and hearing. This combines the fields of computer vision and machine learning to develop new methods for reconstruction, modelling, and understanding natural scenes from video and audio.

He is an internationally recognised expert in 3D and 4D computer vision. His research has contributed to advancing machine perception from controlled static scenes to real-world dynamic

scenes and people. This is a key technology for future intelligent systems allowing human-machine interaction in robotics, health-care, assisted living, entertainment, and immersive experiences.

Adrian has successfully commercialised technologies for 3D and 4D shape capture exploited in entertainment, manufacture and health, receiving two EU IST Innovation Prizes, a Manufacturing Industry Achievement Award, a Royal Society Industry Fellowship with Framestore on Digital Doubles for Film, and a Royal Society Wolfson Research Merit Award in 4D Vision. He is currently Principal Investigator on the EPSRC Programme Grant “S3A Future Spatial Audio” bringing together expertise in audio, vision and human perception to achieve immersive listener experiences at home or on the move.

Automatic Understanding of the Visual World



Prof Cordelia Schmid

Inria

One of the central problems of artificial intelligence is machine perception, i.e. the ability to understand the visual world based on input from sensors such as cameras. In this talk, Prof Cordelia Schmid will present recent progress of her team in this direction. Data play a key role, and the talk will start with presenting results on how to generate additional training data using weak annotations, motion information, and synthetic data. Next, recent results for action recognition will be presented, where human tubes and tubelets have shown to be successful. The tubelets move away

from state-of-the-art frame-based approaches and improve classification and localization by relying on joint information from several frames and the interaction with objects. Finally, some recent results on robot manipulation will be presented.

Prof Cordelia Schmid holds a M.S. degree in Computer Science from the University of Karlsruhe and a Doctorate, also in Computer Science, from the Institut National Polytechnique de Grenoble (INPG). Her doctoral thesis received the best thesis award from INPG in 1996. Dr Schmid was a post-doctoral research assistant in the Robotics Research Group of Oxford University in 1996–1997.

Since 1997 she has held a permanent research position at Inria Grenoble Rhone-Alpes, where she is a research director and directs an Inria team.

Dr Schmid has been an Associate Editor for IEEE PAMI (2001–2005) and for IJCV (2004–2012), editor-in-chief for IJCV (2013–), a program chair of IEEE CVPR 2005 and ECCV 2012 as well as a general chair of IEEE CVPR 2015 and ECCV 2020. In 2006, 2014 and 2016, she was awarded the Longuet-Higgins prize for fundamental contributions in computer vision that have withstood the test of time. She is a fellow of IEEE.

She was awarded an ERC advanced grant in 2013, the Humbolt research award in 2015, and the Inria and French Academy of Science Grand Prix in 2016. She was elected to the German National Academy of Sciences, Leopoldina, in 2017. In 2018 she received the Koenderink prize for fundamental contributions in computer vision that have withstood the test of time. Starting 2018 she holds a joint appointment with Google research.

Dissecting Neural Nets

With the success of deep neural networks and access to image databases with millions of labelled examples, the state of the art in computer vision is advancing rapidly. Even when no examples are available, Generative Adversarial Networks (GANs) have demonstrated a remarkable ability to learn from images and are able to create nearly photorealistic images. The performance achieved by convNets and GANs is remarkable and constitutes the state of the art on many tasks. But why do convNets work so well? What is the nature of the internal representation learned by a convNet in a classification task? How does a GAN represent our visual world internally? In this talk Prof Antonio Torralba will show that the internal representation in both convNets and GANs can be interpretable in some important cases. He will then show several applications for object recognition, computer graphics, and unsupervised learning from images and audio.



Prof Antonio Torralba
Massachusetts Institute of
Technology

Prof Antonio Torralba is a Professor of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT), the MIT director of the MIT-IBM Watson AI Lab, and the inaugural director of the MIT Quest for Intelligence, a MIT campus-wide initiative to discover the foundations of intelligence. He received the degree in telecommunications engineering from Telecom BCN, Spain, in 1994 and the Ph.D. degree in signal, image, and speech processing from the Institut National Polytechnique de Grenoble, France, in 2000. From 2000 to 2005, he spent postdoctoral training at the Brain and Cognitive Science Department and the Computer Science and Artificial Intelligence Laboratory, MIT, where he is now a professor. Prof Torralba is an Associate Editor of the International Journal in Computer Vision, and has served as program chair for the Computer Vision and Pattern Recognition conference in 2015. He received the 2008 National Science Foundation (NSF) Career award, the best student paper award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2009, and the 2010 J. K. Aggarwal Prize from the International Association for Pattern Recognition (IAPR). In 2017, he received the Frank Quick Faculty Research Innovation Fellowship and the Louis D. Smullin ('39) Award for Teaching Excellence.

Tutorials

Computational Face Analysis

In this tutorial Prof Michel Valstar will go through the pipeline of computational face analysis, which includes face detection, face recognition, facial expression recognition, and higher-level behaviour analysis such as the prediction of behaviomedical conditions, for example depression or pain. The tutorial will be focused on practical issues that one needs to consider, and will use the popular open-source toolbox “OpenFace”. Prof Michel Valstar will also go through a number of valuable publicly available face-related databases. While the focus will be on practical implementations, Prof Michel Valstar will refer to and briefly address the literature so attendees can follow up on this at their own leisure after the tutorial.



Prof Michel Valstar
University of Nottingham

Prof Michel Valstar is an associate professor in Computer Science at the University of Nottingham, and member of both the Computer Vision and Mixed Reality Labs. He is an expert the fields of computer vision and pattern recognition, where his main interest and world-leading work is in automatic recognition of human behaviour, specialising in the analysis of facial expressions. Valstar pioneered the concept of Behaviomedics, which aims to diagnose, monitor, and treat medical conditions that alter expressive behaviour by employing objective assessment of that behaviour. Previously he was a Visiting Researcher at MIT’s Media Lab, and a Research Associate in the intelligent Behaviour Understanding Group (iBUG) at Imperial College London. He received his masters degree in Electrical Engineering at Delft University of Technology in 2005 and his PhD at Imperial College London in 2008. He is the founder of the facial expression recognition challenges, FERA 2011/2015/2017, and the Audio-Visual Emotion recognition Challenge series, AVEC 2011–2018. He leads the Objective Assessment research area as the only non-professorial Research Area lead of a £23.6M Biomedical Research Centre, and was the coordinator of the EU Horizon 2020

project ARIA-VALUSPA. Valstar is recipient of Melinda & Bill Gates Foundation funding to help premature babies survive in the developing world. His work has received popular press coverage in The Guardian, Science Magazine, New Scientist, CBC, and on BBC Radio, among others. Valstar is a senior member of the IEEE. He has published over 90 peer-reviewed articles, attracting >7,500 citations and attaining an H-index of 36.

Robust Visual Search and Matching



Prof John Collomosse
University of Surrey



Prof Ondrej Chum
Czech Technical University in
Prague

Visual search and matching are long-standing challenges in computer vision, transformed by deep learning. This tutorial will focus on the latest CNN architectures for robust visual search and retrieval. Prof John Collomosse and Prof Ondrej Chum will analyse design choices and compare various components of the latest descriptor designs, including the aggregation, dimensionality reduction, binarisation, and end-to-end training of large-scale visual search systems. As examples, the REMAP global descriptor, which won the Google Landmark Retrieval Challenge on Kaggle in 2018, and the topic of cross-domain matching through deep representations that disentangle structure and style enabling, for example, sketch based search, will be analysed. The tutorial will cover contemporary methods improving visual search techniques by considering structures, often called manifolds, created by the descriptors of relevant images in the descriptor space. Both, the query-time methods, such as query expansion, and the offline methods, in particular diffusion, which shifts some of the computation into the preprocessing stage, will be considered.

Prof John Collomosse is a Professor of Computer Vision at the Centre for Vision Speech and Signal Processing (CVSSP), and visiting professor at Adobe Research, Creative Intelligence Lab. John

joined CVSSP in 2009. Previously he was an Assistant Professor at the Department of Computer Science, University of Bath, where he completed his PhD in 2004 on the topic of AI for Image Stylization. John has also spent periods of time in commercial R&D, working for IBM UK Labs (Hursley), Vodafone R&D (Munich), Hewlett Packard Labs (Bristol); the latter under a Royal Academy of Engineering fellowship. His research focuses on the interaction of Computer Vision, Graphics, and AI for the creative industries, specifically for human performance capture, video post-production and vfx, and intuitive visual search (particularly sketch search). He also heads up the Surrey Blockchain activity exploring the fusion of AI and Distributed Ledger Technologies. John is a Chartered Engineer (C.Eng, 2013) and since 2018 a member of the EPSRC ICT Strategic Advisory Team (SAT) and UKRI Digital Economy Programme Advisory Board (PAB).

Prof Ondrej Chum is an associate professor at the Czech Technical University in Prague, where he leads a team within the Visual Recognition Group at the Department of Cybernetics, Faculty of Electrical Engineering. He received the MSc degree in computer science from Charles University, Prague, in 2001 and the PhD degree from the Czech Technical University in Prague, in 2005. From 2006 to 2007, he was a postdoctoral researcher at the Visual Geometry Group, University of Oxford, United Kingdom. The research interests include large-scale image and particular object retrieval, object recognition, and robust estimation of geometric models. He is a member of Image and Vision Computing editorial board, and has served in various roles at major international conferences (e.g., ICCV, ECCV, CVPR, and BMVC). Ondrej co-organizes Computer Vision and Sports Summers School in Prague. He was the recipient of the Best Paper Prize at the BMVC in 2002, the Best Science Paper Honorable Mention at BMVC 2017, Longuet-Higgins Prize at CVPR 2017, and the Saburo Tsuji Best Paper Award at ACCV 2018. Ondrej was awarded the 2012 Outstanding Young Researcher in Image & Vision Computing runner up for researchers within seven years of their PhD.


Location Guide



Most of the main conference events (keynote speakers, oral and spotlight presentations, workshops, and tutorials) will take place in the Sir Martin Evans Building (SME). The tea breaks, lunches, poster sessions, and the dinner on Tuesday will take place in the Students' Union (SU). The Snap Welcome Reception (Monday evening) will be held in the Viriamu Jones Gallery (VJG) of the university's Main Building. Finally, the conference banquet on Wednesday will be enjoyed by the delegates at the National Museum Cardiff (NMC). **Registration each morning takes place in the foyer of the Sir Martin Evans Building (SME).**

Programme at a Glance


Registration each morning takes places in the foyer of the Sir Martin Evans Building (SME).

	Mon 9th	Tue 10th	Wed 11th	Thu 12th	
08:00		Registration SME	Registration	Registration	08:00
08:45		Welcome SME			
09:00	Keynotes:	4D Vision in the Wild Adrian Hilton SME	Automatic Understanding of the Visual World Cordelia Schmid SME	Dissecting Neural Nets Antonio Torralba SME	09:00
10:00		Spotlights (1–16) SME	Spotlights (140–155) SME	Orals (291–294) Motion and Flow SME	10:00
11:00		Tea break SU	Tea break SU	Tea break SU	11:00
12:00	Registration SME	Orals (17–22) Deep Learning for Vision SME	Orals (156–161) 3D Computer Vision SME	Orals (295–300) Video Analysis SME	11:45
13:15	Welcome SME	Lunch SU	Lunch SU	Lunch SU	13:15
13:30	Tutorial Computational Face Analysis Michel Valstar SME	Posters (23–131)	Posters (162–282)	Workshops SME	14:00
15:30	Tea break SU	Tea served SU	Tea served SU	Tea break SU	15:45
16:15	Tutorial Robust Visual Search and Matching John Collomosse Ondrej Chum SME	Orals (132–139) Deep Learning for Vision SME	Orals (283–290) Objects, Segmentation, Textures, Colours SME	Workshops SME	16:30
18:15					18:15
19:00	Snap Inc.  Welcome Reception VJG	Dinner SU	Banquet NMC		

Detailed Programme

All presentations (be it orals, spotlights, or posters) are sequentially **numbered** in the programme. The corresponding abstracts can be seen on pages 29–144.

Monday 9 September

12:00 – 13:15	Registration	
13:15 – 13:30	Welcome	
13:30 – 15:30	Tutorial	SME
	Computational Face Analysis	
	Prof. Michel Valstar (University of Nottingham)	
15:30 – 16:15	Tea Break	SU
16:15 – 18:15	Tutorial	SME
	Robust Visual Search and Matching	
	Prof. John Collomosse (University of Surrey)	
	Prof. Ondrej Chum (Czech Technical University in Prague)	
19:00 – late	Reception	VJG
	 Welcome Reception sponsored by Snap Inc.	

Tuesday 10 September

08:00 – 08:45	Registration	
08:45 – 09:00	Welcome	
09:00 – 10:00	Keynote	SME
	4D Vision in the Wild	
	Prof. Adrian Hilton (University of Surrey)	
	Sponsored by Facebook	
10:00 – 11:00	Spotlights (Session 1)	SME
1	Adversarial View-Consistent Learning for Monocular Depth Estimation	
	Yixuan Liu (Tsinghua University), Yuwang Wang (Microsoft Research), Shengjin Wang (Tsinghua University)	
2	Joint Spatial and Layer Attention for Convolutional Networks	

Tony Joseph (University of Ontario Institute of Technology), Konstantinos Derpanis (Ryerson University), Faisal Qureshi (University of Ontario Institute of Technology)

3 A Less Biased Evaluation of Out-of-distribution Sample Detectors

Alireza Shafaei (The University of British Columbia), Mark Schmidt (University of British Columbia), Jim Little (University of British Columbia)

4 Unmasking the Devil in the Details: What Works for Deep Facial Action Coding?

Koichiro Niinuma (Fujitsu Laboratories of America, Inc.), Laszlo Jeni (Carnegie Mellon University), Jeffrey Cohn (University of Pittsburgh), Itir Onal Ertugrul (Carnegie Mellon University)

5 Multi-Weight Partial Domain Adaptation

Jian Hu (Shanghai Jiaotong University), Hongya Tuo (Shanghai Jiaotong University), Chao Wang (Shanghai Jiaotong University), Lingfeng Qiao (Shanghai Jiaotong University), Haowen Zhong (Shanghai Jiaotong University), Zhongliang Jing (Shanghai Jiaotong University)

6 Text Recognition using local correlation

Yujia Li (Institute of Information Engineering, Chinese Academy of Sciences), Hongchao Gao (Institute of Information Engineering, Chinese Academy of Sciences), Xi Wang (Institute of Information Engineering, Chinese Academy of Sciences), Jizhong Han (Institute of Information Engineering, Chinese Academy of Sciences), Ruixuan Li (Huazhong University of Science and Technology)

7 DetectFusion: Detecting and Segmenting Both Known and Unknown Dynamic Objects in Real-time SLAM

Ryo Hachiuma (Keio University), Christian Pirchheim (Graz University of Technology), Dieter Schmalstieg (Graz University of Technology), Hideo Saito (Keio University)

8 Bilinear Siamese Networks with Background Suppression for Visual Object Tracking

Hankyeol Lee (Korea Advanced Institute of Science and Technology), Seokeon Choi (Korea Advanced Institute of Science and Technology), Youngeun Kim (Korea Advanced Institute of Science and Technology), Changick Kim (Korea Advanced Institute of Science and Technology)

9 Adversarial Examples for Handcrafted Features

Muhammad Latif Anjum (NUST), Zohaib Ali (NUST), Wajahat Hussain (NUST - SEecs)

10 One-shot Face Reenactment

Cheng Li (SenseTime Research), Yunxuan Zhang (SenseTime Research), Yue He (SenseTime Research), Siwei Zhang (SenseTime Research), Ziwei Liu (The Chinese University of Hong Kong), Chen Change Loy (Nanyang Technological University)

11 Learning to Focus and Track Extreme Climate Events

Sookyoung Kim (Lawrence Livermore National Laboratory), Sunghyun Park (Korea University), Sunghyo Chung (Korea University), Joonseok Lee (Google Research), Yunsung Lee (Korea University), Hyojin Kim (LLNL), Prabhat (Lawrence Berkeley National Laboratory), Jaegul Choo (Korea University)

12 Revisiting Residual Networks with Nonlinear Shortcuts

Chaoning Zhang (Korea Advanced Institute of Science and Technology), Francois Rameau (Korea Advanced Institute of Science and Technology), Jean-Charles Bazin (Korea Advanced Institute of Science and Technology), Dawit Mureja Argaw (Korea Advanced Institute of Science and Technology), Philipp Benz (Korea Advanced Institute of Science and Technology), Seokju Lee (Korea Advanced Institute of Science and Technology), Junsik Kim (Korea Advanced Institute of Science and Technology), In So Kweon (Korea Advanced Institute of Science and Technology)

13 Unified 2D and 3D Hand Pose Estimation from a Single Visible or X-ray Image

Akila Pemasiri (Queensland University of Technology), Kien Nguyen Thanh (Queensland University of Technology), Sridha Sridharan (Queensland University of Technology), Clinton Fookes (Queensland University of Technology)

- 14 Perspective-n-Learned-Point: Pose Estimation from Relative Depth**
 Nathan Piasco (Univ. Bourgogne Franche-Comte), Désiré Sidibé (Université de Bourgogne), Cedric Demonceaux (Univ. Bourgogne Franche-Comte), Valérie Gouet-Brunet (LASTIG/IGN)
- 15 Joint Multi-view Texture Super-resolution and Intrinsic Decomposition**
 Wei Dong (Carnegie Mellon University), Vagia Tsiminaki (ETH Zurich), Martin R. Oswald (ETH Zurich), Marc Pollefeys (ETH Zurich / Microsoft)
- 16 Class-Distinct and Class-Mutual Image Generation with GANs**
 Takuhiro Kaneko (The University of Tokyo), Yoshitaka Ushiku (The University of Tokyo), Tatsuya Harada (The University of Tokyo / RIKEN)

11:00 – 11:45 Tea Break

SU

11:45 – 13:15 Oral Presentations:

SME

Deep Learning for Vision (Session 1)

Chair: Xianghua Xie

Sponsored by Microsoft

- 17 Guided Zoom: Questioning Network Evidence for Fine-grained Classification**
 Sarah Bargal (Boston University), Andrea Zunino (Istituto Italiano di Tecnologia), Vitali Petsiuk (Boston University), Jianming Zhang (Adobe Research), Kate Saenko (Boston University), Vittorio Murino (Istituto Italiano di Tecnologia), Stan Sclaroff (Boston University)
- 18 Embodied Vision-and-Language Navigation with Dynamic Convolutional Filters**
 Federico Landi (University of Modena and Reggio Emilia), Lorenzo Baraldi (University of Modena and Reggio Emilia), Massimiliano Corsini (University of Modena and Reggio Emilia), Rita Cucchiara (University of Modena and Reggio Emilia)
- 19 Accurate and Compact Convolutional Neural Networks with Trained Binarization**
 Zhe Xu (City University of Hong Kong), Ray Cheung (City University of Hong Kong)
- 20 Show, Infer and Tell: Contextual Inference for Creative Captioning**
 Ankit Khare (University of Texas at Arlington), Manfred Huber (University of Texas at Arlington)
- 21 Push for Quantization: Deep Fisher Hashing**
 Yunqiang Li (Delft University of Technology), Wenjie Pei (Tencent), yufei zha (Air Force Engineering University), Jan van Gemert (Delft University of Technology)
- 22 RecNets: Channel-wise Recurrent Convolutional Neural Networks**
 George Retsinas (National Technical University of Athens), Athena Elafrou (National Technical University of Athens), Georgios Goumas (National Technical University of Athens), Petros Maragos (National Technical University of Athens)

13:15 – 14:00 Lunch

SU

14:00 – 16:15 Posters (Session 1)

SME

23–136, see poster abstracts on pages 29–144.

16:15 – 18:15 Oral Presentations:

SME

Deep Learning for Vision (Session 2)

Chair: Bernard Tiddeman
Sponsored by Snap

- 137 Do Saliency Models Detect Odd-One-Out Targets? New Datasets and Evaluations**
Iuliia Kotseruba (York University), Calden Wloka (York University), Amir Rasouli (York University), John Tsotsos (York University)
- 138 PCAS: Pruning Channels with Attention Statistics for Deep Network Compression**
Kohei Yamamoto (Oki Electric Industry Co., Ltd.), Kurato Maeno (Oki Electric Industry Co., Ltd.)
- 139 Large Margin In Softmax Cross-Entropy Loss**
Takumi Kobayashi (National Institute of Advanced Industrial Science and Technology)
- 140 Differentiable Unrolled Alternating Direction Method of Multipliers for OneNet**
Zoltán Milacski (Eötvös Loránd University), Barnabas Póczos (Carnegie Mellon University), Andras Lorincz (Eötvös Loránd University)
- 141 Graph-based Knowledge Distillation by Multi-head Attention Network**
Seunghyun Lee (Inha University), Byung Cheol Song (Inha University)
- 142 Convolutional CRFs for Semantic Segmentation**
Marvin Teichmann (University of Cambridge), Roberto Cipolla (University of Cambridge)
- 143 Group Based Deep Shared Feature Learning for Fine-grained Image Classification**
Xuelu Li (The Pennsylvania State University), Vishal Monga (Pennsylvania State University)
- 144 Addressing Data Bias Problems for Chest X-ray Image Report Generation**
Philipp Harzig (University of Augsburg), Yan-Ying Chen (FX Pal), Francine Chen (FX Palo Alto Laboratory), Rainer Lienhart (Universität Augsburg)

19:00 – late Dinner

SU

Wednesday 11 September

08:00 – 09:00 Registration

09:00 – 10:00 Keynote

SME

Automatic Understanding of the Visual World

Prof. Cordelia Schmid (Inria)

Sponsored by Scape

10:00 – 11:00 Spotlights (Session 2)

SME

- 145 Object Affordances Graph Network for Action Recognition**

Haoliang Tan (Xi'an Jiaotong University), Le Wang (Xi'an Jiaotong University), Qilin Zhang (HERE Technologies), Zhanning Gao (Alibaba Group), Nanning Zheng (Xi'an Jiaotong University), Gang Hua (Wormpex AI Research)

- 146 Image Captioning with Unseen Objects**
 Berkan Demirel (HAVELSAN Inc. & METU), Ramazan Gokberk Cinbis (METU), Nazli Ikizler-Cinbis (Hacettepe University)
- 147 Residual Multiscale Based Single Image Deraining**
 Yupei Zheng (Beijing Jiaotong University), Xin Yu (Australian National University), Miaomiao Liu (Australian National University), Shunli Zhang (Beijing Jiaotong University)
- 148 Open-set Recognition of Unseen Macromolecules in Cellular Electron Cryo-Tomograms by Soft Large Margin Centralized Cosine Loss**
 Xuefeng Du (Xi'an Jiaotong University), Xiangrui Zeng (Carnegie Mellon University), Bo Zhou (Yale University), Alex Singh (Carnegie Mellon University), Min Xu (Carnegie Mellon University)
- 149 Learnable Gated Temporal Shift Module for Free-form Video Inpainting**
 Ya-Liang Chang (National Taiwan University), Zhe Yu Liu (National Taiwan University), Kuan-Ying Lee (National Taiwan University), Winston Hsu (National Taiwan University)
- 150 MS-GAN: Text to Image Synthesis with Attention-Modulated Generators and Similarity-aware Discriminators**
 Fengling Mao (Chinese Academy of Sciences), Bingpeng Ma (Chinese Academy of Sciences), Hong Chang (Chinese Academy of Sciences), Shiguang Shan (Chinese Academy of Sciences), Xilin Chen (Chinese Academy of Sciences)
- 151 Unsupervised and Explainable Assessment of Video Similarity**
 Konstantinos Papoutsakis (University of Crete & ICS-FORTH, Greece), Antonis Argyros (CSD-UOC and ICS-FORTH)
- 152 AutoCorrect: Deep Inductive Alignment of Noisy Geometric Annotations**
 Honglie Chen (University of Oxford), Weidi Xie (University of Oxford), Andrea Vedaldi (University of Oxford), Andrew Zisserman (University of Oxford)
- 153 Adaptive Compression-based Lifelong Learning**
 Shivangi Srivastava (Wageningen University and Research), Maxim Berman (KU Leuven), Matthew Blaschko (KU Leuven), Devis Tuia (Wageningen University and Research)
- 154 TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition**
 Bishay Mina (Queen Mary University London), Georgios Zoumpourlis (Queen Mary University of London), Ioannis Patras (Queen Mary University of London)
- 155 Adaptive Lighting for Data-Driven Non-Line-of-Sight 3D Localization and Object Identification**
 Sreenithy Chandran (Arizona State University), Suren Jayasuriya (Arizona State University)
- 156 Hybrid Deep Network for Anomaly Detection**
 Trong Nguyen Nguyen (University of Montreal), Jean Meunier (University of Montreal)
- 157 Fast and Multilevel Semantic-Preserving Discrete Hashing**
 Wanqian Zhang (Chinese Academy of Sciences), Dayan Wu (Chinese Academy of Sciences), Jing Liu (Chinese Academy of Sciences), Bo Li (Chinese Academy of Sciences), Xiaoyan Gu (Chinese Academy of Sciences), Weiping Wang (Chinese Academy of Sciences), Dan Meng (Chinese Academy of Sciences)
- 158 Mining Discriminative Food Regions for Accurate Food Recognition**
 Jianing Qiu (Imperial College London), Po Wen Lo (Imperial College London), Yingnan Sun (Imperial College London), Siyao Wang (Imperial College London), Benny Lo (Imperial College London)
- 159 Attentional demand estimation with attentive driving models**
 Petar Palasek (MindVisionLabs), Nilli Lavie (University College London, MindVisionLabs), Luke Palmer (MindVisionLabs)

160 Generalised Visual Microphone

Juhyun Ahn (SUALAB)

11:00 – 11:45 Tea Break

SU

11:45 – 13:15 Oral Presentations:

SME

3D Computer Vision

Chair: Majid Mirmehdi

Sponsored by Roke

161 End-to-End 3D Hand Pose Estimation from Stereo Cameras

Yuncheng Li (Snap Inc.), Zehao Xue (Snap Inc.), Yingying Wang (Snap Inc.), Lihao Ge (Nanyang Technological University), Zhou Ren (Wormpex AI Research), Jonathan Rodriguez (Snap Inc.)

162 Triangulation: Why Optimize?

Seong Hun Lee (University of Zaragoza), Javier Civera (Universidad de Zaragoza)

163 Single-view Object Shape Reconstruction Using Deep Shape Prior and Silhouette

Kejie Li (University of Adelaide), Ravi Garg (University of Adelaide), Ming Cai (The University of Adelaide), Ian Reid (University of Adelaide)

164 Learning Embedding of 3D models with Quadric Loss

Nitin Agarwal (Department of Computer Science, UC-Irvine), Sungeui Yoon (Korea Advanced Institute of Science and Technology), M Gopi (University of California, Irvine)

165 Probabilistic Reconstruction Networks for 3D Shape Inference from a Single Image

Roman Klovov (Inria), Jakob Verbeek (Inria), Edmond Boyer (Inria)

166 Optimal Multi-view Correction of Local Affine Frames

Iván Eichhardt (MTA SZTAKI), Dániel Baráth (MTA SZTAKI, CMP Prague)

13:15 – 14:00 Lunch

SU

14:00 – 16:15 Posters (Session 2)

SME

167–281, see poster abstracts on pages 29–144.**16:15 – 18:15 Oral Presentations:**

SME

Objects, Segmentation, Textures, and Colours

Chair: Roy Davies

Sponsored by Apple

282 Sensor-Independent Illumination Estimation for DNN Models

Mahmoud Afifi (York University), Michael Brown (York University)

283 Robust Synthesis of Adversarial Visual Examples Using a Deep Image Prior

Thomas Gittings (University of Surrey), Steve Schneider (University of Surrey), John Collomosse (University of Surrey)

- 284 Towards Weakly Supervised Semantic Segmentation in 3D Graph-Structured Point Clouds of Wild Scenes**
 Haiyan Wang (City University of New York), Xuejian Rong (City University of New York), Liang Yang (City University of New York), YingLi Tian (City University of New York)
- 285 Orthographic Feature Transform for Monocular 3D Object Detection**
 Thomas Roddick (University of Cambridge), Alex Kendall (University of Cambridge), Roberto Cipolla (University of Cambridge)
- 286 Texel-Att: Representing and Classifying Element-Based Textures by Attributes**
 Marco Godi (University of Verona), Christian Joppi (University of Verona), Andrea Giachetti (University of Verona), Fabio Pellacini (Sapienza University of Rome), Marco Cristani (University of Verona)
- 287 Content and Colour Distillation for Learning Image Translations with the Spatial Profile Loss**
 Saquib Sarfraz (Karlsruhe Institute of Technology), Constantin Seibold (Karlsruhe Institute of Technology), Haroon Khalid (Karlsruhe Institute of Technology), Rainer Stiefelhagen (Karlsruhe Institute of Technology)
- 288 An Empirical Study on Leveraging Scene Graphs for Visual Question Answering**
 Cheng Zhang (Ohio State University), Wei-Lun Chao (Cornell University), Dong Xuan (Ohio State University)
- 289 Fast-SCNN: Fast Semantic Segmentation Network**
 Rudra Poudel (Toshiba Research Europe, Ltd.), Stephan Liwicki (Toshiba Research Europe, Ltd.), Roberto Cipolla (University of Cambridge)

19:00 – late Banquet

NMC

Thursday 12 September

08:00 – 09:00 Registration

09:00 – 10:00 Keynote

SME

Dissecting Neural Nets

Prof. Antonio Torralba (MIT)

Sponsored by Amazon

10:00 – 11:00 Oral Presentations:

SME

Motion and Flow

Chair: Dave Marshall

Sponsored by Huawei

- 290 Relation-aware Multiple Attention Siamese Networks for Robust Visual Tracking**

Fangyi Zhang (Chinese Academy of Sciences), Bingpeng Ma (Chinese Academy of Sciences), Hong Chang (Chinese Academy of Sciences), Shiguang Shan (Chinese Academy of Sciences), Xilin Chen (Institute of Computing Technology, Chinese Academy of Sciences)

291 Spatial Transformer Spectral Kernels for Deformable Image Registration
 Ebrahim Al Safadi (Oregon Health and Science University), Xubo Song (Oregon Health and Science University)

292 Tracking the Known and the Unknown by Leveraging Semantic Information
 Ardhendu Shekhar Tripathi (ETH Zurich), Martin Danelljan (ETH Zurich), Luc Van Gool (ETH Zurich), Radu Timofte (ETH Zurich)

293 Tracking Holistic Object Representations
 Axel Sauer (Technical University of Munich), Elie Aljalbout (Technical University of Munich), Sami Haddadin (Technical University of Munich)

11:00 – 11:45 Tea Break

SU

11:45 – 13:15 Oral Presentations:

SME

Video Analysis

Chair: Paul Rosin

Sponsored by Intel

294 Geometry-Aware Video Object Detection for Static Cameras
 Dan Xu (University of Oxford), Weidi Xie (University of Oxford), Andrew Zisserman (University of Oxford)

295 Spatio-temporal Relational Reasoning for Video Question Answering
 Gursimran Singh (University of British Columbia), Leonid Sigal (University of British Columbia), Jim Little (University of British Columbia)

296 Mutual Suppression Network for Video Prediction using Disentangled Features
 Jungbeom Lee (Seoul National University), Jangho Lee (Seoul National University), Sungmin Lee (Seoul National University), Sungroh Yoon (Seoul National University)

297 Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace
 Dimitrios Kollias (Imperial College London), Stefanos Zafeiriou (Imperial College London)

298 Forecasting Future Action Sequences with Neural Memory Networks
 Harshala Gammulle (Queensland University of Technology), Simon Denman (Queensland University of Technology), Sridha Sridharan (Queensland University of Technology), Clinton Fookes (Queensland University of Technology)

299 Self-supervised Video Representation Learning for Correspondence Flow
 Zihang Lai (University of Oxford), Weidi Xie (University of Oxford)

13:15 – 14:00 Lunch

SU

14:00 – 15:45 Workshops

SME

15:45 – 16:30 Tea Break

SU

16:30 – 18:00 Workshops

SME

Index

The following index maps **presentation numbers** to pages on which the corresponding abstracts can be found.

1	38	30	34	59	49	88	62
2	51	31	35	60	49	89	63
3	81	32	35	61	50	90	63
4	72	33	36	62	50	91	64
5	54	34	36	63	51	92	64
6	65	35	37	64	51	93	65
7	41	36	37	65	52	94	66
8	85	37	38	66	52	95	66
9	33	38	38	67	53	96	67
10	58	39	39	68	53	97	67
11	85	40	39	69	54	98	68
12	59	41	40	70	54	99	68
13	67	42	40	71	55	100	69
14	43	43	41	72	55	101	69
15	44	44	41	73	55	102	70
16	63	45	42	74	56	103	71
17	121	46	42	75	57	104	71
18	46	47	43	76	57	105	72
19	46	48	43	77	58	106	72
20	47	49	44	78	58	107	72
21	48	50	44	79	58	108	73
22	49	51	45	80	59	109	73
23	31	52	45	81	59	110	74
24	31	53	46	82	60	111	74
25	32	54	46	83	60	112	75
26	32	55	47	84	60	113	75
27	33	56	47	85	61	114	76
28	33	57	48	86	61	115	76
29	33	58	48	87	62	116	76

117 77	157 105	197 102	237 122
118 78	158 127	198 103	238 122
119 78	159 94	199 103	239 123
120 79	160 102	200 104	240 123
121 79	161 76	201 104	241 124
122 80	162 36	202 104	242 124
123 80	163 36	203 105	243 125
124 81	164 37	204 105	244 125
125 81	165 37	205 106	245 126
126 82	166 38	206 106	246 126
127 82	167 87	207 107	247 127
128 83	168 88	208 107	248 127
129 83	169 88	209 108	249 128
130 84	170 89	210 108	250 128
131 84	171 89	211 109	251 129
132 85	172 90	212 109	252 129
133 85	173 90	213 110	253 130
134 86	174 91	214 110	254 130
135 86	175 91	215 111	255 131
136 87	176 92	216 111	256 131
137 93	177 92	217 112	257 132
138 122	178 93	218 112	258 132
139 122	179 93	219 113	259 133
140 47	180 94	220 113	260 133
141 48	181 94	221 114	261 134
142 134	182 95	222 114	262 134
143 123	183 95	223 115	263 135
144 49	184 95	224 115	264 135
145 88	185 96	225 116	265 136
146 124	186 96	226 116	266 136
147 103	187 97	227 117	267 137
148 114	188 97	228 117	268 137
149 103	189 98	229 118	269 138
150 104	190 98	230 118	270 138
151 90	191 99	231 119	271 139
152 126	192 100	232 119	272 139
153 132	193 100	233 120	273 140
154 92	194 101	234 120	274 140
155 96	195 101	235 121	275 141
156 140	196 102	236 121	276 141

277 142	283 108	289 130	295 137
278 142	284 129	290 81	296 137
279 142	285 109	291 82	297 68
280 143	286 107	292 82	298 87
281 144	287 96	293 83	299 86
282 95	288 109	294 121	

3D Computer Vision

23 Pixel-Wise Confidences for Stereo Disparities Using Recurrent Neural Networks

Muhammad Shahzeb Khan Gul (Fraunhofer IIS), Michel Bätz (Fraunhofer IIS), Joachim Keinert (Fraunhofer IIS)

One of the inherent problems with stereo disparity estimation algorithms is the lack of reliability information for the computed disparities. As a consequence, errors from the initial disparity maps are propagated to the following processing steps such as view rendering. Nowadays, confidence measures belong to the most popular techniques because of their capability to detect disparity outliers. Recently, convolutional neural network based confidence measures achieved best results by directly processing initial disparity maps. In contrast to existing convolutional neural network based methods, we propose a novel recurrent neural network architecture to compute confidences for different stereo matching algorithms. To maintain a low complexity the confidence for a given pixel is purely computed from its associated matching costs without considering any additional neighbouring pixels. As compared to the state-of-the-art confidence prediction methods leveraging convolutional neural networks, the proposed network is simpler and smaller in terms of size (reduction of the number of trainable parameters by almost 3-4 orders of magnitude). Moreover, the experimental results on three well-known datasets as well as with two popular stereo algorithms clearly highlight that the proposed approach outperforms state-of-the-art confidence estimation techniques.

24 Pan-tilt-zoom SLAM for Sports Videos

Jikai Lu (Zhejiang University), Jianhui Chen (University of British Columbia), Jim Little (University of British Columbia, Canada)

We present an online SLAM system specifically designed to track pan-tilt-zoom (PTZ) cameras in highly dynamic sports such as basketball and soccer games. In these games, PTZ cameras rotate very fast and players cover large image areas. To overcome these challenges, we propose to use

a novel camera model for tracking and to use rays as landmarks in mapping. Rays overcome the missing depth in pure-rotation cameras. We also develop an online pan-tilt forest for mapping and introduce moving objects (players) detection to mitigate negative impacts from foreground objects. We test our method on both synthetic and real datasets. The experimental results show the superior performance of our method over previous methods for online PTZ camera pose estimation.

25 An Evaluation of Feature Matchers for Fundamental Matrix Estimation

JiaWang Bian (The University of Adelaide), Yu-Huan Wu (Nankai University), Ji Zhao (TuSimple), Yun Liu (Nankai University), Le Zhang (Institute for Infocomm Research Agency for Science, Technology and Research (ASTAR)), Ming-Ming Cheng (Nankai University), Ian Reid (University of Adelaide)

Matching two images while estimating their relative geometry is a key step in many computer vision applications. For decades, a well-established pipeline, consisting of SIFT, RANSAC, and 8-point algorithm, has been used for this task. Recently, many new approaches were proposed and shown to outperform previous alternatives on standard benchmarks, including the learned features, correspondence pruning algorithms, and robust estimators. However, whether it is beneficial to incorporate them into the classic pipeline is less-investigated. To this end, we are interested in i) evaluating the performance of these recent algorithms in the context of image matching and epipolar geometry estimation, and ii) leveraging them to design more practical registration systems. The experiments are conducted in four large-scale datasets using strictly defined evaluation metrics, and the promising results provide insight into which algorithms suit which scenarios. According to this, we propose three high-quality matching systems and a Coarse-to-Fine RANSAC estimator. They show remarkable performances and have potentials to a large part of computer vision tasks. To facilitate future research, the full evaluation pipeline and the proposed methods are made publicly available.

26 A Simple Direct Solution to the Perspective-Three-Point Problem

Gaku Nakano (NEC Corporation)

This paper proposes a new direct solution to the perspective-three-point (P3P) problem based on an algebraic approach. The proposed method represents the rotation matrix as a function of distances from the camera center to three 3D points, then, finds the distances by utilizing the orthogonal constraints of the rotation matrix. The formulation can be simply written because it relies only on some simple concepts of linear algebra. According to synthetic data evaluations, the proposed method gives the

second-best performance against the state-of-the-art methods on both numerical accuracy and computational efficiency. In particular, the proposed method is the fastest among the quartic-equation based solvers. Moreover, the experimental results imply that the P3P problem still has an arguable issue on numerical stability regarding a point distribution and a camera pose.

9 Adversarial Examples for Handcrafted Features

27

Muhammad Latif Anjum (NUST), Zohaib Ali (NUST), Wajahat Hussain (NUST - SEecs)

Adversarial examples have exposed the weakness of deep networks. Careful modification of the input fools the network completely. Little work has been done to expose the weakness of handcrafted features in adversarial settings. In this work, we propose novel adversarial perturbations for handcrafted features. Pixel level analysis of handcrafted features reveals simple modifications which considerably degrade their performance. These perturbations generalize over different features, viewpoint and illumination changes. We demonstrate successful attack on several well known pipelines (SLAM, visual odometry, SfM etc.). Extensive evaluation is presented on multiple public benchmarks.

28 Physical Cue based Depth-Sensing by Color Coding with Deaberration Network

Nao Mishima (Toshiba Research and Development Center), Tatsuo Kozakaya (Toshiba), Akihisa Moriya (Toshiba), Ryuzo Okdata (Toshiba), Shinsaku Hiura (University of Hyogo)

Color-coded aperture (CCA) methods can physically measure the depth of a scene given by physical cues from a single-shot image of a monocular camera. However, they are vulnerable to actual lens aberrations in real scenes because they assume an ideal lens for simplifying algorithms. In this paper, we propose physical cue-based deep learning for CCA photography. To address actual lens aberrations, we developed a deep deaberration network (DDN) that is additionally equipped with a self-attention mechanism of position and color channels to efficiently learn the lens aberration. Furthermore, a new Bayes L1 loss function based on Bayesian deep learning enables to handle the uncertainty of depth estimation more accurately. Quantitative and qualitative comparisons demonstrate that our method is superior to conventional methods including real outdoor scenes. Furthermore, compared to a long-baseline stereo camera, the proposed method provides an error-free depth map at close range, as there is no blind spot between the left and right cameras.

29 Merge-SfM: Merging Partial Reconstructions

Meiling Fang (Fraunhofer IOSB), Thomas Pollok (Fraunhofer IOSB), Chengchao Qu (Fraun-

hofer IOSB)

Recovering a 3D scene from unordered photo collections is a long-studied topic in computer vision. Existing reconstruction pipelines, both incremental and global, have already achieved remarkable results. This paper addresses the problem of fusing multiple existing partial 3D reconstructions, in particular finding the overlapping regions and transformations (7 DOF) between partial reconstructions. Unlike the previous methods which have to take the entire epipolar geometry (EG) graph as the input and reconstruct the scene, we propose an approach that reuses the existing reconstructed 3D models as input and merges them by utilizing all the internal information to avoid repeated work. This approach is divided into two steps. The first is to find overlapping areas between partial reconstructions based on Fisher similarity lists. Then, based on those overlaps, pairwise rotation between partial reconstructions is estimated by solving an ℓ_1 approximation optimization problem. After global rotation estimation, translation and scale between each pair of partial reconstructions are computed simultaneously in a global manner. In order to find the optimal transformation path, the maximal spanning tree (MST) is constructed in the second stage. Our approach is evaluated on diverse challenging public datasets and compared to state-of-the-art Structure from Motion (SfM) methods. Experiments show that our merging approach achieves high computational efficiency while preserving similar reconstruction accuracy and robustness. In addition, our method has superior extensibility which can add partial 3D reconstructions gradually to extend an existing 3D scene.

30 **Semi-supervised Macromolecule Structural Classification in Cellular Electron Cryo-Tomograms using 3D Autoencoding Classifier**

Siyuan Liu (Carnegie Mellon University), Xuefeng Du (Xi'an Jiaotong University), Rong Xi (Carnegie Mellon University), Fuya Xu (Carnegie Mellon University), Xiangrui Zeng (Carnegie Mellon University), Bo Zhou (Yale University), Min Xu (Carnegie Mellon University)

Recent advances in the Cellular Electron Cryo-Tomography (CECT) imaging technique have enabled the 3D visualization of macromolecules and other sub-cellular components in single cells in their near-native state. Automatic structural classification of macromolecules is increasingly desirable for researchers to better study and understand the features of different macromolecular complexes. However, accurate classification of macromolecular complexes is still impeded by the lack of annotated training data due to the limited expert resource for labeling full datasets. In this paper, we introduce a semi-supervised classification framework to reduce annotation burden in the macromolecule structural classification tasks. Specifically, we propose a 3D autoencoding classifier framework for simultaneous macromolecule structural reconstruction and classifica-

tion. Our framework jointly optimizes two branches of network using both labeled and unlabeled data during training phase. Extensive experiments demonstrate the effectiveness of our approach against other semi-supervised classification approaches on both real and simulated datasets. Our approach also achieves competitive results in terms of macromolecule reconstruction. To our best knowledge, this is the first work to address the task of semi-supervised macromolecule structural classification in CECT.

31 **Large scale joint semantic re-localisation and scene understanding via globally unique instance coordinate regression**

Ignas Budvytis (Department of Engineering, University of Cambridge), Marvin Teichmann (Machine Intelligence Laboratory, Cambridge University Department of Engineering), Tomas Vojir (University of Cambridge), Roberto Cipolla (University of Cambridge)

In this work we present a novel approach to joint semantic localisation and scene understanding. Our work is motivated by the need for localisation algorithms which not only predict 6-DoF camera pose but also simultaneously recognise surrounding objects and estimate 3D geometry. Such capabilities are crucial for computer vision guided systems which interact with the environment: autonomous driving, augmented reality and robotics. In particular, we propose a two step procedure. During the first step we train a convolutional neural network to jointly predict per-pixel globally unique instance labels and corresponding local coordinates for each instance of a static object (e.g. a building). During the second step we obtain scene coordinates by combining object center coordinates and local coordinates and use them to perform 6-DoF camera pose estimation. We evaluate our approach on real world (CamVid-360) and artificial (SceneCity) autonomous driving datasets. We obtain smaller mean distance and angular errors than state-of-the-art 6-DoF pose estimation algorithms based on direct pose regression and pose estimation from scene coordinates on all datasets. Our contributions include: (i) a novel formulation of scene coordinate regression as two separate tasks of object instance recognition and local coordinate regression and a demonstration that our proposed solution allows to predict accurate 3D geometry of static objects and estimate 6-DoF pose of camera on (ii) maps larger by several orders of magnitude than previously attempted by scene coordinate regression methods, as well as on (iii) lightweight, approximate 3D maps built from 3D primitives such as building-aligned cuboids.

32 **Matching Features without Descriptors: Implicitly Matched Interest Points**

Titus Cieslewski (University of Zurich & ETH Zurich), Michael Bloesch (Deepmind), Davide

Scaramuzza (University of Zurich & ETH Zurich)

The extraction and matching of interest points is a prerequisite for many geometric computer vision problems. Traditionally, matching has been achieved by assigning descriptors to interest points and matching points that have similar descriptors. In this paper, we propose a method by which interest points are instead already implicitly matched at detection time. With this, descriptors do not need to be calculated, stored, communicated, or matched any more. This is achieved by a convolutional neural network with multiple output channels and can be thought of as a collection of a variety of detectors, each specialized to specific visual features. This paper describes how to design and train such a network in a way that results in successful relative pose estimation performance despite the limitation on interest point count. While the overall matching score is slightly lower than with traditional methods, the approach is descriptor free and thus enables localization systems with a significantly smaller memory footprint and multi-agent localization systems with lower bandwidth requirements. The network also outputs the confidence for a specific interest point resulting in a valid match. We evaluate performance relative to state-of-the-art alternatives.

33 **Triangulation: Why Optimize?**

162

Seong Hun Lee (University of Zaragoza), Javier Civera (Universidad de Zaragoza)

For decades, it has been widely accepted that the gold standard for two-view triangulation is to minimize the cost based on reprojection errors. In this work, we challenge this idea. We propose a novel alternative to the classic midpoint method that leads to significantly lower 2D errors and parallax errors. It provides a numerically stable closed-form solution based solely on a pair of backprojected rays. Since our solution is rotationally invariant, it can also be applied for fisheye and omnidirectional cameras. We show that for small parallax angles, our method outperforms the state-of-the-art in terms of combined 2D, 3D and parallax accuracy, while achieving comparable speed.

34 **Single-view Object Shape Reconstruction Using Deep Shape Prior and Silhouette**

163

Kejie Li (University of Adelaide), Ravi Garg (University of Adelaide), Ming Cai (The University of Adelaide), Ian Reid (University of Adelaide)

3D shape reconstruction from a single image is a highly ill-posed problem. Modern deep learning based systems try to solve this problem by learning an end-to-end mapping from image to shape via a deep network. In this paper, we aim to solve this problem via an online optimization framework

inspired by traditional methods. Our framework employs a deep autoencoder to learn a set of latent codes of 3D object shapes, which are fitted by a probabilistic shape prior using Gaussian Mixture Model (GMM). At inference, the shape and pose are jointly optimized guided by both image cues and deep shape prior without relying on an initialization from any trained deep nets. Surprisingly, our method achieves comparable performance to state-of-the-art methods even without training an end-to-end network, which shows a promising step in this direction.

35 Learning Embedding of 3D models with Quadric Loss

164

Nitin Agarwal (Department of Computer Science, UC-Irvine), Sungeui Yoon (Korea Advanced Institute of Science and Technology), M Gopi (University of California, Irvine)

Sharp features such as edges and corners play an important role in the perception of 3D models. In order to capture them better, we propose quadric loss, a point-surface loss function, which minimizes the quadric error between the reconstructed points and the input surface. Computation of Quadric loss is easy, efficient since the quadric matrices can be computed apriori, and is fully differentiable, making quadric loss suitable for training point and mesh based architectures. Through extensive experiments we show the merits and demerits of quadric loss. When combined with Chamfer loss, quadric loss achieves better reconstruction results as compared to any one of them or other point-surface loss functions.

36 Probabilistic Reconstruction Networks for 3D Shape Inference from a Single Image

165

Roman Klokov (Inria), Jakob Verbeek (Inria), Edmond Boyer (Inria)

We study end-to-end learning strategies for 3D shape inference from images, in particular from a single image. Several approaches in this direction have been investigated that explore different shape representations and suitable learning architectures. We focus instead on the underlying probabilistic mechanisms involved and contribute a more principled probabilistic inference-based reconstruction framework, which we coin Probabilistic Reconstruction Networks. This framework expresses image conditioned 3D shape inference through a family of latent variable models, and naturally decouples the choice of shape representations from the inference itself. Moreover, it suggests different options for the image conditioning and allows training in two regimes, using either Monte Carlo or variational approximation of the marginal likelihood. Using our Probabilistic Reconstruction Networks we obtain single image 3D reconstruction results that set a new state of the art on the ShapeNet dataset in terms of the intersection over union and earth mover's distance evaluation metrics. In-

terestingly, we obtain these results using a basic voxel grid representation, improving over recent work based on finer point cloud or mesh based representations.

37 Optimal Multi-view Correction of Local Affine Frames

166

Iván Eichhardt (MTA SZTAKI), Dániel Baráth (MTA SZTAKI, CMP Prague)

A method is proposed for correcting the parameters of a sequence of detected local affine frames through multiple views. The technique requires the epipolar geometry to be pre-estimated between each image pair. It exploits the constraints which the camera movement implies, in order to apply a closed-form correction to the parameters of the input affinities. Also, it is shown that the rotations and scales obtained by partially affine-covariant detectors, e.g. AKAZE or SIFT, can be upgraded to be full affine frames by the proposed algorithm. It is validated both in synthetic experiments and on publicly available real-world datasets that the method almost always improves the output of the evaluated affine-covariant feature detectors. As a by-product, these detectors are compared and the ones obtaining the most accurate affine frames are reported. To demonstrate the applicability in real-world scenarios, we show that the proposed technique improves the accuracy of pose estimation for a camera rig, surface normal and homography estimation.

1 Adversarial View-Consistent Learning for Monocular 38 Depth Estimation

Yixuan Liu (Tsinghua University), Yuwang Wang (Microsoft Research), Shengjin Wang (Tsinghua University)

This paper addresses the problem of Monocular Depth Estimation (MDE). Existing approaches on MDE usually model it as a pixel-level regression problem, ignoring the underlying geometry property. We empirically find this may result in sub-optimal solution: while the predicted depth map presents small loss value in one specific view, it may exhibit large loss if viewed in different directions. In this paper, inspired by multi-view stereo (MVS), we propose an Adversarial View-Consistent Learning (AVCL) framework to force the estimated depth map to be all reasonable viewed from multiple views. To this end, we first design a differentiable depth map warping operation, which is end-to-end trainable, and then propose a pose generator to generate novel views for a given image in an adversarial manner. Collaborating with the differentiable depth map warping operation, the pose generator encourages the depth estimation network to learn from hard views, hence produce view-consistent depth maps. We evaluate our method on NYU Depth V2 dataset and the experimental results show

promising performance gain upon state-of-the-art MDE approaches.

39 Few-Shot Viewpoint Estimation

Hung-Yu Tseng (University of California, Merced), Shalini De Mello (NVIDIA Research), Jonathan Tremblay (NVIDIA), Sifei Liu (NVIDIA), Stan Birchfield (Clemson University), Ming-Hsuan Yang (University of California at Merced), Jan Kautz (NVIDIA)

Viewpoint estimation for known categories of objects has been improved significantly thanks to deep networks and large datasets, but generalization to unknown categories is still very challenging. With an aim towards improving performance on unknown categories, we introduce the problem of category-level few-shot viewpoint estimation. We design a novel framework to successfully train viewpoint networks for new categories with few examples (10 or less). We formulate the problem as one of learning to estimate category-specific 3D canonical shapes, their associated depth estimates, and semantic 2D keypoints. We apply meta-learning to learn weights for our network that are amenable to category-specific few-shot fine-tuning. Furthermore, we design a flexible meta-Siamese network that maximizes information sharing during meta-learning. Through extensive experimentation on the ObjectNet3D and Pascal3D+ benchmark datasets, we demonstrate that our framework, which we call MetaView, significantly outperforms fine-tuning the state-of-the-art models with few examples, and that the specific architectural innovations of our method are crucial to achieving good performance.

40 Differentiable Fixed-Rank Regularisation using Bilinear Parameterisation

Marcus Valtonen Örnå (Lund University), Carl Olsson (Lund University), Anders Heyden (LTH)

Low rank structures are present in many applications of computer vision and machine learning. A popular approach consists of explicitly parameterising the set or matrices with sought rank, leading to a bilinear factorisation, reducing the problem to find the bilinear factors. While such an approach can be efficiently implemented using second-order methods, such as Levenberg-Marquardt (LM) or Variable Projection (VarPro), it suffers from the presence of local minima, which makes theoretical optimality guarantees hard to derive.

Another approach is to penalise non-zero singular values to enforce a low-rank structure. In certain cases, global optimality guarantees are known; however, such methods often lead to non-differentiable (and even discontinuous) objectives, for which it is necessary to use subgradient methods and splitting schemes. If the objective is complex, such as in structure from motion, the convergence rates for such methods can be very slow.

In this paper we show how optimality guarantees can be lifted to meth-

ods that employ bilinear parameterisation when the sought rank is known. Using this approach the best of two worlds are combined: optimality guarantees and superior convergence speeds. We compare the proposed method to state-of-the-art solvers for prior-free non-rigid structure from motion.

41 Mitigating the Hubness Problem for Zero-Shot Learning of 3D Objects

Ali Cheraghian (Australian National University), Shafin Rahman (Australian National University), Dylan Campbell (Australian National University), Lars Petersson (Data61/CSIRO)

The development of advanced 3D sensors has enabled many objects to be captured in the wild at a large scale, and a 3D object recognition system may therefore encounter many objects for which the system has received no training. Zero-Shot Learning (ZSL) approaches can assist such systems in recognizing previously unseen objects. Applying ZSL to 3D point cloud objects is an emerging topic in the area of 3D vision, however, a significant problem that ZSL often suffers from is the so-called hubness problem, which is when a model is biased to predict only a few particular labels for most of the test instances. We observe that this hubness problem is even more severe for 3D recognition than for 2D recognition. One reason for this is that in 2D one can use pre-trained networks trained on large datasets like ImageNet, which produces high-quality features. However, in the 3D case there are no such large-scale, labelled datasets available for pre-training which means that the extracted 3D features are of poorer quality which, in turn, exacerbates the hubness problem. In this paper, we therefore propose a loss to specifically address the hubness problem. Our proposed method is effective for both Zero-Shot and Generalized Zero-Shot Learning, and we perform extensive evaluations on the challenging datasets ModelNet40, ModelNet10, McGill and SHREC2015. A new state-of-the-art result for both zero-shot tasks in the 3D case is established.

42 Optimising 3D-CNN Design towards Human Pose Estimation on Low Power Devices

Manolis Vasileiadis (Imperial College London), Christos-Savvas Bouganis (Imperial College London), Georgios Stavropoulos (Centre for Research and Technology, Hellas, Information Technologies Institute), Dimitrios Tzovaras (Centre for Research and Technology, Hellas)

3D CNN-based architectures have found application in a variety of 3D vision tasks, significantly outperforming earlier approaches. This increase in accuracy, however, has come at the cost of computational complexity, with deep learning models becoming more and more complex, requiring significant computational resources, especially in the case of 3D data. Meanwhile, the growing adoption of low power devices in various technology fields has shifted the research focus towards the implementation

of deep learning on systems with limited resources. While plenty of approaches have achieved promising results in terms of reducing the computational complexity in 2D tasks, their applicability in 3D-CNN designs has not been thoroughly researched. The current work aims at filling this void, by investigating a series of efficient CNN design techniques within the scope of 3D-CNNs, in order to produce guidelines for 3D-CNN design that can be applied to already established architectures, reducing their computational complexity. Following these guidelines, a computationally efficient 3D-CNN architecture for human pose estimation from 3D data is proposed, achieving comparable accuracy to the state-of-the-art. The proposed design guidelines are further validated within the scope of 3D object classification, achieving high accuracy results at a low computational cost.

7 DetectFusion: Detecting and Segmenting Both Known and Unknown Dynamic Objects in Real-time SLAM

Ryo Hachiuma (Keio University), Christian Pirchheim (Graz University of Technology), Dieter Schmalstieg (Graz University of Technology), Hideo Saito (Keio University)

We present DetectFusion, an RGB-D SLAM system that runs in real time and can robustly handle semantically known and unknown objects that can move dynamically in the scene. Our system detects, segments and assigns semantic class labels to known objects in the scene, while tracking and reconstructing them even when they move independently in front of the monocular camera. In contrast to related work, we achieve real-time computational performance on semantic instance segmentation with a novel method combining 2D object detection and 3D geometric segmentation. In addition, we propose a method for detecting and segmenting the motion of semantically unknown objects, thus further improving the accuracy of camera tracking and map reconstruction. We show that our method performs on par or better than previous work in terms of localization and object reconstruction accuracy, while achieving about 20 FPS even if the objects are segmented in each frame.

44 DublinCity: Annotated LiDAR Point Cloud and its Applications

S M Iman Zolanvari (Trinity College Dublin), Susana Ruano (Trinity College Dublin), Aakanksha Rana (Trinity College Dublin), Alan Cummins (Trinity College Dublin), Rogério Eduardo da Silva (University of Houston-Victoria), Morteza Rahbar (CAAD, ITA, ETH Zurich), Aljosa Smolic (Trinity College Dublin)

Scene understanding of full-scale 3D models of an urban area remains a challenging task. While advanced computer vision techniques offer cost-effective approaches to analyse 3D urban elements, a precise and densely labelled dataset is quintessential. The paper presents the first-ever labelled dataset for a highly dense Aerial Laser Scanning (ALS) point

cloud at city-scale. This work introduces a novel benchmark dataset that includes a manually annotated point cloud for over 260 million laser scanning points into 100'000 (approx.) assets from Dublin LiDAR point cloud (Laefer, et al) in 2015. Objects are labelled into 13 classes using hierarchical levels of detail from large (i.e. building, vegetation and ground) to refined (i.e. window, door and tree) elements. To validate the performance of our dataset, two different applications are showcased. Firstly, the labelled point cloud is employed for training Convolutional Neural Networks (CNNs) to classify urban elements. The dataset is tested on the well-known state-of-the-art CNNs (i.e. PointNet, PointNet++ and So-Net). Secondly, the complete ALS dataset is applied as detailed ground truth for city-scale image-based 3D reconstruction.

45 Single Image 3D Hand Reconstruction with Mesh Convolutions

Dominik Kulon (Imperial College London), Haoyang Wang (Imperial College London), Alp Guler (Ariel AI, Imperial College London), Michael Bronstein (Imperial College London), Stefanos Zafeiriou (Imperial College London)

Monocular 3D reconstruction of deformable objects, such as human body parts, has been typically approached by predicting parameters of heavy-weight linear models. In this paper, we demonstrate an alternative solution that is based on the idea of encoding images into a latent non-linear representation of meshes. The prior on 3D hand shapes is learned by training an autoencoder with intrinsic graph convolutions performed in the spectral domain. The pre-trained decoder acts as a non-linear statistical deformable model. The latent parameters that reconstruct the shape and articulated pose of hands in the image are predicted using an image encoder. We show that our system reconstructs plausible meshes and operates in real-time. We evaluate the quality of the mesh reconstructions produced by the decoder on a new dataset and show latent space interpolation results. Our code, data, and models will be made publicly available.

46 MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images

Ammar Qammar (CSD-UOC and ICS-FORTH), Antonis Argyros (CSD-UOC and ICS-FORTH)

We present MocapNET, an ensemble of SNN encoders that estimates the 3D human body pose based on 2D joint estimations extracted from monocular RGB images. MocapNET provides an efficient divide and conquer strategy for supervised learning. It outputs skeletal information directly into the BVH format which can be rendered in real-time or imported without any additional processing in most popular 3D animation software. The proposed architecture achieves 3D human pose estimations at state of the art rates of 400Hz using only CPU processing.

47 An Evaluation of Feature Encoding Techniques for Non-Rigid and Rigid 3D Point Cloud Retrieval

Sindhu Hegde (KLE Technological University), Shankar Gangisetty (KLE Technological University)

In this paper, we address the 3D point cloud based retrieval problem for both non-rigid and rigid 3D data. As powerful computation resources and scanning devices have led to an exponential growth of 3D point cloud data, retrieving the relevant 3D objects from databases is a challenging task. The local descriptors provide only the abstract representations that do not enable the exploration of shape variability to solve the 3D object retrieval problem. Thus, it is not just the local descriptors but also the encoding of local signatures into global descriptors which is of crucial importance for enhancing the performance. To create a compact shape signature that constitutes the 3D object as a whole, various encoding techniques have been proposed in the literature. The most popular among them are bag-of-features, Fisher vector and vector of locally aggregated descriptors. However evaluating the different encoding techniques and analyzing the critical aspects to boost the performance of 3D point cloud retrieval is still an unsolved problem. We propose to provide an exhaustive evaluation of the different encoding techniques when combined with local feature descriptors for solving non-rigid and rigid point cloud retrieval task. We fix improved wave kernel signature and metric tensor & Christoffel symbols local descriptors specifically built for non-rigid and rigid data as given in and respectively. We also present a consistent comparative analysis of our method with the existing benchmarks, the results of which illustrate the robustness of the proposed approach on point cloud data.

14 48 Perspective-n-Learned-Point: Pose Estimation from Relative Depth

Nathan Piasco (Univ. Bourgogne Franche-Comte), Désiré Sidibé (Université de Bourgogne), Cedric Demonceaux (Univ. Bourgogne Franche-Comte), Valérie Gouet-Brunet (LASTIG/IGN)

In this paper we present an online camera pose estimation method that combines Content-Based Image Retrieval (CBIR) and pose refinement based on a learned representation of the scene geometry extracted from monocular images. Our pose estimation method is two-step, we first retrieve an initial 6 Degrees of Freedom (DoF) location of an unknown-pose query by retrieving the most similar candidate in a pool of geo-referenced images. In a second time, we refine the query pose with a Perspective-n-Point (PnP) algorithm where the 3D points are obtained thanks to a generated depth map from the retrieved image candidate. We make our method fast and lightweight by using a common neural network architecture to generate the image descriptor for image indexing and the depth

map used to create the 3D points required in the PnP pose refinement step. We demonstrate the effectiveness of our proposal through extensive experimentation on both indoor and outdoor scenes, as well as generalisation capability of our method to unknown environment. Finally, we show how to deploy our system even if geometric information is missing to train our monocular-image-to-depth neural networks.

15 Joint Multi-view Texture Super-resolution and Intrinsic 49 Decomposition

Wei Dong (Carnegie Mellon University), Vagia Tsiminaki (ETH Zurich), Martin R. Oswald (ETH Zurich), Marc Pollefeys (ETH Zurich / Microsoft)

We aim to recover a high resolution texture representation of objects observed from multiple view points under varying lighting conditions. For many applications the lighting conditions need to be changed and thus require a texture decomposition into shading and albedo components. Both texture super-resolution and intrinsic texture decomposition have been separately studied in the literature. Yet, no method has investigated how these methods can be combined. We propose a framework for joint texture map super-resolution and intrinsic decomposition. To this end, we define shading and albedo maps of the 3D object as the intrinsic properties of its texture and introduce an image formation model to describe the physics of the image generation. Our approach accounts for surface geometry and camera calibration errors and is also applicable to spatio-temporal sequences. Our method achieves state-of-the-art results on a variety of datasets.

50 Learning Depth-aware Heatmaps for 3D Human Pose Es- timation in the Wild

Zerui Chen (Chinese Academy of Sciences), Yiru Guo (Beihang University), Yan Huang (Institute of Automation, Chinese Academy of Sciences), Liang Wang (NLPR, China)

In this paper, we explore to determine 3D human pose directly from monocular image data. While current state-of-the-art approaches employ the volumetric representation to predict per voxel likelihood for each human joint, the network output is memory-intensive, making it hard to function on mobile devices. To reduce the output dimension, we intend to decompose the volumetric representation into 2D depth-aware heatmaps and joint depth estimation. We propose to learn depth-aware 2D heatmaps via associative embeddings to reconstruct the connection between the 2D joint location and its corresponding depth. Our approach achieves a good trade-off between complexity and high performance. We conduct extensive experiments on the popular benchmark Human3.6M and advance the state-of-the-art accuracy for 3D human pose estimation in the wild.

Deep Learning for Vision

51 DwNet: Dense warp-based network for pose-guided human video generation

Polina Zablotnskaia (University of British Columbia), Aliaksandr Siarohin (University of Trento), Leonid Sigal (University of British Columbia), Bo Zhao (University of British Columbia)

Generation of realistic high-resolution videos of human subjects is a challenging and important task in computer vision. In this paper, we focus on human motion transfer - generation of a video depicting a particular subject, observed in a single image, performing a series of motions exemplified in an auxiliary (driving) video. Our GAN-based architecture DwNet leverages dense intermediate pose-guided representation and refinement process to warp the required subject appearance, in the form of the texture, from a source image into a desired pose. Temporal consistency is maintained by further conditioning the decoding process within a GAN on the previously generated frame. In this way a video is generated in an iterative and recurrent fashion. We illustrate the efficacy of our approach by showing state-of-the-art quantitative and qualitative performance on two benchmark datasets: TaiChi and Fashion Modeling. The latter is collected by us and will be made publicly available to the community

52 Annotation-free Quality Estimation of Food Grains using Deep Neural Network

Akankshya Kar (Samsung Research Institute Bangalore), Prakhar Kulshreshtha (Samsung Research Institute Bangalore), Ayush Agrawal (Samsung Research Institute Bangalore), Sandeep Palakkal (Samsung Electronics), Lokesh Boregowda (Samsung Research Institute Bangalore)

We propose a fast and accurate system for automatically estimating the quality of food grains on resource constrained portable devices using computer vision. We are motivated by an urgent need in India for grain quality estimation to ensure transparency in the agricultural supply chain and empower poor farmers to get the correct price for their crops. The system uses instance segmentation of touching grains, followed by classification of each grain according to E-NAM parameters. To the best of our knowledge, this is the first attempt to use Deep Learning to estimate quality of cluttered sample of grains using only mobile phone. Samples are collected from various Agricultural Produce Market Committee (APMC) yards, which are used to generate synthetic data to simulate realistic clutter of grains for training our instance-segmentation network. Novel augmentation techniques while training make the system robust to illumination changes. Our system obtains the state-of-the-art performance and has been tested in various locations in India. At a mAP score of 0.74 and classification accuracy 92%, our system takes less than 100s compared to 15 minutes of manual

quality estimation.

18 Embodied Vision-and-Language Navigation with Dy- 53 namic Convolutional Filters

Federico Landi (University of Modena and Reggio Emilia), Lorenzo Baraldi (University of Modena and Reggio Emilia), Massimiliano Corsini (University of Modena and Reggio Emilia), Rita Cucchiara (University of Modena and Reggio Emilia)

In Vision-and-Language Navigation (VLN), an embodied agent needs to reach a target destination with the only guidance of a natural language instruction. To explore the environment and progress towards the target location, the agent must perform a series of low-level actions, such as rotate, before stepping ahead. In this paper, we propose to exploit dynamic convolutional filters to encode the visual information and the lingual description in an efficient way. Differently from some previous works that abstract from the agent perspective and use high-level navigation spaces, we design a policy which decodes the information provided by dynamic convolution into a series of low-level, agent friendly actions. Results show that our model exploiting dynamic filters performs better than other architectures with traditional convolution, being the new state of the art for embodied VLN in the low-level action space. Additionally, we attempt to categorize recent work on VLN depending on their architectural choices and distinguish two main groups: we call them low-level actions and high-level actions models. To the best of our knowledge, we are the first to propose this analysis and categorization for VLN.

19 Accurate and Compact Convolutional Neural Networks 54 with Trained Binarization

Zhe Xu (City University of Hong Kong), Ray Cheung (City University of Hong Kong)

Although convolutional neural networks (CNNs) are now widely used in various computer vision applications, its huge resource demanding on parameter storage and computation makes the deployment on mobile and embedded devices difficult. Recently, binary convolutional neural networks are explored to help alleviate this issue by quantizing both weights and activations with only 1 single bit. However, there may exist a noticeable accuracy degradation when compared with full-precision models. In this paper, we propose an improved training approach towards compact binary CNNs with higher accuracy. Trainable scaling factors for both weights and activations are introduced to increase the value range. These scaling factors will be trained jointly with other parameters via backpropagation. Besides, a specific training algorithm is developed including tight approximation for derivative of discontinuous binarization function and modified L2 regularization acting on weight scaling factors. With these improvements, the binary CNN achieves 92.3% accuracy on CIFAR-10

with VGG-Small network. On ImageNet, our method also obtains 46.1% top-1 accuracy with AlexNet and 54.2% with Resnet-18 surpassing previous works.

20 **Show, Infer and Tell: Contextual Inference for Creative** 55 **Captioning**

Ankit Khare (University of Texas at Arlington), Manfred Huber (University of Texas at Arlington)

Several attention based encoder-decoder architectures have been geared towards the task of image captioning. Yet, the collocations and contextual inference seen in captions written by humans is not observed in the output of these systems e.g., if we see a lot of different vehicles on the road, we infer “traffic” and say “a lot of traffic on the road”. Further, “hallucination” of commonly seen concepts for fitting the language model is commonly observed in a lot of existing systems. For example, “a group of soldiers cutting a cake with a sword” would be hallucinated as “a boy cutting a cake with a knife”. In this work we construct two simultaneously learning channels, where first channel uses the mean-pooled image feature and learns to associate it with the most relevant words. The second channel, on the other hand, utilizes the spatial features belonging to salient image regions to learn to form meaningful collocations and perform contextual inference. This way, the final language model gets the opportunity to leverage the information from the two channels to learn to generate grammatically correct sentence structures which are more human-like and creative. Our novel “spatial image features to n-gram text features mapping” mechanism not only learns meaningful collocations but also verifies that the caption words correspond to the region(s) of the image, thereby avoiding “hallucination” by the model. We validate the effectiveness of our one pass system on the challenging MS-COCO image captioning benchmark, where our single-model achieves a new state-of-the art 126.3 CIDEr-D on the Karpathy split, and a competitive 124.1 CIDEr-D (c40) on the official server.

56 **Differentiable Unrolled Alternating Direction Method of** 140 **Multipliers for OneNet**

Zoltán Milacski (Eötvös Loránd University), Barnabas Póczos (Carnegie Mellon University), Andras Lorincz (Eötvös Loránd University)

Deep neural networks achieve state-of-the-art results on numerous image processing tasks, but this typically requires training problem-specific networks. Towards multi-task learning, the One Network to Solve Them All (OneNet) method was recently proposed that first pretrains an adversarial denoising autoencoder and subsequently uses it as the proximal operator in Alternating Direction Method of Multipliers (ADMM) solvers of multiple imaging problems. In this work, we highlight training and

ADMM convergence issues of OneNet, and resolve them by proposing an end-to-end learned architecture for training the two steps jointly using Unrolled Optimization with backpropagation. In our experiments, our solution achieves superior or on par results compared to the original OneNet and Wavelet sparsity on four imaging problems (pixelwise inpainting-denoising, blockwise inpainting, scattered inpainting and super resolution) on the MS-Celeb-1M and ImageNet data sets, even with a much smaller ADMM iteration count.

57 Graph-based Knowledge Distillation by Multi-head At- 141 tention Network

Seunghyun Lee (Inha University), Byung Cheol Song (Inha University)

Knowledge distillation (KD) is a technique to derive optimal performance from a small student network (SN) by distilling knowledge of a large teacher network (TN) and transferring the distilled knowledge to the small SN. Since a role of convolutional neural network (CNN) in KD is to embed a dataset so as to perform a given task well, it is very important to acquire knowledge that considers intra-data relations. Conventional KD methods have concentrated on distilling knowledge in data units. To our knowledge, any KD methods for distilling information in dataset units have not yet been proposed. Therefore, this paper proposes a novel method that enables distillation of dataset-based knowledge from the TN using an attention network. The knowledge of the embedding procedure of the TN is distilled to graph by multi-head attention (MHA), and multi-task learning is performed to give relational inductive bias to the SN. The MHA can provide clear information about the source dataset, which can greatly improve the performance of the SN. Experimental results show that the proposed method is 7.05% higher than the SN alone for CIFAR100, which is 2.46% higher than the state-of-the-art.

21 Push for Quantization: Deep Fisher Hashing 58

Yunqiang Li (Delft University of Technology), Wenjie Pei (Tencent), yufei zha (Air Force Engineering University), Jan van Gemert (Delft University of Technology)

Current massive datasets demand light-weight access for analysis. Discrete hashing methods are thus beneficial because they map high-dimensional data to compact binary codes that are efficient to store and process, while preserving semantic similarity. To optimize powerful deep learning methods for image hashing, gradient-based methods are required. Binary codes, however, are discrete and thus have no continuous derivatives. Relaxing the problem by solving it in a continuous space and then quantizing the solution is not guaranteed to yield separable binary codes. The quantiza-

tion needs to be included in the optimization. In this paper we push for quantization: We optimize maximum class separability in the binary space. To do so, we introduce a margin on distances between dissimilar image pairs as measured in the binary space. In addition to pair-wise distances, we draw inspiration from Fisher’s Linear Discriminant Analysis (Fisher LDA) to maximize the binary distances between classes and at the same time minimize the binary distance of images within the same class. Experimental results on CIFAR-10, NUS-WIDE and ImageNet100 show that our approach leads to compact codes and compares favorably to the current state of the art.

59 Addressing Data Bias Problems for Chest X-ray Image 144 Report Generation

Philipp Harzig (University of Augsburg), Yan-Ying Chen (FX Pal), Francine Chen (FX Palo Alto Laboratory), Rainer Lienhart (Universitat Augsburg)

Automatic medical report generation from chest X-ray images is one possibility for assisting doctors to reduce their workload. However, the different patterns and data distribution of normal and abnormal cases can bias machine learning models. Previous attempts did not focus on isolating the generation of the abnormal and normal sentences in order to increase the variability of generated paragraphs. To address this, we propose to separate abnormal and normal sentence generation by using two different word LSTMs in a hierarchical LSTM model. We conduct an analysis on the distinctiveness of generated sentences compared to the BLEU score, which increases when less distinct reports are generated. We hope our findings will help to encourage the development of new metrics to better verify methods of automatic medical report generation.

22 RecNets: Channel-wise Recurrent Convolutional Neural 60 Networks

George Retsinas (National Technical University of Athens), Athena Elafrou (National Technical University of Athens), Georgios Goumas (National Technical University of Athens), Petros Maragos (National Technical University of Athens)

In this paper, we introduce channel-wise recurrent convolutional neural networks (RecNets), a family of novel, compact neural network architectures for computer vision tasks inspired by recurrent neural networks (RNNs). RecNets build upon Channel-wise Recurrent Convolutional (CRC) layers, a novel type of convolutional layer that splits the input channels into disjoint segments and processes them in a recurrent fashion. In this way, we simulate wide, yet compact models, since the number of parameters is vastly reduced via the parameter sharing of the RNN formulation. Experimental results on the CIFAR-10 and CIFAR-100 image classification tasks demonstrate the superior size/accuracy trade-off of RecNets com-

pared to other compact state-of-the-art architectures.

61 **Pose from Shape: Deep Pose Estimation for Arbitrary 3D Objects**

Yang Xiao (École des ponts ParisTech), Xuchong Qiu (École des Ponts ParisTech), Pierre-Alain Langlois (École des Ponts ParisTech), Mathieu Aubry (École des ponts ParisTech), Renaud Marlet (École des Ponts ParisTech)

Most deep pose estimation methods need to be trained for specific object instances or categories. In this work we propose a completely generic deep pose estimation approach, which does not require the network to have been trained on relevant categories, nor objects in a category to have a canonical pose. We believe this is a crucial step to design robotic systems that can interact with new objects “in the wild” not belonging to a predefined category. Our main insight is to dynamically condition pose estimation with a representation of the 3D shape of the target object. More precisely, we train a Convolutional Neural Network that takes as input both a test image and a 3D model, and outputs the relative 3D pose of the object in the input image with respect to the 3D model. We demonstrate that our method boosts performances for supervised category pose estimation on standard benchmarks, namely Pascal3D+, ObjectNet3D and Pix3D, on which we provide results superior to the state of the art. More importantly, we show that our network trained on everyday man-made objects from ShapeNet generalizes without any additional training to completely new types of 3D objects by providing results on the LINEMOD dataset as well as on natural entities such as animals from ImageNet.

62 **XNOR-Net++: Improved binary neural networks**

Adrian Bulat (Samsung AI Center, Cambridge), Georgios Tzimiropoulos (Samsung AI Centre, Cambridge)

This paper proposes an improved training algorithm for binary neural networks in which both weights and activations are binary numbers. A key but fairly overlooked feature of the current state-of-the-art method of XNOR-Net is the use of analytically calculated real-valued scaling factors for re-weighting the output of binary convolutions. We argue that analytic calculation of these factors is sub-optimal. Instead, in this work, we make the following contributions: (a) we propose to fuse the activation and weight scaling factors into a single one that is learned discriminatively via backpropagation. (b) More importantly, we explore several ways of constructing the shape of the scale factors while keeping the computational budget fixed. (c) We empirically measure the accuracy of our approximations and show that they are significantly more accurate than the analytically calculated one. (d) We show that our approach significantly outperforms XNOR-Net within the same computational budget

when tested on the challenging task of ImageNet classification, offering up to 6% accuracy gain.

63 Curriculum based Dropout Discriminator for Domain Adaptation

Vinod Kurmi (IIT Kanpur), Vipul Bajaj (IIT Kanpur), Vinay Namboodiri (IIT Kanpur), K. S. Venkatesh (IIT Kanpur)

Domain adaptation is essential to enable wide usage of deep learning based networks trained using large labeled datasets. Adversarial learning based techniques have shown their utility towards solving this problem using a discriminator that ensures source and target distributions are close. However, here we suggest that rather than using a point estimate it would be useful if a distribution based discriminator could be used to bridge this gap. This could be achieved using multiple classifiers or using traditional ensembles methods. In contrast, we suggest that a Monte Carlo dropout based ensemble discriminator could suffice to obtain the distribution based discriminator. Specifically, we propose a curriculum based dropout discriminator that gradually increases the variance of the sample based distribution and the corresponding reverse gradients are used to align the source and target feature representations. The detailed results and thorough ablation analysis show that our model outperforms state-of-art results.

2 Joint Spatial and Layer Attention for Convolutional Networks

Tony Joseph (University of Ontario Institute of Technology), Konstantinos Derpanis (Ryerson University), Faisal Qureshi (University of Ontario Institute of Technology)

In this paper, we propose a novel approach that learns to sequentially attend to different Convolutional Neural Networks (CNN) layers (i.e., “what” feature abstraction to attend to) and different spatial locations of the selected feature map (i.e., “where”) to perform the task at hand. Specifically, at each Recurrent Neural Network step, both a CNN layer and localized spatial region within it are selected for further processing. We demonstrate the effectiveness of this approach on two computer vision tasks: (i) image-based six degrees of freedom camera pose regression and (ii) indoor scene classification. Empirically, we show that combining the “what” and “where” aspects of attention improves network performance on both tasks. We evaluate our method on standard benchmarks for camera localization (Cambridge, 7-Scenes, and TUM-LSI) and for scene classification (MIT-67 Indoor Scenes). For camera localization, our approach reduces the median error by 18.8% for position and 8.2% for orientation (averaged over all scenes), and for scene classification, it improves the mean accuracy by 3.4% over previous methods.

65 Directed-Weighting Group Lasso For Eltwise Blocked CNN Pruning

Ke Zhan (Beijing University Of Technology), Shimiao Jiang (Alibaba, Inc.), Yu Bai (JD.com, Inc.), Yi Li (JD.com, Inc)

Eltwise layer is a commonly used structure in the multi-branch deep learning network. In a filter-wise pruning procedure, due to the specific operation of the eltwise layer, all its previous convolutional layers should vote for which filters by index to be pruned. Since only an intersection of the voted filters was pruned, the compression rate is limited. The work proposes a method called Directed-Weighting Group Lasso (DWGL), which enforces an index-wise incremental (directed) coefficient on the filter-level group lasso items, so that the low index filters getting high activation tend to be kept while the high index ones tend to be pruned. When using DWGL, less filter is retained during the voting process and the compression rate can be boosted. The paper test the proposed method on ResNet series networks. On CIFAR-10, it achieved a 75.34% compression rate on ResNet-56 with a 0.94% error increment, and a 52.06% compression rate on ResNet-20 with a 0.72% error increment. On ImageNet, it achieved a 53% compression rate with ResNet-50 with a 0.6% error increment, which speed up the network by 2.23 times, it further achieved a 75% compression rate on ResNet-50 with a 1.2% error increment, which speed up the network by 4 times.

66 Camera Style and Identity Disentangling Network for Person Re-identification

Ruochen Zheng (Huazhong University of Science and Technology), Lerenhan Li (Huazhong University of Science and Technology), Chuchu Han (Huazhong University of Science and Technology), Changxin Gao (Huazhong University of Science and Technology), Nong Sang (School of Automation, Huazhong University of Science and Technology)

Camera style (camstyle) is a main factor that affects the performance of person re-identification (ReID). In the past years, existing works mainly exploit implicit solutions from the inputs by designing some strong constraints. However, these methods cannot consistently work as the camstyle still exists in the inputs as well as in the intermediate features. To address this problem, we propose a Camstyle-Identity Disentangling (CID) network for person ReID. More specifically, we disentangle the ID feature and camstyle feature in the latent space. In order to disentangle the features successfully, we present a Camstyle Shuffling and Retraining (CSR) scheme to generate more ID-preserved and camstyle variation samples for training. The proposed scheme ensures the success of disentangling and is able to eliminate the camstyle features in the backbone during the training process. Numerous experimental results on the Market-1501 and DukeMTMC-reID

datasets demonstrate that our network can effectively disentangle the features and facilitate the person ReID networks.

67 **Pseudo-Labeling Curriculum for Unsupervised Domain Adaptation**

Jaehoon Choi (Korea Advanced Institute of Science and Technology), Minki Jeong (Korea Advanced Institute of Science and Technology), Taekyung Kim (Korea Advanced Institute of Science and Technology), Changick Kim (Korea Advanced Institute of Science and Technology)

To learn target discriminative representations, using pseudo-labels is a simple yet effective approach for unsupervised domain adaptation. However, the existence of false pseudo-labels, which may have a detrimental influence on learning target representations, remains a major challenge. To overcome this issue, we propose a pseudo-labeling curriculum based on a density-based clustering algorithm. Since samples with high density values are more likely to have correct pseudo-labels, we leverage these subsets to train our target network at the early stage, and we provide data subsets with low density values at the later stage. We can progressively improve the capability of our network to generate pseudo-labels, and thus these target samples with pseudo-labels are effective for training our model. Moreover, we present a clustering constraint to enhance the discriminative power of the learned target features. Our approach achieves state-of-the-art performance on three benchmarks: Office-31, imageCLEF-DA, and Office-Home.

68 **BioFaceNet: Deep Biophysical Face Image Interpretation**

Sarah Alotaibi (University of York), William Smith (University of York)

In this paper we present BioFaceNet, a deep CNN that learns to decompose a single face image into biophysical parameters maps, diffuse and specular shading maps as well as estimating the spectral power distribution of the scene illuminant and the spectral sensitivity of the camera. The network comprises a fully convolutional encoder for estimating the spatial maps with a fully connected branch for estimating the vector quantities. The network is trained using a self-supervised appearance loss computed via a model-based decoder. The task is highly underconstrained so we impose a number of model-based priors. Skin spectral reflectance is restricted to a biophysical model, we impose a statistical prior on camera spectral sensitivities, a physical constraint on illumination spectra, a sparsity prior on specular reflections and direct supervision on diffuse shading using a rough shape proxy. We show convincing qualitative results on in-the-wild data and introduce a benchmark for quantitative evaluation on this new task.

5 Multi-Weight Partial Domain Adaptation

69

Jian Hu (Shanghai Jiaotong University), Hongya Tuo (Shanghai Jiaotong University), Chao Wang (Shanghai Jiaotong University), Lingfeng Qiao (Shanghai Jiaotong University), Haowen Zhong (Shanghai Jiaotong University), Zhongliang Jing (Shanghai Jiaotong University)

Domain adaptation (DA) plays an important role in transfer learning. However, when target label space is a subset of source label space, standard DA cannot tackle this issue. Partial domain adaptation focuses on how to transfer knowledge from massive labelled dataset to unlabelled miniature one, which attracts extensive research interest. In this paper, we propose a Multi-Weight Partial Domain Adaptation (MWPDA) to solve the problem. We divide the source domain into two parts: shared classes and outlier classes. MWPDA aims to reduce negative transfer caused by outlier classes when transferring knowledge between domains. Based on OTSU-Algorithm, hard shared-class labels are obtained to decrease weights of outlier classes and increase ones of shared classes. A novel shared-sample classifier is trained for shared-sample weights to distinguish outlier samples. Shared-class weights and shared-sample weights are acted on source classifier and domain discriminator to jointly distinguish outlier classes and samples. This kind of multi-weight mechanism can avoid misalignment to outlier classes and promote classification accuracy. Furthermore, our universal network framework is utilized for both partial domain adaptation and standard domain adaptation issues. Extensive experiments on three benchmark domain adaptation datasets illustrate our method achieves state-of-the-art results.

70 Semantically-Aware Attentive Neural Embeddings for 2D Long-Term Visual Localization

Zachary Seymour (SRI International), Karan Sikka (SRI International), Han-Pang Chiu (SRI International), Supun Samarasekera (SRI International), Rakesh Kumar (SRI International)

We present an approach that combines appearance and semantic information for 2D image-based localization (2D-VL) across large perceptual changes and time lags. Compared to appearance features, the semantic layout of a scene is generally more invariant to appearance variations. We use this intuition and propose a novel end-to-end deep attention-based framework that utilizes multimodal cues to generate robust embeddings for 2D-VL. The proposed attention module predicts a shared channel attention and modality-specific spatial attentions to guide the embeddings to focus on more reliable image regions. We evaluate our model against state-of-the-art (SOTA) methods on three challenging localization datasets. We report an average (absolute) improvement of 19% over current SOTA for 2D-VL. Furthermore, we present an extensive study demonstrating the contribution of each component of our model, showing 8 – 15% and 4%

improvement from adding semantic information and our proposed attention module. We finally show the predicted attention maps to offer useful insights into our model.

71 EPNAS: Efficient Progressive Neural Architecture Search

Yanqi Zhou (Google), Peng Wang (Baidu USA LLC.)

In this paper, we propose Efficient Progressive Neural Architecture Search (EPNAS), a neural architecture search (NAS) framework that efficiently handles large search space through a novel progressive search policy with performance prediction based on REINFORCE. EPNAS is designed to search target networks in parallel, which is more scalable on parallel platforms. More importantly, EPNAS can be generalized to architecture search with multiple resource constraints, *e.g.*, model size, compute complexity or intensity, which is crucial for deployment in widespread platforms such as mobile and cloud. We compare EPNAS against other state-of-the-art (SOTA) network architectures (*e.g.*, MobileNetV2) and efficient NAS algorithms (*e.g.*, ENAS, and PNAS) on image recognition tasks using CIFAR10 and ImageNet. On both datasets, EPNAS is superior *w.r.t.* architecture searching speed and recognition accuracy.

72 Batch-wise Logit-Similarity: Generalizing Logit-Squeezing and Label-Smoothing

Ali Shafahi (University of Maryland), Mohammad Amin Ghiasi (University of Maryland), Mahyar Najibi (University of Maryland), Furong Huang (University of Maryland), John Dickerson (University of Maryland), Tom Goldstein (University of Maryland)

We study how cheap regularization methods can increase adversarial robustness. In particular, we introduce logit-similarity which can be seen as a generalization of label-smoothing and logit-squeezing. Our version of logit-squeezing applies a batch-wise penalty and allows penalizing the logits aggressively. By measuring the robustness of our models against various gradient-based and gradient-free attacks, we experimentally show that, with the correct choice of hyper-parameters, regularized models can be as robust as adversarially trained models on the CIFAR-10 and CIFAR-100 datasets when robustness is measured in terms of L-Infinity norm attacks. Unlike conventional adversarial training, regularization methods keep training time short and become robust against L-2 norm attacks in addition to L-Infinity norm.

73 Scrutinizing and De-Biasing Intuitive Physics with Neural Stethoscopes

Fabian Fuchs (Oxford Robotics Insitute), Oliver Groth (Oxford Robotics Insitute), Adam Kosiorek (University of Oxford), Alex Bewley (Google), Markus Wulfmeier (DeepMind), Andrea Vedaldi (University of Oxford), Ingmar Posner (University of Oxford)

Visually predicting the stability of block towers is a popular task in the domain of intuitive physics. While previous work focusses on prediction accuracy, a one-dimensional performance measure, we provide a broader analysis of the learned physical understanding of the final model and how the learning process can be guided. To this end, we introduce neural stethoscopes as a general purpose framework for quantifying the degree of importance of specific factors of influence in deep neural networks as well as for actively promoting and suppressing information as appropriate. In doing so, we unify concepts from multitask learning as well as training with auxiliary and adversarial losses. We apply neural stethoscopes to analyse the state-of-the-art neural network for stability prediction. We show that the baseline model is susceptible to being misled by incorrect visual cues. This leads to a performance breakdown to the level of random guessing when training on scenarios where visual cues are inversely correlated with stability. Using stethoscopes to promote meaningful feature extraction increases performance from 51% to 90% prediction accuracy. Conversely, training on an easy dataset where visual cues are positively correlated with stability, the baseline model learns a bias leading to poor performance on a harder dataset. Using an adversarial stethoscope, the network is successfully de-biased, leading to a performance increase from 66% to 88%.

74 **MixConv: Mixed Depthwise Convolutional Kernels**

Mingxing Tan (Google Brain), Quoc Le (Google Brain)

Depthwise convolution is becoming increasingly popular in modern efficient ConvNets, but its kernel size is often overlooked. In this paper, we systematically study the impact of different kernel sizes, and observe that combining the benefits of multiple kernel sizes can lead to better accuracy and efficiency. Based on this observation, we propose a new mixed depthwise convolution (MDConv), which naturally mixes up multiple kernel sizes in a single convolution. As a simple drop-in replacement of vanilla depthwise convolution, our MDConv improves the accuracy and efficiency for existing MobileNets on both ImageNet classification and COCO object detection.

By integrating MDConv into AutoML search space, we have further developed a new family of models, named as MixNets, which significantly outperform previous models including MobileNetV2 (ImageNet top-1 accuracy +4.2%), ShuffleNetV2 (+3.5%), MnasNet (+1.3%), ProxylessNAS (+2.2%), and FBNet (+2.0%). In particular, our MixNet-L achieves a new state-of-the-art 78.9% ImageNet top-1 accuracy under typical mobile settings (<600M FLOPS). Code is at <https://github.com/tensorflow/tpu/tree/master/models/official/mnasnet/mixnet>.

75 Look and Modify: Modification Networks for Image Captioning

Fawaz Sammani (Multimedia University), Mahmoud Elsayed (Multimedia University)

Attention-based neural encoder-decoder frameworks have been widely used for image captioning. Many of these frameworks deploy their full focus on generating the caption from scratch by relying solely on the image features or the object detection regional features. In this paper, we introduce a framework that learns to modify existing captions from a given framework by modeling the residual information, where at each timestep, the model learns what to keep, remove or add to the existing caption allowing the model to fully focus on “what to modify” rather than on “what to predict”. We evaluate our method on the COCO dataset, trained on top of several image captioning frameworks and show that our model successfully modifies captions yielding better ones with better evaluation scores.

76 Ordinal Pooling

Adrien Deliege (University of Liege), Ashwani Kumar (University of Sheffield), Maxime Istasse (UCLouvain, ICTEAM, ELEN, ISPGROUP), Christophe De Vleeschouwer (Université Catholique de Louvain), Marc Van Droogenbroeck (University of Liege)

In the framework of convolutional neural networks, downsampling is often performed with an average-pooling, where all the activations are treated equally, or with a max-pooling operation that only retains an element with maximum activation while discarding the others. Both of these operations are restrictive and have previously been shown to be sub-optimal. To address this issue, a novel pooling scheme, named ordinal pooling, is introduced in this work. Ordinal pooling rearranges all the elements of a pooling region in a sequence and assigns a different weight to each element based upon its order in the sequence. These weights are used to compute the pooling operation as a weighted sum of the rearranged elements of the pooling region. They are learned via a standard gradient-based training, allowing to learn a behavior anywhere in the spectrum of average-pooling to max-pooling in a differentiable manner. Our experiments suggest that it is advantageous for the networks to perform different types of pooling operations within a pooling layer and that a hybrid behavior between average- and max-pooling is often beneficial. More importantly, they also demonstrate that ordinal pooling leads to consistent improvements in the accuracy over average- or max-pooling operations while speeding up the training and alleviating the issue of the choice of the pooling operations and activation functions to be used in the networks. In particular, ordinal pooling mainly helps on lightweight or quantized deep learning architectures, as typically considered e.g. for embedded applications.

77 Defending against adversarial examples using defense kernel network

Yuying Hao (TBSI, Tsinghua), Tuanhui Li (Tsinghua University), Yong Jiang (Tsinghua University), Xuanye Cheng (SenseTime Research), Li Li (Graduate School at Shenzhen, Tsinghua University)

Deep neural networks have been widely used in recent years. Thus, the security of deep neural networks is crucial for practical applications. Most of previous defense methods are not robust for diverse adversarial perturbations and rely on some specific structure or properties of the attacked model. In this work, we propose a novel defense kernel network to convert the adversarial examples to images with evident classification features. Our method is robust to variety adversarial perturbations and can be independently apply to different attacked model. Experiments on two benchmarks demonstrate that our method has competitive defense ability against existing state-of-the-art defense methods.

10 One-shot Face Reenactment 78

Cheng Li (SenseTime Research), Yunxuan Zhang (SenseTime Research), Yue He (SenseTime Research), Siwei Zhang (SenseTime Research), Ziwei Liu (The Chinese University of Hong Kong), Chen Change Loy (Nanyang Technological University)

To enable realistic shape (e.g. pose and expression) transfer, existing face reenactment methods rely on a set of target faces for learning subject-specific traits. However, in real-world scenario end-users often only have one target face at hand, rendering existing methods inapplicable. In this work, we bridge this gap by proposing a novel one-shot face reenactment learning framework. Our key insight is that the one-shot learner should be able to disentangle and compose appearance and shape information for effective modeling. Specifically, the target face appearance and the source face shape are first projected into latent spaces with their corresponding encoders. Then these two latent spaces are associated by learning a shared decoder that aggregates multi-level features to produce the final reenactment results. To further improve the synthesizing quality on mustache and hair regions, we additionally propose FusionNet which combines the strengths of our learned decoder and the traditional warping method. Extensive experiments show that our one-shot face reenactment system achieves superior transfer fidelity as well as identity preserving capability than alternatives. More remarkably, our approach trained with only one target image per subject achieves competitive results to those using a set of target images, demonstrating the practical merit of this work.

79 Predicting Visual Memory Schemas with Variational Autoencoders

Cameron Kyle-Davidson (University of York), Adrian Bors (University of York), Karla Evans (University of York)

Visual memory schema (VMS) maps show which regions of an image cause that image to be remembered or falsely remembered. Previous work has succeeded in generating low resolution VMS maps using convolutional neural networks. We instead approach this problem as an image-to-image translation task making use of a variational autoencoder. This approach allows us to generate higher resolution dual channel images that represent visual memory schemas, allowing us to evaluate predicted true memorability and false memorability separately. We also evaluate the relationship between VMS maps, predicted VMS maps, ground truth memorability scores, and predicted memorability scores.

12 Revisiting Residual Networks with Nonlinear Shortcuts

80

Chaoning Zhang (Korea Advanced Institute of Science and Technology), Francois Rameau (Korea Advanced Institute of Science and Technology), Jean-Charles Bazin (Korea Advanced Institute of Science and Technology), Dawit Mureja Argaw (Korea Advanced Institute of Science and Technology), Philipp Benz (Korea Advanced Institute of Science and Technology), Seokju Lee (Korea Advanced Institute of Science and Technology), Junsik Kim (Korea Advanced Institute of Science and Technology), In So Kweon (Korea Advanced Institute of Science and Technology)

Residual networks (ResNets) with an identity shortcut have been widely used in various computer vision tasks due to their compelling performance and simple design. In this paper we revisit ResNet identity shortcut and propose RGSNets which are based on a new nonlinear ReLU Group Normalization (RG) shortcut, outperforming the existing ResNet by a relatively large margin. Our work is inspired by previous findings that there is a trade-off between representational power and gradient stability in deep networks and that the identity shortcut reduces the representational power. Our proposed nonlinear RG shortcut can contribute to effectively utilizing the representational power of relatively shallow networks and outperform much (3 or 4 times) deeper ResNets, which demonstrates the high efficiency of RG shortcut. Moreover, we have explored variations of RGSNets, and our experimental result shows that Res-RGSNet combining the proposed RG shortcut with the existing identity shortcut achieves the best performance and is robust to network depth. Our code and model will be publicly available.

81 PMC-GANs: Generating Multi-Scale High-Quality Pedestrian with Multimodal Cascaded GANs

Jie Wu (China Electronics Technology Cyber Security Co., Ltd.), Ying Peng (China Electronics Technology Cyber Security Co., Ltd.), Chenghao Zheng (China Electronics Technology Cyber Security Co., Ltd.), Zongbo Hao (UESTC), Zhang Jian (China Electronics Technology Cyber Security Co., Ltd)

Recently, generative adversarial networks (GANs) have shown great ad-

vantages in synthesizing images, leading to a boost of explorations of using faked images to augment data. This paper proposes a multimodal cascaded generative adversarial networks (PMC-GANs) to generate realistic and diversified pedestrian images and augment pedestrian detection data. The generator of our model applies a residual U-net structure, with multi-scale residual blocks to encode features, and attention residual blocks to help decode and rebuild pedestrian images. The model constructs in a coarse-to-fine fashion and adopts cascade structure, which is beneficial to produce high-resolution pedestrians. PMC-GANs outperforms baselines, and when used for data augmentation, it improves pedestrian detection results.

82 Contrastive Learning for Lifted Networks

Christopher Zach (Chalmers University), Virginia Estellers (Microsoft)

In this work we address supervised learning via lifted network formulations. Lifted networks are interesting because they allow training on massively parallel hardware and assign energy models to discriminatively trained neural networks. We demonstrate that training methods for lifted networks proposed in the literature have significant limitations, and therefore we propose to use a contrastive loss to train lifted networks. We show that this contrastive training approximates back-propagation in theory and in practice, and that it is superior to the regular training objective for lifted networks.

83 Adaptive Graphical Model Network for 2D Handpose Estimation

Deying Kong (University of California, Irvine), Yifei Chen (Tencent), Haoyu Ma (Southeast University), Xiangyi Yan (Southern University of Science and Technology), Xiaohui Xie (University of California, Irvine)

In this paper, we propose a new architecture called Adaptive Graphical Model Network (AGMN) to tackle the challenging task of 2D hand pose estimation from a monocular RGB image. The AGMN consists of two branches of deep convolutional neural networks (DCNNs) for calculating unary and pairwise potential functions, followed by a graphical model inference module for integrating unary and pairwise potentials. Unlike existing architectures proposed to combine DCNNs with graphical models, our AGMN is novel in that the parameters of its graphical model are conditioned on and fully adaptive to individual input images. Experiments show that our approach outperforms the state-of-the-art method used in 2D hand keypoints estimation by a notable margin on two public datasets.

84 Bag of Negatives for Siamese Architectures

Bojana Gajic (Computer Vision Center), Ariel Amato (Vintra, Inc.), Ramón Baldrich (Computer Vision Center), Carlo Gatta (Vintra, Inc.)

Training a Siamese architecture for re-identification with a large number of identities is a challenging task due to the difficulty of finding relevant negative samples efficiently. In this work we present Bag of Negatives (BoN), a method for accelerated and improved training of Siamese networks that scales well on datasets with a very large number of identities. BoN is an efficient and loss-independent method, able to select a bag of “high quality negatives”, based on a novel online hashing strategy.

85 **SC-RANK: Improving Convolutional Image Captioning with Self-Critical Learning and Ranking Metric-based Reward**

Shiyang Yan (Queen's University Belfast), Yang Hua (Queen's University Belfast), Neil Robertson (Queen's University Belfast)

Image captioning usually employs a Recurrent Neural Network (RNN) to decode the image features from a Convolutional Neural Network (CNN) into a sentence. This RNN model is trained under Maximum Likelihood Estimation (MLE). However, inherent issues like the complex memorising mechanism of the RNNs and the exposure bias introduced by MLE exist in this approach. Recently, the convolutional captioning model shows advantages with a simpler architecture and a parallel training capability. Nevertheless, the MLE training brings the exposure bias which still prevents the model from achieving better performance. In this paper, we prove that the self-critical algorithm can optimise the CNN-based model to alleviate this problem. A ranking metric-based reward, denoted as SC-RANK, is proposed with the sentence embeddings from a pre-trained language model to generate more diversified captions. Applying SC-RANK can avoid the tedious tuning of the specially-designed language model and the knowledge transferred from a pre-trained language model proves to be helpful for image captioning tasks. State-of-the-art results have been obtained in the MSCOCO dataset by proposed SC-RANK.

86 **SO(2)-equivariance in Neural networks using tensor nonlinearity**

Muthuvel Murugan Issakkimuthu (Chennai Mathematical Institute), K V Subrahmanyam (Chennai Mathematical Institute)

Inspired by recent work of Kondor and Cohen and Welling, we build rotation equivariant autoencoders to obtain a basis of images adapted to the group of planar rotations $SO(2)$, directly from the data. We do this in an unsupervised fashion, working in the Fourier domain of $SO(2)$. Working in the Fourier domain we build a rotation equivariant classifier to classify images. As in the recent papers of Thomas et al. and Kondor et al. we use

tensor product nonlinearity to build our autoencoders and classifiers. We discover the basis using a small sample of inputs. As a consequence our classifier is robust to rotations - the classifier trained on upright images, classifies rotated versions of images, achieving state of the art. In order to deal with images under different scales simultaneously, we define the notion of a coupled-bases and show that a coupled-bases can be learned using tensor nonlinearity.

87 Discriminative Features Matter: Multi-layer Bilinear Pooling for Camera Localization

Xin Wang (Beihang University), Xiang Wang (Beihang University), Chen Wang (Beihang University), Xiao Bai (Beihang University), Jing Wu (Cardiff University), Edwin Hancock (University of York)

Deep learning based camera localization from a single image has been explored recently since these methods are computationally efficient. However, existing methods only provide general global representations, from which an accurate pose estimation can not be reliably derived. We claim that effective feature representations for accurate pose estimation shall be both “informative” (focusing on geometrically meaningful regions) and “discriminative” (accounting for different poses of similar images). Therefore, we propose a novel multi-layer factorized bilinear pooling module for feature aggregation. Specifically, informative features are selected via bilinear pooling, and discriminative features are highlighted via multi-layer fusion. We develop a new network for camera localization using the proposed feature pooling module. The effectiveness of our approach is demonstrated by experiments on an outdoor Cambridge Landmarks dataset and an indoor 7 Scenes dataset. The results show that focusing on discriminative features significantly improves the network performance of camera localization in most cases.

88 ProSe: Product of Orthogonal Spheres Parameterization for Disentangled Representation Learning

Ankita Shukla (Indraprastha Institute of Information Technology), Shagun Uppal (Indraprastha Institute of Information Technology), Sarthak Bhagat (Indraprastha Institute of Information Technology), Saket Anand (Indraprastha Institute of Information Technology), Pavan Turaga (Arizona State University)

Learning representations that can disentangle explanatory attributes underlying the data improves interpretability as well as provides control on data generation. Various learning frameworks such as VAEs, GANs and autoencoders have been used in the literature to learn such representations. Most often, the latent space is constrained to a partitioned representation or structured by a prior to impose disentangling. In this work, we advance the use of a latent representation based on a product space of Orthogonal Spheres ProSe. The ProSe model is motivated by the reasoning that

latent-variables related to the physics of image-formation can under certain relaxed assumptions lead to spherical-spaces. Orthogonality between the spheres is motivated via physical independence models. Imposing the orthogonal-sphere constraint is much simpler than other complicated physical models, is fairly general and flexible, and extensible beyond the factors used to motivate its development. Under further relaxed assumptions of equal-sized latent blocks per factor, the constraint can be written down in closed form as an ortho-normality term in the loss function. We show that our approach improves the quality of disentanglement significantly. We find consistent improvement in disentanglement compared to several state-of-the-art approaches, across several benchmarks and metrics.

89 **Dynamic Neural Network Channel Execution for Efficient Training**

Simeon Spasov (University of Cambridge), Pietro Lió (University of Cambridge)

Existing methods for reducing the computational burden of neural networks at run-time, such as parameter pruning or dynamic computational path selection, focus solely on improving computational efficiency during inference. On the other hand, in this work, we propose a novel method which reduces the memory footprint and number of computing operations required for training and inference. Our framework efficiently integrates pruning as part of the training procedure by exploring and tracking the relative importance of convolutional channels. At each training step, we select only a subset of highly salient channels to execute according to the combinatorial upper bound confidence algorithm, and run a forward and backward pass only on these activated channels, hence learning their parameters. Consequently, we enable the efficient discovery of compact models. We validate our approach empirically on state-of-the-art CNNs - VGGNet, ResNet and DenseNet, and on several image classification datasets. Results demonstrate our framework for dynamic channel execution reduces computational cost up to 4x and parameter count up to 9x, thus reducing the memory and computational demands for discovering and training compact neural network models.

16 **Class-Distinct and Class-Mutual Image Generation with GANs**

Takuhiko Kaneko (The University of Tokyo), Yoshitaka Ushiku (The University of Tokyo), Tatsuya Harada (The University of Tokyo / RIKEN)

Class-conditional extensions of generative adversarial networks (GANs), such as auxiliary classifier GAN (AC-GAN) and conditional GAN (cGAN), have garnered attention owing to their ability to decompose representations into class labels and other factors and to boost the training stability.

However, a limitation is that they assume that each class is separable and ignore the relationship between classes even though class overlapping frequently occurs in a real-world scenario when data are collected on the basis of diverse or ambiguous criteria. To overcome this limitation, we address a novel problem called class-distinct and class-mutual image generation, in which the goal is to construct a generator that can capture between-class relationships and generate an image selectively conditioned on the class specificity. To solve this problem without additional supervision, we propose classifier’s posterior GAN (CP-GAN), in which we redesign the generator input and the objective function of AC-GAN for class-overlapping data. Precisely, we incorporate the classifier’s posterior into the generator input and optimize the generator so that the classifier’s posterior of generated data corresponds with that of real data. We demonstrate the effectiveness of CP-GAN using both controlled and real-world class-overlapping data with a model configuration analysis and comparative study. Our code is available at <https://github.com/takuhirok/CP-GAN/>.

91 **Classification is a Strong Baseline for Deep Metric Learning**

Hao-Yu Wu (Pinterest, Inc.), Andrew Zhai (Pinterest, Inc.)

Deep metric learning aims to learn a function mapping image pixels to embedding feature vectors that model the similarity between images. Two major applications of metric learning are content-based image retrieval and face verification. For the retrieval tasks, the majority of current state-of-the-art (SOTA) approaches are triplet-based non-parametric training. For the face verification tasks, however, recent SOTA approaches have adopted classification-based parametric training. In this paper, we look into the effectiveness of classification based approaches on image retrieval datasets. We evaluate on several standard retrieval datasets such as CAR-196, CUB-200-2011, Stanford Online Product, and In-Shop datasets for image retrieval and clustering, and establish that our classification-based approach is competitive across different feature dimensions and base feature networks. We further provide insights into the performance effects of subsampling classes for scalable classification-based training, and the effects of binarization, enabling efficient storage and computation for practical applications.

92 **Functionality-Oriented Convolutional Filter Pruning**

Zhuwei Qin (George Mason University), Fuxun Yu (George Mason University), Chenchen Liu (Clarkson University), Xiang Chen (George Mason University)

The sophisticated structure of Convolutional Neural Network (CNN) models allows for outstanding performance, but at the cost of intensive

computation load. To reduce this cost, many model compression works have been proposed to eliminate insignificant model structures, such as pruning the convolutional filters which have smaller absolute weights. However, most of these works merely depend on quantitative significance ranking without qualitative filter functionality interpretation or thorough model structure analysis, resulting in considerable model retraining cost. Different from previous works, we interpret the functionalities of the convolutional filters and identify the model structural redundancy as repetitive filters with similar feature preferences. In this paper, we proposed a functionality-oriented filter pruning method, which can precisely remove the redundant filters without compromising the model functionality integrity and accuracy performance. Experiments with multiple CNN models and databases testified the unreliability of conventional weight-ranking based filter pruning methods, and demonstrate our method's advantages in terms of computation load reduction (at most 68.88% FLOPs), accuracy retaining ($<0.34\%$ accuracy drop), and expected retraining independence.

Document Processing

6 Text Recognition using local correlation 93

Yujia Li (Institute of Information Engineering, Chinese Academy of Sciences), Hongchao Gao (Institute of Information Engineering, Chinese Academy of Sciences), Xi Wang (Institute of Information Engineering, Chinese Academy of Sciences), Jizhong Han (Institute of Information Engineering, Chinese Academy of Sciences), Ruixuan Li (Huazhong University of Science and Technology)

In this paper, we propose an improved text recognition method by considering the local correlation of the character region. Fractal theory indicates that most images have self-similarity properties including scene text images. The recent methods always extract the features of word region through a Convolution Neural Network(CNN) which uses fixed kernels. The self-similarity of the image is not fully used. In our paper, we propose Local Correlation(LC) layer which represents the self-similarity of text image by considering the local correlation of the character region. This layer weight the input by computing the correlation. This mechanism not only brings significant improvement of recognition results but also can be easy to embed in other recognition architectures. After we embed this layer in scene text recognition architecture, the experiment shows that the proposed model gains better representations of the scene images and achieves the state-of-the-art results on several benchmark datasets includ-

ing IIIT-5K, SVT, CUTE80, SVT-Perspective and ICDAR.

94 **A Learning-based Text Synthesis Engine for Scene Text Detection**

Xiao Yang (Pennsylvania State University), Dafang He (Pennsylvania State University), Dan Kifer (Pennsylvania State University), Lee Giles (Pennsylvania State University)

Scene text detection and recognition methods have recently greatly improved with the use of synthetic training data playing an important role. That being said, for text detection task the performance of a model that is trained solely on large-scale synthetic data is significantly worse than one trained on a few real-world data samples. However, state-of-the-art performance on text recognition can be achieved by only training on synthetic data. This shows the limitations in only using large-scale synthetic data for scene text detection. In this work, we propose the first learning-based, data-driven text synthesis engine for scene text detection task. Our text synthesis engine is decomposed into two modules: 1) a *location* module that learns the distribution of text locations on the image plane, and 2) an *appearance* module that translates the text-inserted images to realistic-looking ones that are essentially indistinguishable from real-world scene text images. Evaluation of our created synthetic data on ICDAR 2015 Incidental Scene Text dataset outperforms previous text synthesis methods.

95 **Document Binarization using Recurrent Attention Generative Model**

Shuchun Liu (ele AI Lab), Feiyun Zhang (ele AI Lab), Pan He (University of Florida), Mingxi Chen (Tongji University), Yufei Xie (East China Normal University), Jie Shao (Fudan University)

Image binarization is an elementary pre-processing step in the document image analysis and recognition pipeline. It is well-known that contextual and semantic information is beneficial to the separation of foreground text from complex background. We develop a simple general deep learning approach, by introducing a recurrent attention generative model with adversarial training. The DB-RAM model comprises three contributions: First, to suppress the interference from complex background, non-local attention blocks are incorporated to capture spatial long-range dependencies. Second, we explore the use of Spatial Recurrent Neural Networks (SRNNs) to pass spatially varying contextual information across an image, which leverages the prior knowledge of text orientation and semantics. Third, to validate the effectiveness of our proposed method, we further synthetically generate two comprehensive subtitle datasets that cover various real-world conditions. Evaluated on various standard benchmarks, our proposed method significantly outperforms state-of-the-art binarization

methods both quantitatively and qualitatively. Experiment results show that the proposed method can also improve the recognition rate. Moreover, the proposed method performs well in the task of image unshadowing, which evidently verifies its generality.

96 **End-to-End Information Extraction by Character-Level Embedding and Multi-Stage Attentional U-Net**

Tuan Anh Nguyen Dang (Cinnamon), Dat Nguyen Thanh (Cinnamon)

Information extraction from document images has received a lot of attention recently, due to the need for digitizing a large volume of unstructured documents such as invoices, receipts, bank transfers, etc. In this paper, we propose a novel deep learning architecture for end-to-end information extraction on the 2D character-grid embedding of the document, namely the “Multi-Stage Attentional U-Net”. To effectively capture the textual and spatial relations between 2D elements, our model leverages a specialized multi-stage encoder-decoders design, in conjunction with efficient uses of the self-attention mechanism and the box convolution. Experimental results on different datasets show that our model outperforms the baseline U-Net architecture by a large margin while using 40% less parameters. Moreover, it also significantly improved the baseline in erroneous OCR and limited training data scenario, thus becomes practical for real-world applications.

Face and Gesture

13 **Unified 2D and 3D Hand Pose Estimation from a Single Visible or X-ray Image**

Akila Pemasiri (Queensland University of Technology), Kien Nguyen Thanh (Queensland University of Technology), Sridha Sridharan (Queensland University of Technology), Clinton Fookes (Queensland University of Technology)

Robust detection of the keypoints of the human hand from a single 2D image is a crucial step in many applications including medical image processing, where X-ray images play a vital role. In this paper, we address the challenging problem of 2D and 3D hand pose estimation from a single hand image, where the image can be either in the visible spectrum or an X-ray. In contrast to the state-of-the-art methods, which are for hand pose estimation on visible images, in this work, we do not incorporate the depth images to the training model, thereby making the pose estimation more appealing for the situations where the access to the depth images is not viable. Besides, by training a unified model for both X-ray

and visible images, where each modality captures different information which complements each other, we elevate the accuracy of the overall model. We present a cascaded network architecture which utilizes a template mesh to estimate the deformations in the 2D images where the estimation is propagated in different cascaded levels to increase the accuracy.

98 Expression, Affect, Action Unit Recognition: Aff-Wild2, 297 Multi-Task Learning and ArcFace

Dimitrios Kollias (Imperial College London), Stefanos Zafeiriou (Imperial College London)

Affective computing has been largely limited in terms of available data resources. The need to collect and annotate diverse in-the-wild datasets has become apparent with the rise of deep learning models, as the default approach to address any computer vision task. Some in-the-wild databases have been recently proposed. However: i) their size is small, ii) they are not audiovisual, iii) only a small part is manually annotated, iv) they contain a small number of subjects, or v) they are not annotated for all main behavior tasks (valence-arousal estimation, action unit detection and basic expression classification). To address these, we substantially extend the largest available in-the-wild database (Aff-Wild) to study continuous emotions such as valence and arousal. Furthermore, we annotate parts of the database with basic expressions and action units. As a consequence, for the first time, this allows the joint study of all three types of behavior states. We call this database Aff-Wild2. We conduct extensive experiments with CNN and CNN-RNN architectures that use visual and audio modalities; these networks are trained on Aff-Wild2 and their performance is then evaluated on 10 publicly available emotion databases. We show that the networks achieve state-of-the-art performance for the emotion recognition tasks. Additionally, we adapt the ArcFace loss function in the emotion recognition context and use it for training two new networks on Aff-Wild2 and then re-train them in a variety of diverse expression recognition databases. The networks are shown to improve the existing state-of-the-art. The database, emotion recognition models and source code are available at <http://ibug.doc.ic.ac.uk/resources/aff-wild2>.

99 Two-stage Image Classification Supervised by a Single Teacher Single Student Model

Jianhang Zhou (University of Macau), Shaoning Zeng (University of Macau, Huizhou University), Bob Zhang (University of Macau)

The two-stage strategy has been widely used in image classification. However, these methods barely take the classification criteria of the first stage into consideration in the second prediction stage. In this paper, we propose a novel two-stage representation method (TSR), and convert it to

a Single-Teacher Single-Student (STSS) problem in our two-stage image classification framework. We seek the nearest neighbours of the test sample to choose candidate target classes. Meanwhile, the first stage classifier is formulated as the teacher, which holds the classification scores. The samples of the candidate classes are utilized to learn a student classifier based on L2-minimization in the second stage. The student will be supervised by the teacher classifier, which approves the student only if it obtains a higher score. In actuality, the proposed framework generates a stronger classifier by staging two weaker classifiers in a novel way. The experiments conducted on several face and object databases show that our proposed framework is effective and outperforms multiple popular classification methods.

100 MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language

HAMID VAEZI JOZE (Microsoft), Oscar Koller (Microsoft)

Sign language recognition is a challenging and often underestimated problem comprising multi-modal articulators (handshape, orientation, movement, upper body and face) that integrate asynchronously on multiple streams. Learning powerful statistical models in such a scenario requires much data, particularly to apply recent advances of the field. However, labeled data is a scarce resource for sign language due to the enormous cost of transcribing these unwritten languages. We propose the first real-life large-scale sign language data set comprising over 25,000 annotated videos, which we thoroughly evaluate with state-of-the-art methods from sign and related action recognition. Unlike the current state-of-the-art, the data set allows to investigate the generalization to unseen individuals (signer-independent test) in a realistic setting with over 200 signers. Previous work mostly deals with limited vocabulary tasks, while here, we cover a large class count of 1000 signs in challenging and unconstrained real-life recording conditions. We further propose I3D, known from video classifications, as a powerful and suitable architecture for sign language recognition, outperforming the current state-of-the-art by a large margin. The data set is publicly available to the community.

101 Trajectory Space Factorization for Deep Video-Based 3D Human Pose Estimation

Jiahao Lin (National University of Singapore), Gim Hee Lee (National University of Singapore)

Existing deep learning approaches on 3d human pose estimation for videos are either based on Recurrent or Convolutional Neural Networks (RNNs or CNNs). However, RNN-based frameworks can only tackle sequences with limited frames because sequential models are sensitive to bad frames

and tend to drift over long sequences. Although existing CNN-based temporal frameworks attempt to address the sensitivity and drift problems by concurrently processing all input frames in the sequence, the existing state-of-the-art CNN-based framework is limited to 3d pose estimation of a single frame from a sequential input. In this paper, we propose a deep learning-based framework that utilizes matrix factorization for sequential 3d human poses estimation. Our approach processes all input frames concurrently to avoid the sensitivity and drift problems, and yet outputs the 3d pose estimates for every frame in the input sequence. More specifically, the 3d poses in all frames are represented as a motion matrix factorized into a trajectory bases matrix and a trajectory coefficient matrix. The trajectory bases matrix is precomputed from matrix factorization approaches such as Singular Value Decomposition (SVD) or Discrete Cosine Transform (DCT), and the problem of sequential 3d pose estimation is reduced to training a deep network to regress the trajectory coefficient matrix. We demonstrate the effectiveness of our framework on long sequences by achieving state-of-the-art performances on multiple benchmark datasets. Our source code is available at: <https://github.com/jiahaoLjh/trajectory-pose-3d>.

102 **BIRD: Learning Binary and Illumination Robust Descriptor for Face Recognition**

Zhuo Su (University of Oulu), Matti Pietikäinen (University of Oulu), Li Liu (University of Oulu)

Recently face recognition has made significantly progress due to the advancement of large scale Deep Convolutional Neural Network (DeepCNNs). Despite the great success, the known deficiencies of DeepCNNs have not been addressed, such as the need for too much labeled training data, energy hungry, lack of theoretical interpretability, lack of robustness to image transformations and degradations, and vulnerable to attacks, which limits DeepCNNs to be used in many real world applications. Therefore, these factors make previous predominating Local Binary Patterns (LBP) based face recognition methods still irreplaceable.

In this paper we propose a novel approach called BIRD (learning Binary and Illumination Robust Descriptor) for face representation, which nicely balances the three criteria: distinctiveness, robustness, and computationally inexpensive cost. We propose to learn discriminative and compact binary codes directly from six types of Pixel Difference Vectors (PDVs). For each type of binary codes, we cluster and pool these compact binary codes to obtain a histogram representation of each face image. Six global histograms derived from six types of learned compact binary codes are fused for the final face recognition. Experimental results on the CAS_PERL_R1 and LFW databases indicate the performance of our BIRD surpasses all previous binary based face recognition methods on the two evaluated datasets. More impressively, the proposed

BIRD is shown to be highly robust to illumination changes, and produces 89.5% on the CAS_PEAL_R1 illumination subset, which, we believe, is so far the best reported results on this dataset. Our code is made available.

103 Construct Dynamic Graphs for Hand Gesture Recognition via Spatial-Temporal Attention

Yuxiao Chen (Rutgers University), Long Zhao (Rutgers University), Xi Peng (University of Delaware), Jianbo Yuan (University of Rochester), Dimitris Metaxas (Rutgers University)

We propose a Dynamic Graph-Based Spatial-Temporal Attention (DG-STA) method for hand gesture recognition. The key idea is to first construct a fully-connected graph from a hand skeleton, where the node features and edges are then automatically learned via a self-attention mechanism that performs in both spatial and temporal domains. We further propose to leverage the spatial-temporal cues of joint positions to guarantee robust recognition in challenging conditions. In addition, a novel spatial-temporal mask is applied to significantly cut down the computational cost by 99%. We carry out extensive experiments on benchmarks (DHG-14/28 and SHREC'17) and prove the superior performance of our method compared with the state-of-the-art methods. The source code can be found at <https://github.com/yuxiaochen1103/DG-STA>.

104 Annealed Label Transfer for Face Expression Recognition

Corneliu Florea (University Politehnica of Bucharest), Laura Florea (University Politehnica of Bucharest), Mihai Badea (Image Processing and Analysis Laboratory, University Politehnica of Bucharest), Constantin Vertan (University Politehnica of Bucharest), Andrei Racoviteanu (University Politehnica of Bucharest)

In this paper we propose a method for recognizing facial expressions using information from a pair of domains: one has labelled data and one with unlabelled data. As the two domains may differ in distribution, we depart from the traditional semi-supervised framework towards a transfer learning approach. In our method, which we call Annealed Label Transfer, the deep learner explores and predicts labels on the unsupervised part, yet, in order to prevent too much confidence in its predictions (as domains are not identical), the global error is regularized with a randomization input via an annealing process. The method's evaluation is carried out on a set of four scenarios. The first two are standard benchmarks with expression faces in the wild, while the latter two have been little attempted before: face expression recognition in children and the study of the separability of anxiety-originated expressions in the wild. In all cases we show the superiority of the proposed method with respect to the strong baselines.

105 **FlickerNet: Adaptive 3D Gesture Recognition from Sparse Point Clouds**

Yuecong Min (Institute of Computing Technology, Chinese Academy of Sciences), Xiujuan Chai (Agricultural Information Institute), Lei Zhao (HUAWEI Technologies Co., Ltd.), Xilin Chen (Institute of Computing Technology, Chinese Academy of Sciences)

Recent studies on gesture recognition use deep convolutional neural networks (CNNs) to extract spatiotemporal features from individual frames or short video clips. However, extracting features frame-by-frame will bring a lot of redundant and ambiguous gesture information. Inspired by the flicker fusion phenomena, we propose a simple but efficient network, called FlickerNet, to recognize gesture from a sequence of sparse point clouds sampled from depth videos. Different from the existing CNN-based methods, FlickerNet can adaptively recognize hand postures and hand motions from the flicker of gestures: the point clouds of the stable hand postures and the sparse point-cloud motion for fast hand motions. Notably, FlickerNet significantly outperforms the previous state-of-the-art approaches on two challenging datasets with much higher computational efficiency.

106 **Pose-Aware Face Alignment based on CNN and 3DMM**

Songjiang Li (Peking University), Honggai Li (Peking University), Jinshi Cui (Peking University), Hongbin Zha (Peking University)

Pose variation is one of the tough challenges in the area of face alignment. In this paper, we showed how a framework based on convolutional neural networks (CNN) and 3D morphable models (3DMM), can explicitly handle pose variations for robust facial landmark localization. Since human faces are usually horizontally symmetric, a left-looking face (from the viewer's perspective) is equivalent to a right-looking face after a horizontal flip. Based on the symmetry, we focus on frontal and right-looking faces. We divided landmarks into two categories, SL (stable landmarks) and UL (unstable landmarks), according to their visibility across poses. A sophisticated CNN model was trained to directly estimate the SLs, whereas a following 3DMM model generated the remaining ULs. A series of experiments were conducted on popular datasets, such as 300-W, COFW, and AFLW. The results showed that the proposed method reduced errors for large-pose samples without degrading the performance of semi-frontal faces, thus demonstrating the superiority and robustness of our method.

4 **Unmasking the Devil in the Details:What Works for Deep** 107 **Facial Action Coding?**

Koichiro Niinuma (Fujitsu Laboratories of America, Inc.), Laszlo Jeni (Carnegie Mellon University), Jeffrey Cohn (University of Pittsburgh), Itir Onal Ertugrul (Carnegie Mellon University)

The performance of automated facial expression coding has improving

steadily as evidenced by results of the latest Facial Expression Recognition and Analysis (FERA 2017) Challenge. Advances in deep learning techniques have been key to this success. Yet the contribution of critical design choices remains largely unknown. Using the FERA 2017 database, we systematically evaluated design choices in pre-training, feature alignment, model size selection, and optimizer details. Our findings vary from the counter-intuitive (e.g., generic pre-training outperformed face-specific models) to best practices in tuning optimizers. Informed by what we found, we developed an architecture that exceeded state-of-the-art on FERA 2017. We achieved a 3.5% increase in F1 score for occurrence detection and a 5.8% increase in ICC for intensity estimation.

108 Large Margin Loss for Learning Facial Movements from Pseudo-Emotions

Andrei Racoviteanu (University Politehnica of Bucharest), Mihai Badea (Image Processing and Analysis Laboratory, University Politehnica of Bucharest), Corneliu Florea (University Politehnica of Bucharest), Laura Florea (University Politehnica of Bucharest), Constantin Vertan (University Politehnica of Bucarest)

In this paper we propose a large margin based loss function that enables information transfer from an unsupervised domain to a supervised one. The proposed methodology is applied in the context of face expression analysis. Categorical expressions are easier to understand and mutually exclusive, yet annotation is difficult and arguable. In contrast, facial movements encoded as action units have gained wider acceptance. Our strategy assumes self labeling images in the wild with pseudo-emotions to better learn action units. The proposed method is tested in two challenging scenarios with expressions in the wild, showing improved performance with respect to the baseline.

109 Body Part Alignment and Temporal Attention Pooling for Video-Based Person Re-Identification

Michael Jones (Mitsubishi Electric Research Laboratories), Sai Saketh Rambhatla (University of Maryland)

We present a novel deep neural network for video-based person re-identification that is designed to address two of the major issues that make this problem difficult. The first is dealing with misalignment between cropped images of people. For this we take advantage of the OpenPose network to localize different body parts so that corresponding regions of feature maps can be compared. The second is dealing with bad frames in a video sequence. These are typically frames in which the person is occluded, poorly localized or badly blurred. For this we design a temporal attention network that analyzes feature maps of multiple frames to assign different weights to each frame. This allows more useful frames to receive

more weight when creating an aggregated feature vector representing an entire sequence. Our resulting deep network improves over the state of the art on all three standard test sets for video-based person re-id (PRID2011, iLIDS-VID and MARS).

110 Automatic 4D Facial Expression Recognition via Collaborative Cross-domain Dynamic Image Network

Muzammil Behzad (University of Oulu), Nhat Vo (University of Oulu), Xiaobai Li (University of Oulu), Guoying Zhao (University of Oulu)

This paper proposes a novel 4D Facial Expression Recognition (FER) method using Collaborative Cross-domain Dynamic Image Network (CCDN). Given a 4D data of face scans, we first compute its geometrical images, and then combine their correlated information in the proposed cross-domain image representations. The acquired set is then used to generate cross-domain dynamic images (CDI) via rank pooling that encapsulates facial deformations over time in terms of a single image. For the training phase, these CDIs are fed into an end-to-end deep learning model, and the resultant predictions collaborate over multi-views for performance gain in expression classification. Furthermore, we propose a 4D augmentation scheme that not only expands the training data scale but also introduces significant facial muscle movement patterns to improve the FER performance. Results from extensive experiments on the commonly used BU-4DFE dataset under widely adopted settings show that our proposed method outperforms the state-of-the-art 4D FER methods by achieving an accuracy of 96.5% indicating its effectiveness.

111 Enhanced Normalized Mean Error loss for Robust Facial Landmark detection

Shenqi Lai (MeituanDianping Group), Zhenhua Chai (MeituanDianping Group), Huanhuan Meng (MeituanDianping Group), Shengxi Li (MeituanDianping Group), Mengzhao Yang (MeituanDianping Group), Xiaoming Wei (MeituanDianping Group)

Normalized Mean Error (NME) is one of the most popular evaluation metrics in facial landmark detection benchmark. However, the commonly used loss functions (L1 and L2) are not designed to optimize NME directly, and thus there might be a gap between optimizing the distance losses for regressing the parameters of landmark coordinates and minimizing this metric value. In this paper, we will try to address this issue, and propose a novel loss function named Enhanced Normalized Mean Error (ENME) loss, which will consider both the final metric and the attention mechanism for different NME intervals. In order to evaluate the effectiveness of our proposed loss, we design and train a light-weight regressing model we call Thin Residual Network (TRNet). Extensive experiments are conducted on three popular public datasets such as AFLW, COFW and challenging 300W,

and the results show that TRNet when trained with the enhanced NME loss will exhibit better performance than the state of the art methods.

112 SRN: Stacked Regression Network for Real-time 3D Hand Pose Estimation

Pengfei Ren (Beijing University of Posts and Telecommunications), Haifeng Sun (Beijing University of Posts and Telecommunications), Jingyu Wang (Beijing University of Posts and Telecommunications), Qi Qi (Beijing University of Posts and Telecommunications), Weiting Huang (Beijing University of Posts and Telecommunications)

Recently, most of state-of-the-art methods are based on 3D input data, because 3D data capture more spatial information than the depth image. However, these methods either require a complex network structure or time-consuming data preprocessing and post-processing. We present a simple and accurate method for 3D hand pose estimation from a 2D depth image. This is achieved by a differentiable re-parameterization module, which constructs 3D heatmaps and unit vector fields from joint coordinates directly. Taking the spatial-aware representations as intermediate features, we can easily stack multiple regression modules to capture spatial structures of depth data efficiently for accurate and robust estimation. Furthermore, we explore multiple good practices to improve the performance of the 2D CNN for 3D hand pose estimation. Experiments on four challenging hand pose datasets show that our proposed method outperforms all state-of-the-art methods.

113 Face Anti-Spoofing via Sample Learning Based Recurrent Neural Network (RNN)

Usman Muhammad (University of Oulu), Abdenour Hadid (University of Oulu), Wheidima Melo (University of Oulu), Tuomas Kristian Holmberg (University of Oulu)

Face biometric systems are vulnerable to spoofing attacks because of criminals who are developing different techniques such as print attack, replay attack, 3D mask attack, etc. to easily fool the face recognition systems. To improve the security measures of biometric systems, we propose a simple and effective architecture called sample learning based recurrent neural network (SLRNN). The proposed sample learning is based on sparse filtering which is applied for augmenting the features by leveraging Residual Networks (ResNet). The augmented features form as a sequence, which are fed into a Long Short-Term Memory (LSTM) network for constructing the final representation. We show that for face anti-spoofing task, incorporating sample learning into recurrent structures learn more meaningful representations to LSTM with much fewer model parameters. Experimental studies on MSU and CASIA dataset demonstrate that the proposed SLRNN has a superior performance than state-of-the-art methods used

now.

114 PAttNet: Patch-attentive deep network for action unit detection

Itir Onal Ertugrul (Carnegie Mellon University), Laszlo Jeni (Carnegie Mellon University), Jeffrey Cohn (University of Pittsburgh)

Facial action units (AUs) refer to specific facial locations. Recent efforts in automatic AU detection have focused on learning their representations. Two factors have limited progress. One is that current approaches implicitly assume that facial patches are robust to head rotation. The other is that the relation between patches and AUs is pre-defined or ignored. Both assumptions are problematic. We propose a patch-attentive deep network called PAttNet for AU detection that learns mappings of patches and AUs, controls for 3D head and face rotation, and exploits co-occurrence among AUs. We encode patches with separate convolutional neural networks (CNNs) and weight the contribution of each patch to detection of specific AUs using a sigmoid patch attention mechanism. Unlike conventional softmax attention mechanisms, a sigmoidal attention mechanism allows multiple patches to contribute to detection of specific AUs. The latter is important because AUs often co-occur and multiple patches may be needed to detect them reliably. On the BP4D dataset, PAttNet improves upon state-of-the-art by 3.7%. Visualization of the learned attention maps reveal power of this patch-based approach.

115 End-to-End 3D Hand Pose Estimation from Stereo Cameras

Yuncheng Li (Snap Inc.), Zehao Xue (Snap Inc.), Yingying Wang (Snap Inc.), Liuhao Ge (Nanyang Technological University), Zhou Ren (Wormpex AI Research), Jonathan Rodriguez (Snap Inc.)

This work proposes an end-to-end approach to estimate full 3D hand pose from stereo cameras. Most existing methods of estimating hand pose from stereo cameras apply stereo matching to obtain depth map and use depth-based solution to estimate hand pose. In contrast, we propose to bypass the stereo matching and directly estimate the 3D hand pose from the stereo image pairs. The proposed neural network architecture extends from any keypoint predictor to estimate the sparse disparity of the hand joints. In order to effectively train the model, we propose a large scale synthetic dataset that is composed of stereo image pairs and ground truth 3D hand pose annotations. Experiments show that the proposed approach outperforms the existing methods based on the stereo depth.

116 TAGAN: Tonality Aligned Generative Adversarial Networks for Realistic Hand Pose Synthesis

Liangjian Chen (University of California, Irvine), Shih-Yao Lin (Tencent Medical AI Lab), Yusheng Xie (Tencent Medical AI Lab), Hui Tang (Tencent Medical AI Lab), Yufan Xue (workday), Yen-Yu Lin (Academia Sinica), Xiaohui Xie (University of California, Irvine), Wei Fan (Tencent)

Despite recent progress, estimating 3D hand poses from single RGB images remains challenging. One of the major limiting factors is the lack of sufficiently large hand pose datasets with accurate 3D hand keypoint annotations. To address this limitation, we present an efficient method for generating realistic hand poses, and show that existing algorithms for hand pose estimation can be greatly improved by augmenting training data with images of the synthetic hand poses, which come naturally with ground truth annotations. More specifically, we adopt an augmented reality simulator to synthesize hand poses with accurate 3D hand-keypoint annotations. However, these synthesized hand poses look unnatural. To produce more realistic hand poses, we propose to blend each synthetic hand pose with a real background. To this end, we develop tonality-aligned generative adversarial networks (TAGAN), which align the tonality and color distributions between synthetic hand poses and real backgrounds, and can generate high-quality hand poses. TAGAN is evaluated on the RHP, STB, and CMU-PS hand pose datasets. With the aid of the synthesized poses, our method performs favorably against the state-of-the-arts in both 2D and 3D hand pose estimation.

117 **Frustratingly Easy Person Re-Identification: Generalizing Person Re-ID in Practice**

Jieru Jia (Beijing Jiaotong University), Qiuqi Ruan (Beijing Jiaotong University), Timothy Hospedales (Edinburgh University)

Contemporary person re-identification (Re-ID) methods usually require access to data from the deployment camera network during training in order to perform well. This is because contemporary Re-ID models trained on one dataset do not generalise to other camera networks due to the domain-shift between datasets. This requirement is often the bottleneck for deploying Re-ID systems in practical security or commercial applications, as it may be impossible to collect this data in advance or prohibitively costly to annotate it. This paper alleviates this issue by proposing a simple baseline for domain generalizable (DG) person re-identification. That is, to learn a Re-ID model from a set of source domains that is suitable for application to unseen datasets out-of-the-box, without any model updating. Specifically, we observe that the domain discrepancy in Re-ID is due to style and content variance across datasets and demonstrate appropriate Instance and Feature Normalization alleviates much of the resulting domain-shift in Deep Re-ID models. Instance Normalization (IN) in early layers filters out style statistic variations and Feature Normalization (FN) in deep layers is able to further eliminate disparity in content statistics. Compared to con-

temporary alternatives, this approach is extremely simple to implement, while being faster to train and test, thus making it an extremely valuable baseline for implementing Re-ID in practice. With a few lines of code, it increases the rank 1 Re-ID accuracy by 11.8%, 33.2%, 12.8% and 8.5% on the VIPeR, PRID, GRID, and i-LIDS benchmarks respectively. Source codes are available at https://github.com/BJTUJia/person_reID_DualNorm.

Statistics and Machine Learning

118 Delving Deep into Least Square Regression Model for Subspace Clustering

Masataka Yamaguchi (NTT Corporation), Go Irie (NTT Communication Science Laboratories), Takahito Kawanishi (NTT Corporation), Kunio Kashino (NTT Corporation)

Subspace clustering is the problem of clustering data drawn from a union of multiple subspaces. The most popular subspace clustering framework in recent years is the spectral clustering-based approach, which performs subspace clustering by first computing an affinity matrix and then applying spectral clustering to it. One of the representative methods for computing an affinity matrix is the least square regression (LSR) model, which is based on the idea of self-representation. Although its efficiency and effectiveness have been empirically validated, it lacks some theoretical analysis and practicality, e.g.: absence of interpretations, lack of theoretical analysis on its robustness, absence of guidelines for choosing the hyper-parameter, and the scalability. This paper aims at providing novel insights for better understanding on LSR, and also improving its practicality. For this purpose, we present four contributions: first, we present a novel interpretation of LSR, which is based on random sampling perspective. Second, we provide novel theoretical analysis on LSR's robustness toward outliers. Third, we theoretically and empirically demonstrate that selecting a larger value for the hyper-parameter tends to result in good clustering results. Finally, we derive another equivalent form of the LSR's solution, which can be computed with less time complexity than the original form regarding the data size.

119 Variational Saccading: Efficient Inference for Large Resolution Images

Jason Ramapuram (University of Geneva), Russ Webb (Apple), Maurits Diephuis (University of Geneva), Alexandros Kalousis (AU Geneva), Frantzeska Lavda (University of Geneva)

Image classification with deep neural networks is typically restricted to images of small dimensionality such as 224×224 in Resnet models. This

limitation excludes the 4000×3000 dimensional images that are taken by modern smartphone cameras and smart devices. In this work, we aim to mitigate the prohibitive inferential and memory costs of operating in such large dimensional spaces. To sample from the high-resolution original input distribution, we propose using a smaller proxy distribution to learn the co-ordinates that correspond to regions of interest in the high-dimensional space. We introduce a new principled variational lower bound that captures the relationship of the proxy distribution's posterior and the original image's co-ordinate space in a way that maximizes the conditional classification likelihood. We empirically demonstrate on one synthetic benchmark and one real world large resolution DSLR camera image dataset that our method produces comparable results with 10x faster inference and lower memory consumption than a model that utilizes the entire original input distribution. Finally, we experiment with a more complex setting using mini-maps from Starcraft II to infer the number of characters in a complex 3d-rendered scene. Even in such complicated scenes our model provides strong localization: a feature missing from traditional classification models.

120 An Acceleration Scheme for Mini-batch, Streaming PCA

Salaheddin Alakkari (Trinity College Dublin), John Dingliana (Trinity College Dublin)

In this paper, we propose an acceleration scheme for mini-batch streaming PCA methods that are based on the Stochastic Gradient Approximation. Our scheme converges to the first $k > 1$ eigenvectors in a single data pass even when using a very small batch size. We provide empirical convergence results of our scheme based on the spiked covariance model. Our scheme does not require any prior knowledge of the data distribution and hence is well suited for streaming data scenarios. Furthermore, based on empirical evaluations using the spiked covariance model and large-scale benchmark datasets, we find that our acceleration scheme outperforms related state-of-the-art online PCA approaches including SGA, Incremental PCA and Candid Covariance-free Incremental PCA.

121 A Generic Active Learning Framework for Class Imbalance Applications

Aditya Bhattacharya (Florida State University), Ji Liu (University of Rochester), Shayok Chakraborty (Florida State University)

Active learning algorithms automatically identify the most informative samples from large amounts of unlabeled data and tremendously reduce human annotation effort in inducing a robust machine learning model. Real-world data often exhibit significantly skewed class distributions, where samples from one class dominate over the other. While active learning has been extensively studied, there have been limited research efforts

to develop active learning algorithms specifically for class imbalance applications. In this paper, we propose a novel framework to address this research challenge. We pose the active sample selection as a constrained optimization problem and derive a linear programming relaxation to select a batch of samples. Contrary to existing algorithms, our framework is generic and is applicable to both binary and multi-class problems, where the imbalance may exist across multiple classes. Our extensive empirical studies on four vision datasets spanning three different application domains (face, facial expression and handwritten digits recognition) with varied degrees of class imbalance demonstrate the promise and potential of the method for real-world imbalanced data applications.

122 Transductive Learning Via Improved Geodesic Sampling

Youshan Zhang (Lehigh University), Brian Davison (Lehigh University), Sihong Xie (Lehigh University)

Transductive learning exploits the connection between training and test data to improve classification performance, and the geometry of the manifold underlying the training and the test data is essential to make this connection explicit. However, existing approaches primarily focused on the Grassmannian manifold, while much is less known for other manifolds which can bring better computational and learning performance. In this paper, we define a novel, more general formulation of geodesic sampling on Riemannian manifolds (GSM), which is applicable to manifolds beyond Grassmannian. We demonstrate the use of the GSM model on three manifolds. To provide practical guidance for classification, we explore hyperparameter settings with extensive experiments and propose a Target-focused GSM (TGSM) with a single sample that is close to the target (test data) on a spherical manifold. These choices produce the highest accuracy and least computation time over state-of-the-art methods.

123 A General Transductive Regularizer for Zero-Shot Learning

Huaqi Mao (Nanjing University of Science and Technology), Haofeng Zhang (Nanjing University of Science and Technology), Shidong Wang (University of East Anglia), Yang Long (Newcastle University), Longzhi Yang (Northumbria University)

Zero Shot Learning (ZSL) has attracted much attention due to its ability to recognize objects of unseen classes, which is realized by transferring knowledge from seen classes through semantic embeddings. Since the seen classes and unseen classes usually have different distributions, conventional inductive ZSL often suffers from the domain shift problem. Transductive ZSL is a type of method for solving such a problem. However, the regularizers of conventional transductive methods are different

from each other, and cannot be applied to other methods. In this paper, we propose a General Transductive Regularizer (GTR), which assigns each unlabeled sample to a fixed attribute by defining a Kullback-Leibler Divergence (KLD) objective. To this end, GTR can be easily applied to many compatible linear and deep inductive ZSL models. Extensive experiments on both linear and deep methods are conducted on four popular datasets, and the results show that GTR can significantly improve the performance comparing to its original inductive method, and also outperform some state-of-the-art methods, especially the extension on deep model.

3 A Less Biased Evaluation of Out-of-distribution Sample 124 Detectors

Alireza Shafaei (The University of British Columbia), Mark Schmidt (University of British Columbia), Jim Little (University of British Columbia)

In the real world, a learning system could receive an input that is unlike anything it has seen during training. Unfortunately, out-of-distribution samples can lead to unpredictable behaviour. We need to know whether any given input belongs to the population distribution of the training/evaluation data to prevent unpredictable behaviour in deployed systems. A recent surge of interest in this problem has led to the development of sophisticated techniques in the deep learning literature. However, due to the absence of a standard problem definition or an exhaustive evaluation, it is not evident if we can rely on these methods. What makes this problem different from a typical supervised learning setting is that the distribution of outliers used in training may not be the same as the distribution of outliers encountered in the application. Classical approaches that learn inliers vs. outliers with only two datasets can yield optimistic results. We introduce OD-test, a three-dataset evaluation scheme as a more reliable strategy to assess progress on this problem. We present an exhaustive evaluation of a broad set of methods from related areas on image classification tasks. Contrary to the existing results, we show that for realistic applications of high-dimensional images the previous techniques have low accuracy and are not reliable in practice.

Motion, Flow and Tracking

125 Relation-aware Multiple Attention Siamese Networks 290 for Robust Visual Tracking

Fangyi Zhang (Chinese Academy of Sciences), Bingpeng Ma (Chinese Academy of Sciences), Hong Chang (Chinese Academy of Sciences), Shiguang Shan (Chinese Academy of Sciences),

Xilin Chen (Institute of Computing Technology, Chinese Academy of Sciences)

Partial occlusion is a challenging problem in visual object tracking. Neither Siamese network based trackers nor conventional part-based trackers can address this problem successfully. In this paper, inspired by the fact that attentions can make the model focus on the most salient regions of an image, we propose a new method named Relation-aware Multiple Attention (RMA) to address the partial occlusion problem. In the RMA module, part features generated from a set of attention maps can represent the discriminative parts of the target and ignore the occluded ones. Meanwhile, an attention regularization term is proposed to force the multiple attention maps to localize diverse local patterns. Besides, we incorporate relation-aware compensation to adaptively aggregate and distribute part features to capture the semantic dependency among them. We integrate the RMA module into Siamese matching networks and verify the superior performance of the RMA-Siam tracker on five visual tracking benchmarks, including VOT-2016, VOT-2017, LaSOT, OTB-2015 and TrackingNet.

126 Spatial Transformer Spectral Kernels for Deformable Im- 291 age Registration

Ebrahim Al Safadi (Oregon Health and Science University), Xubo Song (Oregon Health and Science University)

Recent advances in kernel methods have made them more attractive tools for spatial transformation models. In this work, Spatial Transformer Spectral Kernels are introduced as a framework for Deformable Image Registration. The transformation is restricted to live in a Reproducing Kernel Hilbert Space, and Generalized Spectral Mixture kernels are used as the reproducing kernels. This combination results in a powerful but simple regularization model that can adapt to many deformation scenarios with nonstationary and possibly long range nonmonotonic relations across the pixels. Our formulation leads to a Kernel Ridge Regression transform that is pre-computed once before optimization, and unlike most developments in image registration the loss function explicitly pairs this transform with a specific interpolation function. We derive a closed-form gradient of the loss function with respect to the spatial transformation and interpolation function which enhances registration results. Based on our evaluation, while being simpler our method can perform comparably to more complex Large Deformation Diffeomorphic Metric Mapping models in terms of reducing the intensity sum of squared differences, and can provide a more accurate estimate of the underlying displacement field.

127 Tracking the Known and the Unknown by Leveraging Se- 292 mantic Information

Ardhendu Shekhar Tripathi (ETH Zurich), Martin Danelljan (ETH Zurich), Luc Van Gool (ETH Zurich), Radu Timofte (ETH Zurich)

Current research in visual tracking is largely focused on the generic case, where no prior knowledge about the target object is assumed. However, many real-world tracking applications stem from specific scenarios where the class or type of object is known. In this work, we propose a tracking framework that can exploit this semantic information, without sacrificing the generic nature of the tracker. In addition to the target-specific appearance, we model the class of the object through a semantic module that provides complementary class-specific predictions. By further integrating a semantic classification module, we can utilize the learned class-specific models even if the target class is unknown. Our unified tracking architecture is trained end-to-end on large scale tracking datasets by exploiting the available semantic metadata. Comprehensive experiments are performed on five tracking benchmarks. Our approach achieves state-of-the-art performance while operating at real-time frame-rates. The code and the trained models are available at <https://tracking.vision.ee.ethz.ch/track-known-unknown/>.

128 Tracking Holistic Object Representations

293

Axel Sauer (Technical University of Munich), Elie Aljalbout (Technical University of Munich), Sami Haddadin (Technical University of Munich)

Recent advances in visual tracking are based on siamese feature extractors and template matching. For this category of trackers, latest research focuses on better feature embeddings and similarity measures. In this work, we focus on building holistic object representations for tracking. We propose a framework that is designed to be used on top of previous trackers without any need for further training of the siamese network. The framework leverages the idea of obtaining additional object templates during the tracking process. Since the number of stored templates is limited, our method only keeps the most diverse ones. We achieve this by providing a new diversity measure in the space of siamese features. The obtained representation contains information beyond the ground truth object location provided to the system. It is then useful for tracking itself but also for further tasks which require a visual understanding of objects. Strong empirical results on tracking benchmarks indicate that our method can improve the performance and robustness of the underlying trackers while barely reducing their speed. In addition, our method is able to match current state-of-the-art results, while using a simpler and older network architecture and running three times faster.

129 Video Upright Adjustment and Stabilization

Jucheol Won (DGIST), Sunghyun Cho (POSTECH)

We propose a novel video upright adjustment method that can reliably correct slanted video contents. Our approach combines deep learning and Bayesian inference to estimate accurate rotation angles from video frames. We train a convolutional neural network to obtain initial estimates of the rotation angles of input video frames. The initial estimates are temporally inconsistent and inaccurate. To resolve this, we use Bayesian inference. We analyze estimation errors of the network, and derive an error model. Based on the error model, we formulate video upright adjustment as a maximum a posteriori problem where we estimate consistent rotation angles from the initial estimates. Finally, we propose a joint approach to video stabilization and upright adjustment to minimize information loss. Experimental results show that our video upright adjustment method can effectively correct slanted video contents, and our joint approach can achieve visually pleasing results from shaky and slanted videos.

130 Video Stitching for Linear Camera Arrays

Wei-Sheng Lai (University of California, Merced), Orazio Gallo (NVIDIA Research), Jinwei Gu (NVIDIA), Deqing Sun (Google), Ming-Hsuan Yang (University of California at Merced), Jan Kautz (NVIDIA)

Despite the long history of image and video stitching research, existing academic and commercial solutions still produce strong artifacts. In this work, we propose a wide-baseline video stitching algorithm that is temporally stable and tolerant to strong parallax. Our key insight is that stitching can be cast as a problem of learning a smooth spatial interpolation between the input videos. To solve this problem, inspired by pushbroom cameras, we introduce a fast pushbroom interpolation layer and propose a novel pushbroom stitching network, which learns a dense flow field to smoothly align the multiple input videos with spatial interpolation. Our approach outperforms the state-of-the-art by a significant margin, as we show with a user study, and has immediate applications in many areas such as virtual reality, immersive telepresence, autonomous driving, and video surveillance.

131 Learning Target-aware Attention for Robust Tracking with Conditional Adversarial Network

Xiao Wang (Anhui University), Rui Yang (Anhui university), Tao Sun (Anhui university), Bin Luo (Anhui University)

Many of current visual trackers are based on tracking-by-detection framework which attempts to search target object within a local search window for each frame. Although they have achieved appealing performance, however, their localization and scale handling often perform poorly in extremely challenging scenarios, such as heavy occlusion and large deformation.

tion due to two major reasons: i) They simply set a local searching window using temporal context, which may not cover the target at all and therefore cause tracking failure. ii) Some of them adopt image pyramid strategy to handle scale variations, which heavily relies on target localization, and thus can be easily disturbed when the localization is unreliable. To handle these issues, this paper presents a novel and general target-aware attention learning approach to simultaneously achieve target localization and scale handling. Through conditional generative adversarial network (CGAN), attention maps are produced to generate the proposals with high-quality locations and scales, and perform object tracking via multi-domain CNN. The proposed approach is efficient and effective, needs small amount of training data, and improves the tracking-by-detection framework significantly. Extensive experiments have shown the proposed approach outperforms most of recent state-of-the-art trackers on several visual tracking benchmarks, and provides improved robustness for fast motion, scale variation as well as heavy occlusion. The project page of this paper can be found at: <https://sites.google.com/view/globalattentiontracking/home>.

8 Bilinear Siamese Networks with Background Suppression for Visual Object Tracking

Hankyeol Lee (Korea Advanced Institute of Science and Technology), Seokeon Choi (Korea Advanced Institute of Science and Technology), Youngeun Kim (Korea Advanced Institute of Science and Technology), Changick Kim (Korea Advanced Institute of Science and Technology)

In recent years, siamese networks have shown to be useful for visual tracking with high accuracy and real-time speed. However, since the networks only use the output of the last convolution layer, low-level feature maps which provide important spatial details for visual tracking are ignored. In this paper, we propose bilinear siamese networks for visual object tracking to take into account both high- and low-level feature maps. To effectively incorporate feature maps extracted from multiple layers, we adopt factorized bilinear pooling into our network. Also, we introduce a novel background suppression module to reduce the background interference. This module collects negative feature maps for the background in the first frame and suppresses the background information during tracking. Therefore, the module makes the tracker more robust to the background interference. Experimental results on the OTB-50 and OTB-100 benchmarks demonstrate that the proposed tracker has comparable performance with that of the state-of-the-art trackers while running in real-time.

11 Learning to Focus and Track Extreme Climate Events

Sookyung Kim (Lawrence Livermore National Laboratory), Sunghyun Park (Korea University), Sunghyo Chung (Korea University), Joonseok Lee (Google Research), Yunsung Lee (Korea Uni-

versity), Hyojin Kim (LLNL), Prabhat (Lawrence Berkeley National Laboratory), Jaegul Choo (Korea University)

This paper tackles the task of extreme climate event tracking. It has unique challenges compared to other visual object tracking problems, including a wider range of spatio-temporal dynamics, the unclear boundary of the target, and the shortage of a labeled dataset. We propose a simple but robust end-to-end model based on multi-layered ConvLSTMs, suitable for climate event tracking. It first learns to imprint the location and the appearance of the target at the first frame in an auto-encoding fashion. Next, the learned feature is fed to the tracking module to track the target in subsequent time frames. To tackle the data shortage problem, we propose data augmentation based on conditional generative adversarial networks. Extensive experiments show that the proposed framework significantly improves tracking performance of a hurricane tracking task over several state-of-the-art methods.

134 Features for Ground Texture Based Localization - A Survey

Jan Fabian Schmid (Robert Bosch GmbH), Stephan F. Simon (Robert Bosch GmbH), Rudolf Mester (NTNU Trondheim)

Ground texture based vehicle localization using feature-based methods is a promising approach to achieve infrastructure-free high-accuracy localization. In this paper, we provide the first extensive evaluation of available feature extraction methods for this task, using separately taken image pairs as well as synthetic transformations. We identify AKAZE, SURF and CenSurE as best performing keypoint detectors, and find pairings of CenSurE with the ORB, BRIEF and LATCH feature descriptors to achieve greatest success rates for incremental localization, while SIFT stands out when considering severe synthetic transformations as they might occur during absolute localization.

135 Self-supervised Video Representation Learning for Correspondence Flow

Zihang Lai (University of Oxford), Weidi Xie (University of Oxford)

The objective of this paper is self-supervised learning of feature embeddings from video, suitable for correspondence flow. We leverage the natural spatial-temporal coherence of appearance in videos, to create a model that learns to reconstruct a target frame by copying colors from a reference frame. We make three contributions: First, we introduce a simple information bottleneck that enforces the model to learn robust features for correspondence matching, and avoids it learning trivial solutions, e.g. matching by low-level color information. Second, we propose

to learn the matching over a longer temporal window in videos. To make the model more robust to complex object deformation, occlusion, i.e. the problem of tracker drifting, we formulate a recursive model and trained with scheduled sampling and cycle consistency. Third, we evaluate the approach by first training on the Kinetics dataset using self-supervised learning, and then directly applied for DAVIS video segmentation and JH-MDB keypoint tracking. On both tasks, our approach has outperformed all previous methods by a significant margin. The source code will be released at <https://github.com/zlai0/CorrFlow>.

136 **PMnet: Learning of Disentangled Pose and Movement for Unsupervised Motion Retargeting**

Jongin Lim (Seoul National University), Hyung Jin Chang (University of Birmingham), Jin Young Choi (Seoul National University)

In this paper, we propose a deep learning framework for unsupervised motion retargeting. In contrast to the existing method, we decouple the motion retargeting process into two parts that explicitly learn poses and movements of a character. Here, the first part retargets the pose of the character at each frame, while the second part retargets the character's overall movement. To realize these two processes, we develop a novel architecture referred to as the pose-movement network (PMnet), which separately learns frame-by-frame poses and overall movement. At each frame, to follow the pose of the input character, PMnet learns how to make the input pose first and then adjusts it to fit the target character's kinematic configuration. To handle the overall movement, a normalizing process is introduced to make the overall movement invariant to the size of the character. Along with the normalizing process, PMnet regresses the overall movement to fit the target character. We then introduce a novel loss function that allows PMnet to properly retarget the poses and overall movement. The proposed method is verified via several self-comparisons and outperforms the state-of-the-art (sota) method by reducing the motion retargeting error (average joint position error) from 7.68 (sota) to 1.95 (ours).

Action and Event Recognition

167 **Forecasting Future Action Sequences with Neural Memory Networks**

Harshala Gammulle (Queensland University of Technology), Simon Denman (Queensland University of Technology), Sridha Sridharan (Queensland University of Technology), Clinton

Fookes (Queensland University of Technology)

We propose a novel neural memory network based framework for future action sequence forecasting. This is a challenging task where we have to consider short-term, within sequence relationships as well as relationships in between sequences, to understand how sequences of actions evolve over time. To capture these relationships effectively, we introduce neural memory networks to our modelling scheme. We show the significance of using two input streams, the observed frames and the corresponding action labels, which provides different information cues for our prediction task. Furthermore, through the proposed method we effectively map the long-term relationships among individual input sequences through separate memory modules, which enables better fusion of the salient features. Our method outperforms the state-of-the-art approaches by a large margin on two publicly available datasets: Breakfast and 50 Salads.

145 Object Affordances Graph Network for Action Recognition

Haoliang Tan (Xi'an Jiaotong University), Le Wang (Xi'an Jiaotong University), Qilin Zhang (HERE Technologies), Zhanning Gao (Alibaba Group), Nanning Zheng (Xi'an Jiaotong University), Gang Hua (Wormpex AI Research)

Human actions often involve interactions with objects, and such action possibilities of objects were termed “affordances” in human-computer interaction (HCI) literature. To facilitate action recognition with object affordances, we propose the Object Affordances Graph (OAG), which cast human-object interaction cues into video representations via an iterative refinement procedure. With the spatio-temporal co-occurrences between human and objects captured, the Object Affordances Graph Network (OAGN) is subsequently proposed. To provide a fair evaluation of the role that object affordances could play on human action recognition, we have assembled a new dataset with additional annotated object bounding boxes to account for human-object interactions. Multiple experiments on this proposed Object-Charades dataset verify the value of including object affordances in human action recognition, specifically via the proposed OAGN, which outperforms existing state-of-the-art affordance-less action recognition methods.

169 Zero-Shot Sign Language Recognition: Can Textual Data Uncover Sign Languages?

Yunus Can Bilge (Hacettepe University), Nazli Ikizler-Cinbis (Hacettepe University), Ramazan Gokberk Cinbis (METU)

We introduce the problem of zero-shot sign language recognition (ZSSLR), where the goal is to leverage models learned over the seen sign class examples to recognize the instances of unseen signs. To this end, we propose to

utilize the readily available descriptions in sign language dictionaries as an intermediate-level semantic representation for knowledge transfer. We introduce a new benchmark dataset called ASL-Text that consists of 250 sign language classes and their accompanying textual descriptions. Compared to the ZSL datasets in other domains (such as object recognition), our dataset consists of limited number of training examples for a large number of classes, which imposes a significant challenge. We propose a framework that operates over the body and hand regions by means of 3D-CNNs, and models longer temporal relationships via bidirectional LSTMs. By leveraging the descriptive text embeddings along with these spatio-temporal representations within a zero-shot learning framework, we show that textual data can indeed be useful in uncovering sign languages. We anticipate that the introduced approach and the accompanying dataset will provide a basis for further exploration of this new zero-shot learning problem.

170 An Efficient 3D CNN for Action/Object Segmentation in Video

Rui Hou (UCF), Chen Chen (University of North Carolina at Charlotte), Mubarak Shah (University of Central Florida), Rahul Sukthankar (Google)

Convolutional Neural Network (CNN) based image segmentation has made great progress in recent years. However, video object segmentation remains a challenging task due to its high computational complexity. Most of the previous methods employ a two-stream CNN framework to handle spatial and motion features separately. In this paper, we propose an end-to-end encoder-decoder style 3D CNN for video object segmentation. The proposed approach leverages 3D separable convolutions and drastically reduces the number of trainable parameters. Instead of the popular two-stream framework, we adopt 3D CNN to aggregate spatial and temporal information. To efficiently process video, we propose 3D separable convolution for the pyramid pooling module and decoder, which dramatically reduces the number of operations while maintaining the performance. Additionally, we also extend our framework to video action segmentation by adding an extra classifier to predict the action label for actors in videos. Extensive experiments on several video datasets demonstrate the superior performance of the proposed approach for action and object segmentation compared to the state-of-the-art.

171 Pedestrian Action Anticipation using Contextual Feature Fusion in Stacked RNNs

Amir Rasouli (York University), Iuliia Kotseruba (York University), John Tsotsos (York University)

One of the major challenges for autonomous vehicles in urban environments is to understand and predict other road users' actions, in particular,

pedestrians at the point of crossing. The common approach to solving this problem is to use the motion history of the agents to predict their future trajectories. However, pedestrians exhibit highly variable actions most of which cannot be understood without visual observation of the pedestrians themselves and their surroundings. To this end, we propose a solution for the problem of pedestrian action anticipation at the point of crossing. Our approach uses a novel stacked RNN architecture in which information collected from various sources, both scene dynamics and visual features, is gradually fused into the network at different levels of processing. We show, via extensive empirical evaluations, that the proposed algorithm achieves a higher prediction accuracy compared to alternative recurrent network architectures. We conduct experiments to investigate the impact of the length of observation, time to event and types of features on the performance of the proposed method. Finally, we demonstrate how different data fusion strategies impact prediction accuracy.

172 Focused Attention for Action Recognition

Vladyslav Sydorov (Inria), Karteek Alahari (Inria), Cordelia Schmid (Inria)

Current state-of-the-art approaches to action recognition emphasize learning ConvNets on large amounts of training data, using 3D convolutions to process the temporal dimension. This approach is expensive in terms of memory usage and constitutes a major performance bottleneck of existing approaches. Further, video input data points typically include irrelevant information, along with useful features, which limits the level of detail that networks can process, regardless of the quality of the original video. Hence, models that can focus computational resources on relevant training signal are desirable. To address this problem, we rely on network-specific saliency outputs to drive an attention model that provides tighter crops around relevant video regions. We experimentally validate this approach and show how this strategy improves performance for the action recognition task.

151 Unsupervised and Explainable Assessment of Video 173 Similarity

Konstantinos Papoutsakis (University of Crete & ICS-FORTH, Greece), Antonis Argyros (CSD-UOC and ICS-FORTH)

We propose a novel unsupervised method that assesses the similarity of two videos on the basis of the estimated relatedness of the objects and their behavior and provides arguments supporting this assessment. A video is represented as a complete undirected action graph that encapsulates information on the types of objects and the way they (inter)act. The similarity of a pair of videos is estimated based on the bipartite Graph EditDistance

(GED) of the corresponding action graphs. As a consequence, on-top of estimating a quantitative measure of video similarity, our method establishes spatiotemporal correspondences between objects across videos if these objects are semantically related, if/when they interact similarly, or both. We consider this an important step towards explainable assessment of video and action similarity. The proposed method is evaluated on a publicly available dataset on the tasks of activity classification and ranking and is shown to compare favorably to state of the art supervised learning methods.

174 Dynamic Graph Modules for Modeling Object-Object Interactions in Activity Recognition

Hao Huang (University of Rochester), Luowei Zhou (University of Michigan), Wei Zhang (University of Rochester), Jason Corso (University of Michigan), Chenliang Xu (University of Rochester)

Video action recognition, as a critical problem in video understanding, has been gaining increasing attention. To identify actions induced by complex object-object interactions, we need to consider not only spatial relations among objects in a single frame but also temporal relations among different or the same objects across multiple frames. However, existing approaches modeling video representations and non-local features are either incapable of explicitly modeling relations at the object-object level or unable to handle streaming videos. In this paper, we propose a novel dynamic hidden graph module to model complex object-object interactions in videos, of which two instantiations are considered: a visual graph that captures appearance/motion changes among objects and a location graph that captures relative spatiotemporal position changes among objects. Besides, the proposed graph module allows us to process streaming videos, setting it apart from existing methods. Experimental results on two benchmark datasets, Something-Something and ActivityNet, show the competitive performance of our methods.

175 Weakly-Supervised 3D Pose Estimation from a Single Image using Multi-View Consistency

Guillaume Rochette (University of Surrey), Chris Russell (University of Surrey), Richard Bowden (University of Surrey)

We present a novel data-driven regularizer for weakly-supervised learning of 3D human pose estimation that eliminates the drift problem that effects existing approaches. We do this by moving the stereo reconstruction problem into the loss of the network itself. This avoids the need to reconstruct data prior to training and unlike previous semi-supervised approaches, avoids the need for a warm-up period of supervised training. The conceptual and implementational simplicity of our approach is fundamental to its appeal. Not only is it straightforward to augment many weakly-supervised

approaches with our additional re-projection based loss, but it is obvious how it shapes reconstructions and prevents drift. As such we believe it will be a valuable tool for any researcher working in weakly-supervised 3D reconstruction. Evaluating on Panoptic, the largest multi-camera and markerless dataset available, we obtain an accuracy that is essentially indistinguishable from a fully supervised approach making full use of 3D ground truth in training.

176 Learning Visual Actions Using Multiple Verb-Only Labels

Michael Wray (University of Bristol), Dima Damen (University of Bristol)

This work introduces verb-only representations for both recognition and retrieval of visual actions. Current methods neglect legitimate semantic ambiguities between verbs, instead choosing unambiguous subsets of verbs along with objects to disambiguate the actions. By combining multiple verbs, through hard or soft assignment, as a regression, we are able to learn a much larger vocabulary of verbs, including contextual overlaps of these verbs. We collect multi-verb annotations for three action datasets and evaluate the verb-only labelling representations for action recognition and cross-modal retrieval (video-to-text and text-to-video). We demonstrate that multi-label verb-only representations outperform conventional single verb labels. We also explore other benefits of a multi-verb representation including cross-dataset retrieval and verb type (manner and result verb types) retrieval.

154 TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition

Bishay Mina (Queen Mary University London), Georgios Zoumpourlis (Queen Mary University of London), Ioannis Patras (Queen Mary University of London)

In this paper we propose a novel Temporal Attentive Relation Network (TARN) for the problems of few-shot and zero-shot action recognition. At the heart of our network is a meta-learning approach that learns to compare representations of variable temporal length, that is, either two videos of different length (in the case of few-shot action recognition) or a video and a semantic representation such as word vector (in the case of zero-shot action recognition). By contrast to other works in few-shot and zero-shot action recognition, we a) utilise attention mechanisms so as to perform temporal alignment, and b) learn a deep-distance measure on the aligned representations at video segment level. We adopt an episode-based training scheme and train our network in an end-to-end manner. The proposed method does not require any fine-tuning in the target domain or maintaining additional representations as is the case of memory networks.

Experimental results show that the proposed architecture outperforms the state of the art in few-shot action recognition, and achieves competitive results in zero-shot action recognition.

178 A Spatiotemporal Pre-processing Network for Activity Recognition under Rain

Minah Lee (Georgia Institute of Technology), Burhan Mudassar (Georgia Institute of Technology), Taesik Na (Georgia Institute of Technology), Saibal Mukhopadhyay (Georgia Institute of Technology)

This paper presents a deep neural network (DNN) based fully spatiotemporal rain removal network, MoPE-Spatiotemporal, to enhance accuracy of activity recognition in rainy videos. The proposed network utilizes spatiotemporal information of an image sequence to detect rain streaks and recover the non-rainy image. We also present rain alert network that detects the rain fall and informs the reduction of recognition confidence under rain. Experimental results show that heavy rain can highly degrade activity recognition accuracy. MoPE-Spatiotemporal removes heavy rain better than state-of-the-art methods, and significantly improves (0.15) activity recognition accuracy in rainy videos with minimal impact on recognition accuracy in clean videos.

Biologically Inspired Vision

137 Do Saliency Models Detect Odd-One-Out Targets? New Datasets and Evaluations

Iuliia Kotseruba (York University), Calden Wloka (York University), Amir Rasouli (York University), John Tsotsos (York University)

Recent advances in the field of saliency have concentrated on fixation prediction, with benchmarks reaching saturation. However, there is an extensive body of works in psychology and neuroscience that describe aspects of human visual attention that might not be adequately captured by current approaches. Here, we investigate singleton detection, which can be thought of as a canonical example of salience. We introduce two novel datasets, one with psychophysical patterns and one with natural odd-one-out stimuli. Using these datasets we demonstrate through extensive experimentation that nearly all saliency algorithms do not adequately respond to singleton targets in synthetic and natural images. Furthermore, we investigate the effect of training state-of-the-art CNN-based saliency models on these types of stimuli and conclude that the additional training data does not lead to a significant improvement of their ability to find

odd-one-out targets.

159 Attentional demand estimation with attentive driving models

Petar Palasek (MindVisionLabs), Nilli Lavie (University College London, MindVisionLabs), Luke Palmer (MindVisionLabs)

The task of driving can sometimes require the processing of large amounts of visual information; such situations can overload the perceptual systems of human drivers leading to 'inattention blindness', where potentially critical visual information is overlooked. This phenomenon of 'looking but failing to see' is the third largest contributor to traffic accidents in the UK. In this work we develop a method to identify these particularly demanding driving scenes using an end-to-end driving architecture, imbued with a spatial attention mechanism and trained to mimic ground-truth driving controls from video input. At test time, the network's attention distribution is segmented to identify relevant items in the driving scene which are used to estimate the attentional demand on the driver according to an established model in cognitive neuroscience. Without collecting any ground-truth attentional demand data - instead using readily available odometry data in a novel way - our approach is shown to outperform several baselines on a new dataset of 1200 driving scenes labelled for attentional demand in driving.

181 Edge Detection for Event Cameras using Intra-pixel-area Events

Sangil Lee (Seoul National University), Haram Kim (Seoul National University), Hyoun Jin Kim (Seoul National University)

In this work, we propose an edge detection algorithm by estimating a lifetime of an event produced from dynamic vision sensor (DVS), also known as event camera. The event camera, unlike traditional CMOS camera, generates sparse event data at a pixel whose log-intensity changes. Due to this characteristic, theoretically, there is only one or no event at the specific time, which makes it difficult to grasp the world captured by the camera at a particular moment. In this work, we present an algorithm that keeps the event alive until the corresponding event is generated in a nearby pixel so that the shape of an edge is preserved. Particularly, we consider a pixel area to fit a plane on Surface of Active Events (SAE) and call the point inside the pixel area closest to the plane as a intra-pixel-area event. These intra-pixel-area events help the fitting plane algorithm to estimate lifetime robustly and precisely. Our algorithm performs better in terms of sharpness and similarity metric than the accumulation of events over fixed counts or time intervals, when compared with the existing edge detection algorithms, both qualitatively and quantitatively.

182 Simple vs complex temporal recurrences for video saliency prediction

Panagiotis Linardos (Insight Center for Data Analytics), Eva Moledano (Insight Center for Data Analytics), Juan Jose Nieto (Insight Center for Data Analytics), Noel O'Connor (Dublin City University (DCU)), Xavier Giro-i-Nieto (Universitat Politècnica de Catalunya), Kevin McGuinness (Insight Centre for Data Analytics)

This paper investigates modifying an existing neural network architecture for static saliency prediction using two types of recurrences that integrate information from the temporal domain. The first modification is the addition of a ConvLSTM within the architecture, while the second is a conceptually simple exponential moving average of an internal convolutional state. We use weights pre-trained on the SALICON dataset and fine-tune our model on DHF1K. Our results show that both modifications achieve state-of-the-art results and produce similar saliency maps. Source code is available at <https://git.io/fjPiB>.

Illumination and Reflectance

183 Sensor-Independent Illumination Estimation for DNN Models

Mahmoud Afifi (York University), Michael Brown (York University)

While modern deep neural networks (DNNs) achieve state-of-the-art results for illuminant estimation, it is currently necessary to train a separate DNN for each type of camera sensor. This means when a camera manufacturer uses a new sensor, it is necessary to retrain an existing DNN model with training images captured by the new sensor. This paper addresses this problem by introducing a novel sensor-independent illuminant estimation framework. Our method learns a sensor-independent working space that can be used to canonicalize the RGB values of any arbitrary camera sensor. Our learned space retains the linear property of the original sensor raw-RGB space and allows unseen camera sensors to be used on a single DNN model trained on this working space. We demonstrate the effectiveness of this approach on several different camera sensors and show it provides performance on par with state-of-the-art methods that were trained per sensor.

184 Convolutional Mean: A Simple Convolutional Neural Network for Illuminant Estimation

Han Gong (University of East Anglia)

We present Convolutional Mean (CM) – a simple and fast convolutional neural network for illuminant estimation. Our proposed method only requires a small neural network model (1.1K parameters) and a 48×32 thumbnail input image. Our unoptimized Python implementation takes 1 ms/image, which is arguably $3\text{--}3750\times$ faster than the current leading solutions with similar accuracy. Using two public datasets, we show that our proposed light-weight method offers accuracy comparable to the current leading methods’ (which consist of thousands/millions of parameters) across several measures.

155 Adaptive Lighting for Data-Driven Non-Line-of-Sight 3D 185 Localization and Object Identification

Sreenithy Chandran (Arizona State University), Suren Jayasuriya (Arizona State University)

Non-line-of-sight (NLOS) imaging of objects not visible to either the camera or illumination source is a challenging task with vital applications including surveillance and robotics. Recent NLOS reconstruction advances have been achieved using time-resolved measurements which requires expensive and specialized detectors and laser sources. In contrast, we propose a data-driven approach for NLOS 3D localization and object identification requiring only a conventional camera and projector. To generalize to complex line-of-sight (LOS) scenes with non-planar surfaces and occlusions, we introduce an adaptive lighting algorithm. This algorithm, based on radiosity, identifies and illuminates scene patches in the LOS which most contribute to the NLOS light paths, and can factor in system power constraints. We achieve an average identification of 87.1% object identification for four classes of objects, and average localization of the NLOS object’s centroid with a mean-squared error (MSE) of 1.97 cm in the occluded region for real data taken from a hardware prototype. These results demonstrate the advantage of combining the physics of light transport with active illumination for data-driven NLOS imaging.

Image Processing Techniques

186 Content and Colour Distillation for Learning Image 287 Translations with the Spatial Profile Loss

Saqib Sarfraz (Karlsruhe Institute of Technology), Constantin Seibold (Karlsruhe Institute of Technology), Haroon Khalid (Karlsruhe Institute of Technology), Rainer Stiefelhagen (Karlsruhe Institute of Technology)

Generative adversarial networks has emerged as a defacto standard for image translation problems. To successfully drive such models, one has

to rely on additional networks e.g., discriminators and/or perceptual networks. Training these networks with pixel based losses alone are generally not sufficient to learn the target distribution. In this paper, we propose a novel method of computing the loss directly between the source and target images that enable proper distillation of shape/content and colour/style. We show that this is useful in typical image-to-image translations allowing us to successfully drive the generator without relying on additional networks. We demonstrate this on many difficult image translation problems such as image-to-image domain mapping, single image super-resolution and photo realistic makeup transfer. Our extensive evaluation shows the effectiveness of the proposed formulation and its ability to synthesize realistic images.

187 Quantitative Analysis of Similarity Measures of Distributions

Eric Bazán (PSL Research University - MINES ParisTech), Petr Dokládál (PSL Research University - MINES ParisTech), Eva Dokládalová (Université Paris-Est, LIGM, UMR 8049, ESIEE Paris)

There are many measures of dissimilarity that, depending on the application, do not always have optimal behavior. In this paper, we present a qualitative analysis of the similarity measures most used in the literature and the Earth Mover's Distance (EMD). The EMD is a metric based on the theory of optimal transport with interesting geometrical properties for the comparison of distributions. However, the use of this measure is limited in comparison with other similarity measures. The main reason was, until recently, the computational complexity. We show the superiority of the EMD through three different experiments. First, analyzing the response of the measures in the simplest of cases; one-dimension synthetic distributions. Second, with two image retrieval systems; using colour and texture features. Finally, using a dimensional reduction technique for a visual representation of the textures. We show that today the EMD is a measure that better reflects the similarity between two distributions.

188 Gated Multiple Feedback Network for Image Super-Resolution

Qilei Li (Sichuan University), Zhen Li (Sichuan University), Lu Lu (Sichuan University), Gwanggil Jeon (Incheon National University), Kai Liu (Sichuan University), Xiaomin Yang (Sichuan University)

The rapid development of deep learning (DL) has driven single image super-resolution (SR) into a new era. However, in most existing DL based image SR networks, the information flows are solely feedforward, and the high-level features cannot be fully explored. In this paper, we propose the gated multiple feedback network (GMFN) for accurate image SR, in which the representation of low-level features are efficiently enriched by

rerouting multiple high-level features. We cascade multiple residual dense blocks (RDBs) and recurrently unfolds them across time. The multiple feedback connections between two adjacent time steps in the proposed GMFN exploits multiple high-level features captured under large receptive fields to refine the low-level features lacking enough contextual information. The elaborately designed gated feedback module (GFM) efficiently selects and further enhances useful information from multiple rerouted high-level features, and then refine the low-level features with the enhanced high-level information. Extensive experiments demonstrate the superiority of our proposed GMFN against state-of-the-art SR methods in terms of both quantitative metrics and visual quality. Code is available at <https://github.com/liqilei/GMFN>.

189 Wide Activation for Efficient Image and Video Super-Resolution

Jiahui Yu (University of Illinois at Urbana-Champaign), Yuchen Fan (University of Illinois at Urbana-Champaign), Thomas Huang (University of Illinois at Urbana-Champaign)

In this work we demonstrate that with same parameters and computational budgets, models with wider features before ReLU activation have significantly better performance for image and video super-resolution. The resulted SR residual network has a slim identity mapping pathway with wider ($2\times$ to $4\times$) channels before activation in each residual block. To further widen activation ($6\times$ to $9\times$) without computational overhead, we introduce linear low-rank convolution into SR networks and achieve even better accuracy-efficiency tradeoffs. In addition, compared with batch normalization or no normalization, we find training with weight normalization leads to better accuracy for deep super-resolution networks. Our proposed SR network WDSR achieves better results on large-scale DIV2K image super-resolution benchmark in terms of PSNR, under same or lower computational complexity. Based on WDSR, our method won **1st places** in *NTIRE 2018 Challenge on Single Image Super-Resolution* in all three realistic tracks. Moreover, a simple frame-concatenation based WDSR achieved **2nd places** in three out of four tracks of *NTIRE 2019 Challenge for Video Super-Resolution and Deblurring*. Our experiments and ablation studies support the importance of wide activation. Code and models will be publicly available.

190 Higher order Dictionary Learning for Compressed Sensing based Dynamic MRI reconstruction

Minha Mubarak (Indian Institute of Space Science and Technology), Thomas James Thomas (Indian Institute of Space Science and Technology), Sheeba Rani J (Indian Institute of Space Science and Technology), Deepak Mishra (Indian Institute of Space Science and Technology)

Compressed sensing (CS) is one of the predominant tools used presently

to explore the possibilities in accelerating cardiac and body Magnetic Resonance (MR) imaging for achieving shorter scans and accommodating a wider patient group. CS accomplishes this by manoeuvring the scanning hardware to make much fewer measurements and imposing sparsity of the MR image in a known basis on the reconstruction process. Prior works have adopted fixed transforms as well as dictionaries learned on image patches to incorporate the sparsifying basis. Despite the obvious merits of 1-D dictionary learning methods, they suffer from high computational and memory complexity when extended to higher patch sizes and are thus restricted to capturing only local features. Thus, there arises the need for an efficient framework to extend the advantages of dictionary learning to higher dimensional applications like dynamic MRI (dMRI) where there exists a strong correlation across successive frames. This work employs a tensor decomposition based dictionary learning approach to effectively extend CS to dMRI and exploit the temporal gradient (TG) sparsity, thereby retaining maximum spectral and temporal resolution at higher under-sampling factors. The proposed technique is experimentally validated to achieve significant improvement in reconstruction quality over the current state-of-art in dynamic MRI under distinct sampling trajectories and noisy conditions. Further, it facilitates faster reconstructions of the dMRI volume than existing methods, rendering it an ideal choice in critical scenarios which demand a swift diagnosis.

191 Robust Joint Image Reconstruction from Color and Monochrome Cameras

Muxingzi Li (Inria), Peihan Tu (University of Tokyo), Wolfgang Heidrich (KAUST)

Recent years have seen an explosion of the number of camera modules integrated into individual consumer mobile devices, including configurations that contain multiple different types of image sensors. One popular configuration is to combine an RGB camera for color imaging with a monochrome camera that has improved performance in low-light settings, as well as some sensitivity in the infrared. In this work we introduce a method to combine simultaneously captured images from such a two-camera stereo system to generate a high-quality, noise reduced color image. To do so, pixel-to-pixel alignment has to be constructed between the two captured monochrome and color images, which however, is prone to artifacts due to parallax. The joint image reconstruction is made robust by introducing a novel artifact-robust optimization formulation. We provide extensive experimental results based on the two-camera configuration of a commercially available cell phone.

192 Progressive Face Super-Resolution via Attention to Facial Landmark

Deokyun Kim (Korea Advanced Institute of Science and Technology), Minseon Kim (Korea Advanced Institute of Science and Technology), Gihyun Kwon (Korea Advanced Institute of Science and Technology), Daeshik Kim (Korea Advanced Institute of Science and Technology)

Face Super-Resolution (SR) is a subfield of the SR domain that specifically targets the reconstruction of face images. The main challenge of face SR is to restore essential facial features without distortion. We propose a novel face SR method that generates photo-realistic $8\times$ super-resolved face images with fully retained facial details. To that end, we adopt a progressive training method, which allows stable training by splitting the network into successive steps, each producing output with a progressively higher resolution. We also propose a novel facial attention loss and apply it at each step to focus on restoring facial attributes in greater details by multiplying the pixel difference and heatmap values. Lastly, we propose a compressed version of the state-of-the-art face alignment network (FAN) for landmark heatmap extraction. With the proposed FAN, we can extract the heatmaps suitable for face SR and also reduce the overall training time. Experimental results verify that our method outperforms state-of-the-art methods in both qualitative and quantitative measurements, especially in perceptual quality.

193 An Unsupervised Subspace Ranking Method for Continuous Emotions in Face Images

Pooyan Balouchian (University of Central Florida), Marjaneh Safaei (University of Central Florida), Xiaochun Cao (Chinese Academy of Sciences), Hassan Foroosh (University of Central Florida)

Continuous dimensional models of human affect have shown to offer a higher accuracy in identifying a broad range of emotions compared to the discrete categorical approaches dealing only with emotion categories such as joy, sadness, anger, etc. Unlike the majority of existing works on dimensional models of human affect (VAD; i.e. Valence-Arousal-Dominance) that rely on training-based approaches, here we propose an unsupervised and novel approach for ranking of continuous emotions in images using canonical polyadic decomposition. To better portray the efficacy of our proposed approach, we provide theoretical and empirical proof that our system is capable of generating a Pearson Correlation Coefficient that outperforms the state of the art by a large margin; i.e. improved from 0.407 to 0.6721 in one experiment and from 0.35 to 0.7143 in another, when our method was applied to valence rank estimation. Towards this aim, we run experiments on four major emotion recognition datasets; i.e. CK+, AFEW-VA, SEMAINE and AffectNet, and provide analysis on the observed results accordingly. Our datasets are selected in a way to include images

collected under controlled environments such as a laboratory setting; e.g. CK+ and SEMAINE, images collected from semi-controlled environments; e.g. AFEW-VA, and images collected under uncontrolled environments (from the wild); e.g. AffectNet.

194 Deep Learning for Robust end-to-end Tone Mapping

Alexia Briassouli (Maastricht University), Rico Montulet (Maastricht University)

Low-light images require localised processing to enhance details, contrast and lighten dark regions without affecting the appearance of the entire image. A range of tone mapping techniques have been developed to achieve this, with the latest state-of-the-art methods leveraging deep learning. In this work, a new end-to-end tone mapping approach based on Deep Convolutional Adversarial Networks (DCGANs) is introduced along with a data augmentation technique, and shown to improve upon the latest state-of-the-art on benchmarking datasets. We carry out comparisons using the MIT-Adobe FiveK (MIT-5K) and the LOL datasets, as they provide benchmark training and testing data, which is further enriched with data augmentation techniques to increase diversity and robustness. A U-net is used in the generator and a patch-GAN in the discriminator, while a perceptually-relevant loss function based on VGG is used in the generator. The results are visually pleasing, and shown to improve upon the state-of-the-art Deep Retinex, Deep Photo Enhancer and GLADNet on the most widely used benchmark dataset MIT-5K and LOL, without additional computational requirements.

195 Base-detail image inpainting

Ruonan Zhang (Peng Cheng Laboratory), Yurui Ren (Shenzhen Graduate School, Peking University), Ge Li (SECE, Shenzhen Graduate School, Peking University), Jingfei Qiu (Peng Cheng Laboratory)

Recent advances in image inpainting have shown exciting promise with learning-based methods. Though they are effective in capturing features with some prior techniques, most of them fail to reconstruct reasonable base and detail information, so that the inpainted regions appear blurry, over-smoothed, and weird. Therefore, we propose a new “Divider and Conquer” model called Base-Detail Image Inpainting, which combines the reconstructed base and detail layers to generate the final subjective perception images. The base layer with low-frequency information can grasp the basic distribution while the detail layer with high-frequency information assists with the details. The joint generator overall would benefit from these two as guided anchors. In addition, we evaluate our two models over three publicly available datasets, and our experiments demonstrate that our method outperforms current state-of-the-art techniques quantitatively and qualitatively.

160 Generalised Visual Microphone

196

Juhyun Ahn (SUALAB)

Visual microphone (VM) recovers the sound from a silent video, which extracts subtle motion signals from the video using quadrature filter pairs of the complex steerable pyramid (CSP), and then recovers the sound signal by the weighted sum of subtle motion signals. We observe that (1) the subtle motion extraction and the weighted sum in the VM can be treated as a convolution operation, (2) the selection of the sound with the least expected noise can be seen as a pooling of the sound with maximum expected ratio of the sound and noise signals, (3) the VM needs to utilise the past signal to obtain the zero DB level sound signal, and (4) the VM cannot recover the sound sufficiently in the case of using normal frame rate camera. These observations motivate the generalised VM that has the following features: (1) it has the sound recovery convolutional neural networks (SR-CNN) that can learn the ideal filter weights from the training data, (2) it has the DC blocker recurrent neural network (DB-RNN) that recovers the signal with zero DC level, and (3) it has the bandwidth extension residual network (BE-ResNet) to extend the bandwidth of the recovered sound by double. Experiment results show that the proposed generalised VM achieves 12.52% higher segmented SNR than the conventional VM in the case of using the normal frame rate camera.

197 Guide Your Eyes: Learning Image Manipulation under Saliency Guidance

Yen-Chung Chen (National Chiao Tung University), Keng-Jui Chang (National Chiao Tung University), Yi-Hsuan Tsai (NEC Labs America), Yu-Chiang Frank Wang (National Taiwan University), Wei-Chen Chiu (National Chiao Tung University)

In this paper, we tackle the problem of saliency-guided image manipulation for adjusting the saliency distribution over image regions. Conventional approaches ordinarily utilize explicit operations on altering the low-level features based on the selected saliency computation. However, it is difficult to generalize such methods for various saliency estimations. To address this issue, we propose a deep learning-based model that bridges between any differentiable saliency estimation methods and a neural network which applies image manipulation. Thus, the manipulation can be directly optimized in order to satisfy saliency-guidance. Extensive experiment results verify the capacity of our model in saliency-driven image editing and show favorable performance against numerous baselines.

147 **Residual Multiscale Based Single Image Deraining**

198

Yupei Zheng (Beijing Jiaotong University), Xin Yu (Australian National University), Miaomiao Liu (Australian National University), Shunli Zhang (Beijing Jiaotong University)

Rain streaks deteriorate the performance of many computer vision algorithms. Previous methods represent rain streaks by different rain layers and then separate those layers from the background image. However, it is rather difficult to decouple a rain image into rain and background layers due to the complexity of real-world rain, such as various shapes, directions, and densities of rain streaks. In this paper, we propose a residual multiscale pyramid based single image deraining method to alleviate the difficulty of rain image decomposition. In particular, we remove rain streaks in a coarse-to-fine manner. In this fashion, the heavy rain can be significantly removed in the coarse-resolution level of the pyramid first, and the light rain will then be further removed in the high-resolution level. This allows us to avoid distinguishing the densities of rain streaks explicitly since the inaccurate classification of rain densities may lead to over- or insufficient-removal of rain. Furthermore, the residual between a recovered image and its corresponding rain image can provide vital clues of rain streaks. We therefore exploit such residual as an attention map for deraining in its consecutive finer-level. Benefiting from the residual attention maps, rain layers can be better extracted from a higher-resolution input image. Extensive experimental results on synthetic and real datasets demonstrate that our method outperforms the state of the art significantly.

149 **Learnable Gated Temporal Shift Module for Free-form**

199 **Video Inpainting**

Ya-Liang Chang (National Taiwan University), Zhe Yu Liu (National Taiwan University), Kuan-Ying Lee (National Taiwan University), Winston Hsu (National Taiwan University)

How to efficiently utilize temporal information to recover videos in a consistent way is the main issue for video inpainting problems. Conventional 2D CNNs have achieved good performance on image inpainting but often lead to temporally inconsistent results where frames will flicker when applied to videos; 3D CNNs can capture temporal information but are computationally intensive and hard to train. In this paper, we present a novel component termed Learnable Gated Temporal Shift Module (LGTSM) for video inpainting models that could effectively tackle arbitrary video masks without additional parameters from 3D convolutions. LGTSM is designed to let 2D convolutions make use of neighboring frames more efficiently, which is crucial for video inpainting. Specifically, in each layer, LGTSM learns to shift some channels to its temporal neighbors so that 2D convolutions could be enhanced to handle temporal information. Meanwhile, a gated convolution is applied to the layer identify the masked

areas that are poisoning for conventional convolutions. On the FaceForensics and Free-form Video Inpainting (FVI) dataset, our model achieves state-of-the-art results with simply 33% of parameters and inference time.

150 MS-GAN: Text to Image Synthesis with Attention-Modulated Generators and Similarity-aware Discriminators

Fengling Mao (Chinese Academy of Sciences), Bingpeng Ma (Chinese Academy of Sciences), Hong Chang (Chinese Academy of Sciences), Shiguang Shan (Chinese Academy of Sciences), Xilin Chen (Chinese Academy of Sciences)

Existing approaches for text-to-image synthesis often produce images that either contain artifacts or do not well match the text, when the input text description is complex. In this paper, we propose a novel model named MS-GAN, composed of multi-stage attention-Modulated generators and Similarity-aware discriminators, to address these problems. Our proposed generator consists of multiple convolutional blocks that are modulated by both globally and locally attended features calculated between the output image and the text. With such an attention-modulation, our generator can better preserve the semantic information of the text during the text-to-image transformation. Moreover, we propose a similarity-aware discriminator to explicitly constrain the semantic consistency between the text and the synthesized image. Experimental results on Caltech-UCSD Birds and MS-COCO datasets demonstrate that our model can generate images that look more realistic and better match the given text description, compared to the state-of-the-art models.

201 Element-Embedded Style Transfer Networks for Style Harmonization

Hwai-Jin Peng (National Taiwan University), Chia-Ming WANG (National Taiwan University), Yu-Chiang Frank Wang (National Taiwan University)

Neural image style transfer has been receiving increasing attention on the creation of artistic images. Given a reference image with style of interest, image style harmonization aims to blend an element from one image into this reference, achieving harmonization for the stylized output. We present an Element-Embedded Style Transfer Network (E2STN) for addressing this task. Our proposed network uniquely integrates style transfer and image matting modules. Together with global and local discriminators, both context and style information can be properly preserved in the embedded output. In the experiments, we show that our proposed network performs favorably against existing style transfer models and is able to produce results with satisfactory quality.

202 Harmonic Networks for Image Classification

Matej Ulicny (Trinity College Dublin), Vladimir Krylov (Trinity College Dublin), Rozenn Dahyot (Trinity College Dublin)

Convolutional neural networks (CNNs) learn filters in order to capture local correlation patterns in feature space. In contrast, in this paper we propose harmonic blocks that produce features by learning optimal combinations of responses to preset spectral filters. We rely on the use of the Discrete Cosine Transform filters which have excellent energy compaction properties and are widely used for image compression. The proposed harmonic blocks are intended to replace conventional convolutional layers to produce partially or fully harmonic versions of new or existing CNN architectures. We demonstrate how the harmonic networks can be efficiently compressed by exploiting redundancy in spectral domain and truncating high-frequency information. We extensively validate our approach and show that the introduction of harmonic blocks into state-of-the-art CNN models results in improved classification performance on CIFAR and ImageNet datasets.

203 **Semi-supervised Feature-Level Attribute Manipulation for Fashion Image Retrieval**

Minchul Shin (Search Solutions Inc.), Sanghyuk Park (NAVER Clova Vision), Taeksoo Kim (Naver Corporation)

With a growing demand for the search by image, many works have studied the task of fashion instance-level image retrieval (FIR). Furthermore, the recent works introduce a concept of fashion attribute manipulation (FAM) which manipulates a specific attribute (e.g color) of a fashion item while maintaining the rest of the attributes (e.g shape, and pattern). In this way, users can search not only “the same” items but also “similar” items with the desired attributes. FAM is a challenging task in that the attributes are hard to define, and the unique characteristics of a query are hard to be preserved. Although both FIR and FAM are important in real-life applications, most of the previous studies have focused on only one of these problem. In this study, we aim to achieve competitive performance on both FIR and FAM. To do so, we propose a novel method that converts a query into a representation with the desired attributes. We introduce a new idea of attribute manipulation at the feature level, by matching the distribution of manipulated features with real features. In this fashion, the attribute manipulation can be done independently from learning a representation from the image. By introducing the feature-level attribute manipulation, the previous methods for FIR can perform attribute manipulation without sacrificing their retrieval performance.

157 **Fast and Multilevel Semantic-Preserving Discrete Hashing**

Wanqian Zhang (Chinese Academy of Sciences), Dayan Wu (Chinese Academy of Sciences), Jing Liu (Chinese Academy of Sciences), Bo Li (Chinese Academy of Sciences), Xiaoyan Gu (Chinese Academy of Sciences), Weiping Wang (Chinese Academy of Sciences), Dan Meng (Chinese Academy of Sciences)

Deep hashing methods have achieved great success in multi-label image retrieval due to its computation and storage efficiency. However, most existing methods adopt a relaxation-and-quantization optimization strategy, which inevitably degrades the performance. Besides, existing discrete hashing methods are very time-consuming because of the bit-wise learning strategy. To tackle these issues, we propose a novel deep asymmetric discrete hashing method, called Fast and Multilevel semantic-preserving Discrete Hashing (FMDH). FMDH makes the best of supervised information to preserve the multilevel semantic similarities between multi-label images, and further accelerates the training process. Extensive experiments on two widely used multi-label image datasets demonstrate that FMDH can achieve the state-of-the-art performance on both accuracy and training time efficiency.

205 Blind Image Deconvolution using Pretrained Generative Priors

Muhammad Asim (Information Technology University, Lahore), Fahad Shamshad (Information Technology University, Lahore), Ali Ahmed (Information Technology University, Lahore)

This paper proposes a novel approach to regularize the ill-posed blind image deconvolution (blind image deblurring) problem using deep generative networks. We employ two separate deep generative models — one trained to produce sharp images while the other trained to generate blur kernels from lower-dimensional parameters. To deblur, we propose an alternating gradient descent scheme operating in the latent lower-dimensional space of each of the pretrained generative models. Our experiments show excellent deblurring results even under large blurs and heavy noise. To improve the performance on rich image datasets not well learned by the generative networks, we present a modification of the proposed scheme that governs the deblurring process under both generative and classical priors.

206 Single Image Super-Resolution via CNN Architectures and TV-TV Minimization

Marija Vella (Heriot-Watt University), Joao F.C. Mota (Heriot-Watt University)

Super-resolution (SR) is a technique that allows increasing the resolution of a given image. Having applications in many areas, from medical imaging to consumer electronics, several SR methods have been proposed. Currently, the best performing methods are based on convolutional neural networks (CNNs) and require extensive datasets for training. However, at test time, they fail to impose consistency between the super-resolved image

and the given low-resolution image, a property that classic reconstruction-based algorithms naturally enforce in spite of having poorer performance. Motivated by this observation, we propose a new framework that joins both approaches and produces images with superior quality than any of the prior methods. Although our framework requires additional computation, our experiments on Set5, Set14, and BSD100 show that it systematically produces images with better peak signal to noise ratio (PSNR) and structural similarity (SSIM) than the current state-of-the-art CNN architectures for SR.

207 **PtychoNet: Fast and High Quality Phase Retrieval for Ptychography**

Ziqiao Guan (Stony Brook University), Esther Tsai (Brookhaven National Laboratory), Xiaojing Huang (Brookhaven National Laboratory), Kevin Yager (Brookhaven National Laboratory), Hong Qin (Stony Brook University)

Ptychography is a coherent diffractive imaging (CDI) method that captures multiple diffraction patterns of a sample with a set of shifted localized illuminations (“probes”). The reconstruction problem, known as “phase retrieval”, is conventionally solved by iterative algorithms. In this paper, we propose PtychoNet, a deep learning based method to perform phase retrieval for ptychography in a non-iterative manner. We devise a generative network to encode a full ptychography scan, reverse the diffractions at each scanning point and compute the amplitude and phase of the object. We demonstrate successful reconstruction using PtychoNet as well as recovering fine features in the case of extreme sparse scanning where conventional methods fail to give recognizable features.

Objects and Textures

208 **Texel-Att: Representing and Classifying Element-Based** 286 **Textures by Attributes**

Marco Godi (University of Verona), Christian Joppi (University of Verona), Andrea Giachetti (University of Verona), Fabio Pellacini (Sapienza University of Rome), Marco Cristani (University of Verona)

Element-based texture is a kind of texture formed by nameable elements, the texels, distributed according to specific statistical distributions; it is of primary importance in many sectors, namely textile, fashion and interior design industry. State-of-the art texture descriptors fail to properly characterize element-based texture, so we present Texel-Att to fill this gap. Texel-Att is the first fine-grained, attribute-based representation and

classification framework for element-based textures. It first individuates texels, characterizing them with individual attributes; subsequently, texels are grouped and characterized through layout attributes, which give the Texel-Att representation. Texels are detected by a Mask-RCNN, trained on a brand-new element-based texture dataset, ElBa, containing 30K texture images with 3M fully-annotated texels. Examples of individual and layout attributes are exhibited to give a glimpse on the level of achievable graininess. In the experiments, we present detection results to show that texels can be precisely individuated, even on textures “in the wild”; to this sake, we individuate the elementbased classes of the Describable Texture Dataset (DTD), where almost 900K texels have been manually annotated, leading to the Element-based DTD (E-DTD). Subsequently, classification and ranking results demonstrate the expressivity of Texel-Att on ElBa and E-DTD, overcoming the alternative features and relative attributes, doubling the best performance in some cases; finally, we report interactive search results on ElBa and E-DTD: with Texel-Att on the E-DTD dataset we are able to individuate within 10 iterations the desired texture in the 90% of cases, against the 71% obtained with a combination of the finest existing attributes so far. Dataset and code is available at <https://github.com/godimarcovr/Texel-Att>.

209 Style-Guided Zero-Shot Sketch-based Image Retrieval

Titir Dutta (Indian Institute of Science, Bangalore), Soma Biswas (Indian Institute of Science, Bangalore)

Given a sketch query from a previously unseen category, the goal of zero-shot sketch-based image retrieval (ZS-SBIR) is to retrieve semantically meaningful images from a given database. The knowledge-gap between the seen and unseen categories along with sketch-image domain shift makes this an extremely challenging problem. In this work, we propose a novel framework which decomposes each image and sketch into its domain-independent content and a domain, as well as data-dependent variation/style component. Specifically, given a query sketch and a search set of images, we utilize the image specific styles to guide the generation of fake images using the query content to be used for retrieval. Extensive experiments on two large-scale sketch-image datasets, Sketchy extended and TU-Berlin show that the proposed approach performs better or comparable to the state-of-the-art in both ZS-SBIR and generalized ZS-SBIR protocols.

210 Robust Synthesis of Adversarial Visual Examples Using 283 a Deep Image Prior

Thomas Gittings (University of Surrey), Steve Schneider (University of Surrey), John Collosse (University of Surrey)

We present a novel method for generating robust adversarial image examples building upon the recent ‘deep image prior’ (DIP) that exploits convolutional network architectures to enforce plausible texture in image synthesis. Adversarial images are commonly generated by perturbing images to introduce high frequency noise that induces image misclassification, but that is fragile to subsequent digital manipulation of the image. We show that using DIP to reconstruct an image under adversarial constraint induces perturbations that are more robust to affine deformation, whilst remaining visually imperceptible. Furthermore we show that our DIP approach can also be adapted to produce local adversarial patches (‘adversarial stickers’). We demonstrate robust adversarial examples over a broad gamut of images and object classes drawn from the ImageNet dataset.

211 **Orthographic Feature Transform for Monocular 3D Ob-** 285 **ject Detection**

Thomas Roddick (University of Cambridge), Alex Kendall (University of Cambridge), Roberto Cipolla (University of Cambridge)

3D object detection from monocular images has proven to be an enormously challenging task, with the performance of leading systems not yet achieving even 10% of that of LiDAR-based counterparts. One explanation for this performance gap is that existing systems are entirely at the mercy of the perspective image-based representation, in which the appearance and scale of objects varies drastically with depth and meaningful distances are difficult to infer. In this work we argue that the ability to reason about the world in 3D is an essential element of the 3D object detection task. To this end, we introduce the orthographic feature transform, which maps image-based features into an orthographic 3D space, enabling us to reason holistically about the spatial configuration of the scene. We apply this transformation as part of an end-to-end deep learning architecture and demonstrate our approach on the KITTI 3D object benchmark.

212 **An Empirical Study on Leveraging Scene Graphs for Vi-** 288 **sual Question Answering**

Cheng Zhang (Ohio State University), Wei-Lun Chao (Cornell University), Dong Xuan (Ohio State University)

Visual question answering (Visual QA) has attracted significant attention these years. While a variety of algorithms have been proposed, most of them are built upon different combinations of image and language features as well as multi-modal attention and fusion. In this paper, we investigate an alternative approach inspired by conventional QA systems that operate on knowledge graphs. Specifically, we investigate the use of scene graphs derived from images for Visual QA: an image is abstractly represented by

a graph with nodes corresponding to object entities and edges to object relationships. We adapt the recently proposed graph network (GN) to encode the scene graph and perform structured reasoning according to the input question. Our empirical studies demonstrate that scene graphs can already capture essential information of images and graph networks have the potential to outperform state-of-the-art Visual QA algorithms but with a much cleaner architecture. By analyzing the features generated by GNs we can further interpret the reasoning process, suggesting a promising direction towards explainable Visual QA.

213 Generalized Zero-shot Learning using Open Set Recognition

Omkar Gune (Indian Institute of Technology Bombay), Amit More (Indian Institute of Technology Bombay), Biplab Banerjee (Indian Institute of Technology Bombay), Subhasis Chaudhuri (Indian Institute of Technology Bombay)

Generalized Zero-shot Learning (GZSL) aims at identifying the test samples which can belong to previously *seen* (training) or *unseen* visual categories by leveraging the side information present in the form of class semantics. In general, GZSL is a difficult problem in comparison to the standard Zero-shot Learning (ZSL) given the model bias towards the seen classes. In this paper, we follow an intuitive approach to solve the GZSL problem by adhering ideas from the Open Set Recognition (OSR) literature. To this end, the proposed model acts as a pre-processing module for the GZSL inference stage which decides whether a given test sample belongs to seen or unseen class (domain). In order to comprehend the same, we generate *pseudo unseen visual* samples from the available seen data and further train a domain classifier for on-the-fly domain label assignment for the test samples. The domain specific inference modules are then applied subsequently for improved classification. We experiment on standard benchmark AWA1, APY, FLO, and CUB datasets which confirm superior performance over the existing state of the art.

214 High Frequency Residual Learning for Multi-Scale Image Classification

Bowen Cheng (UIUC), Rong Xiao (Ping An), Jianfeng Wang (Microsoft Research), Thomas Huang (UIUC), Lei Zhang (Microsoft)

We present a novel high frequency residual learning framework, which leads to a highly efficient multi-scale network (MSNet) architecture for mobile and embedded vision problems. The architecture utilizes two networks: a low resolution network to efficiently approximate low frequency components and a high resolution network to learn high frequency residuals by reusing the upsampled low resolution features. With a classifier calibration module, MSNet can dynamically allocate computation resources

during inference to achieve a better speed and accuracy trade-off. We evaluate our methods on the challenging ImageNet-1k dataset and observe consistent improvements over different base networks. On ResNet-18 and MobileNet with $\alpha=1.0$, MSNet gains 1.5% over both architectures without increasing computations. On the more efficient MobileNet with $\alpha=0.25$, our method gains 3.8% with the same amount of computations.

215 Learning Efficient Detector with Semi-supervised Adaptive Distillation

Shitao Tang (SenseTime Research), Litong Feng (SenseTime Research), Wenqi Shao (The Chinese University of HongKong), Zhanghui Kuang (SenseTime Ltd.), Wayne Zhang (SenseTime Research), Zheng Lu (University of Nottingham, Ningbo China)

Convolutional Neural Networks based object detection techniques produce accurate results but often time consuming. Knowledge distillation has been popular for model compression to speed up. In this paper, we propose a Semi-supervised Adaptive Distillation (SAD) framework to accelerate single-stage detectors while still improving the overall accuracy. We introduce our Adaptive Distillation Loss (ADL) that enables student model to mimic teacher's logits adaptively with more attention paid on two types of hard samples, hard-to-learn samples predicted by teacher model with low certainty and hard-to-mimic samples with a large gap between the teacher's and the student's prediction. We then show that student model can be improved further in the semi-supervised setting with the help of ADL. Our experiments validate that for distillation on unlabeled data, ADL achieves better performance than existing data distillation using both soft and hard targets. On the COCO database, SAD makes a student detector with a backbone of ResNet-50 out-perform its teacher with a backbone of ResNet-101, while the student has half of the teacher's computation complexity.

216 Knowledge Distillation for End-to-End Person Search

Bharti Munjal (OSRAM), Fabio Galasso (OSRAM), Sikandar Amin (OSRAM)

We introduce knowledge distillation for end-to-end person search. End-to-End methods are the current state-of-the-art for person search that solve both detection and re-identification jointly. These approaches for joint optimization show their largest drop in performance due to a sub-optimal detector. We propose two distinct approaches for extra supervision of end-to-end person search methods in a teacher-student setting. The first is adopted from state-of-the-art knowledge distillation in object detection. We employ this to supervise the detector of our person search model at various levels using a specialized detector. The second approach is new, simple and yet considerably more effective. This distills knowledge from a teacher re-identification technique via a pre-computed look-up table of

ID features. It relaxes the learning of identification features and allows the student to focus on the detection task. This procedure not only helps fixing the sub-optimal detector training in the joint optimization and simultaneously improving the person search, but also closes the performance gap between the teacher and the student for model compression in this case. Over-all, we demonstrate significant improvements for two recent state-of-the-art methods using our proposed knowledge distillation approach on two benchmark datasets. Moreover, on the model compression task our approach brings the performance of smaller models on par with the larger models.

217 One-Shot Scene-Specific Crowd Counting

Mohammad Hossain (HUAWEI Technologies Co, LTD.), Mahesh Kumar K (University of Manitoba), Mehrdad Hosseinzadeh (University of Manitoba), Omit Chanda (University of Manitoba), Yang Wang (University of Manitoba)

We consider the problem of crowd counting in static images. Given an image, the goal is to estimate a density map of this image, where each value in the density map indicates the density level of the corresponding location in the image. In particular, we consider a novel problem setting which we call the one-shot scene-specific crowd counting. During training, we assume that we have labeled images collected from different scenes. Each scene corresponds to a camera at a fixed location and angle. Given a target scene, we assume that we have one single labeled image collected from that scene. Our goal is to adapt the crowd counting model to this specific scene based on this single example. We argue that this setting is more realistic in terms of deploying crowd counting algorithms in real-world applications. We propose a novel one-shot learning approach for learning how to adapt to a target scene using one labeled example. Our experiment results demonstrate that our proposed approach outperforms other alternative methods.

218 Exploring the Vulnerability of Single Shot Module in Object Detectors via Imperceptible Background Patches

Yuezun Li (University at Albany), Xiao Bian (GE Global Research), Ming-Ching Chang (University at Albany), Siwei Lyu (University at Albany)

Recent works succeeded to generate adversarial perturbations on the entire image or the object of interests to corrupt CNN based object detectors. In this paper, we focus on exploring the vulnerability of the Single Shot Module (SSM) commonly used in recent object detectors, by adding small perturbations to patches in the background outside the object. The SSM is referred to the Region Proposal Network used in a two-stage object detector or the single-stage object detector itself. The SSM is typically a fully convolutional neural network which generates output in a single

forward pass. Due to the excessive convolutions used in SSM, the actual receptive field is larger than the object itself. As such, we propose a novel method to corrupt object detectors by generating imperceptible patches only in the background. Our method can find a few background patches for perturbation, which can effectively decrease true positives and dramatically increase false positives. Efficacy is demonstrated on 5 two-stage object detectors and 8 single-stage object detectors on the MS COCO 2014 dataset. Results indicate that perturbations with small distortions outside the bounding box of object region can still severely damage the detection performance.

219 **Balancing Specialization, Generalization, and Compression for Detection and Tracking**

Dotan Kaufman (Amazon), Koby Bibas (Amazon), Eran Borenstein (Amazon), Michael Chertok (Amazon, Lab126), Tal Hassner (Amazon)

We propose a method for specializing deep detectors and trackers to restricted settings. Our approach is designed with the following goals in mind: (a) Improving accuracy in restricted domains; (b) preventing overfitting to new domains and forgetting of generalized capabilities; (c) aggressive model compression and acceleration. To this end, we propose a novel loss that balances compression and acceleration of a deep learning model vs. loss of generalization capabilities. We apply our method to the existing tracker and detector models. We report detection results on the VIRAT and CAVIAR data sets. These results show our method to offer unprecedented compression rates along with improved detection. We apply our loss for tracker compression at test time, as it processes each video. Our tests on the OTB2015 benchmark show that applying compression during test time actually improves tracking performance.

220 **Efficient Coarse-to-Fine Non-Local Module for the Detection of Small Objects**

Hila Levi (Weizmann Institute of Science), Shimon Ullman (Weizmann Institute of Science)

An image is not just a collection of objects, but rather a graph where each object is related to other objects through spatial and semantic relations. Using relational reasoning modules, such as the non-local module, can therefore improve object detection. Current schemes apply such dedicated modules either to a specific layer of the bottom-up stream, or between already-detected objects. We show that the relational process can be better modeled in a coarse-to-fine manner and present a novel framework, applying a non-local module sequentially to increasing resolution feature maps along the top-down stream. In this way, information can naturally be passed from larger objects to smaller related ones. Applying the module

to fine feature maps further allows the information to pass between the small objects themselves, exploiting repetitions of instances from of the same class. In practice, due to the expensive memory utilization of the non-local module, it is infeasible to apply the module as currently used to high-resolution feature maps. We redesigned the non local module, improved it in terms of memory and number of operations, allowing it to be placed anywhere along the network. We further incorporated relative spatial information into the module, in a manner that can be incorporated into our efficient implementation. We show the effectiveness of our scheme by improving the results of detecting small objects on COCO by 1-2 AP points over Faster and Mask RCNN and by 1 AP over using non-local module on the bottom-up stream.

148 Open-set Recognition of Unseen Macromolecules in Cellular Electron Cryo-Tomograms by Soft Large Margin Centralized Cosine Loss

Xuefeng Du (Xi'an Jiaotong University), Xiangrui Zeng (Carnegie Mellon University), Bo Zhou (Yale University), Alex Singh (Carnegie Mellon University), Min Xu (Carnegie Mellon University)

Cellular Electron Cryo-Tomography (CECT) is a 3D imaging tool that visualizes the structure and spatial organization of macromolecules at sub-molecular resolution in a near native state, allowing systematic analysis of seen and unseen macromolecules. Methods for high-throughput subtomogram classification on known macromolecules based on deep learning have been developed. However, the learned features guided by either the regular Softmax loss or traditional feature descriptors are not well applicable in the open-set recognition scenarios where the testing data and the training data have a different label space. In other words, the testing data contain novel structural classes unseen in the training data. In this paper, we propose a novel loss function for deep neural networks to extract discriminative features for unseen macromolecular structure recognition in CECT, called Soft Large Margin Centralized Cosine Loss (Soft LMCCL). Our Soft LMCCL projects 3D images into a normalized hypersphere that generates features with a large inter-class variance and a low intra-class variance, which can better generalize across data with different classes and in different datasets. Our experiments on CECT subtomogram recognition tasks using both simulation data and real data demonstrate that we are able to achieve significantly better verification accuracy and reliability compared to classic loss functions. In summary, our Soft LMCCL is a useful design in our detection task of unseen structures and is potentially useful in other similar open-set scenarios.

222 Multi-scale Template Matching with Scalable Diversity Similarity in an Unconstrained Environment

Yi Zhang (Iwate University), Chao Zhang (University of Fukui), Takuya Akashi (Iwate University)

We propose a novel multi-scale template matching method which is robust against both scaling and rotation in unconstrained environments. The key component behind is a similarity measure referred to as scalable diversity similarity (SDS). Specifically, SDS exploits bidirectional diversity of the nearest neighbor (NN) matches between two sets of points. To address the scale-robustness of the similarity measure, local appearance and rank information are jointly used for the NN search. Furthermore, by introducing penalty term on the scale change, and polar radius term into the similarity measure, SDS is shown to be a well-performing similarity measure against overall size and rotation changes, as well as non-rigid geometric deformations, background clutter, and occlusions. The properties of SDS are statistically justified, and experiments on both synthetic and real-world data show that SDS can significantly outperform state-of-the-art methods.

223 Improving Multi-stage Object Detection via Iterative Proposal Refinement

Jicheng Gong (Westwell-lab), Zhao Zhao (Westwell-lab), Nic Li (Westwell-lab)

For object detection tasks, multi-stage detection frameworks have achieved excellent detection performance (e.g., Cascade R-CNN) compared to those one and two-stage frameworks (e.g., FPN). In this work, we introduce an LSTM-based proposal refinement module that iteratively refines proposed bounding boxes. This module can naturally be integrated with different frameworks. And the number of iterative steps is flexible and can differ between training and testing stages. In this work, we focus on improving the widely used two-stage frameworks by replacing the original bounding box regression head with our proposed module. To verify the efficacy of our method, we perform extensive experiments on PASCAL VOC and MS COCO benchmarks with both ResNet-50 and ResNet-101 backbones. The results show that by having our LSTM based module it achieves significantly higher mAP than the vanilla R-FCN and FPN on both benchmarks. Meanwhile, it outperforms the existing state-of-the-art method Cascade R-CNN especially under high IoU thresholds.

224 Generating Expensive Relationship Features from Cheap Objects

Xiaogang Wang (National University of Singapore), Qianru Sun (Singapore Management University), Marcelo Ang (National University of Singapore), Tat-Seng Chua (National University of Singapore)

We investigate the problem of object relationship classification of visual scenes. For a relationship object1-predicate-object2 that captures the object interaction, its representation is composed by the combination of

object1 and object2 features. As a result, relationship classification models usually bias to the frequent objects, leading to poor generalization to rare or unseen objects. Inspired by the data augmentation methods, we propose a novel Semantic Transform Generative Adversarial Network (ST-GAN) that synthesizes relationship features for rare objects, conditioned on the features from random instances of the objects. Specifically, ST-GAN essentially offers a semantic transform function from cheap object features to expensive relationship features. Here, “cheap” means any easy-to-collect object which possesses an original but undesired relationship attribute, e.g., a sitting person; “expensive” means a target relationship on this object, e.g., person-riding-horse. By generating massive triplet combinations from any object pair with larger variance, ST-GAN can reduce the data bias. Extensive experiments on two benchmarks – Visual Relationship Detection (VRD) and Visual Genome (VG), show that using our synthesized features for data augmentation, the relationship classification model can be consistently improved in various settings such as zero-shot and low-shot.

225 Soft Sampling for Robust Object Detection

Zhe Wu (University of Maryland), Navaneeth Bodla (University of Maryland), Bharat Singh (Amazon), Mahyar Najibi (University of Maryland), Rama Chellappa (University of Maryland), Larry Davis (University of Maryland)

We study the robustness of object detection under the presence of missing annotations. In this setting, the unlabeled object instances will be treated as background, which will generate an incorrect training signal for the detector. Interestingly, we observe that after dropping 30% of the annotations (and labeling them as background), the performance of CNN-based object detectors like Faster-RCNN only drops by 5% on the PASCAL VOC dataset. We provide a detailed explanation for this result. To further bridge the performance gap, we propose a simple yet effective solution, called Soft Sampling. Soft Sampling re-weights the gradients of RoIs as a function of overlap with positive instances. This ensures that the uncertain background regions are given a smaller weight compared to the hard-negatives. Extensive experiments on curated PASCAL VOC datasets demonstrate the effectiveness of the proposed Soft Sampling method at different annotation drop rates. Finally, we show that on OpenImagesV3, which is a real-world dataset with missing annotations, Soft Sampling outperforms standard detection baselines by over 3%. It was also included in the top performing entries in the OpenImagesV4 challenge conducted during ECCV 2018.

226 Joint Learning of Attended Zero-Shot Features and Visual-Semantic Mapping

Yanan Li (Zhejiang Lab), Donghui Wang (Zhejiang University)

Zero-shot learning (ZSL) aims to recognize unseen categories by associating image features with semantic embeddings of class labels and its performance can be improved progressively through learning better features and more generalized visual-semantic mapping (V-S mapping) to unseen classes. Current methods typically learn feature extractors and V-S mapping independently. In this work, we propose a simple but effective joint learning framework with fused autoencoder (AE) paradigm, which can simultaneously learn features specific to ZSL task as well as V-S mapping inseparable to learning features. In particular, the encoder in AE can not only transfer semantic knowledge to the feature space, but also achieve semantics-guided attended feature learning. At the same time, the decoder in AE can be used as a V-S mapping, which further improves the generalization ability to unseen classes. Extensive experiments show that the proposed approach can achieve promising results.

227 Cascade RetinaNet: Maintaining Consistency for Single-Stage Object Detection

Hongkai Zhang (Chinese Academy of Sciences), Hong Chang (Chinese Academy of Sciences), Bingpeng Ma (Chinese Academy of Sciences), Shiguang Shan (Chinese Academy of Sciences), Xilin Chen (Chinese Academy of Sciences)

Recent researches attempt to improve the detection performance by adopting the idea of cascade for single-stage detectors. In this paper, we analyze and discover that inconsistency is the major factor limiting the performance. The refined anchors are associated with the feature extracted from the previous location and the classifier is confused by misaligned classification and localization. Further, we point out two main designing rules for the cascade manner: improving consistency between classification confidence and localization performance, and maintaining feature consistency between different stages. A multistage object detector named Cas-RetinaNet, is then proposed for reducing the misalignments. It consists of sequential stages trained with increasing IoU thresholds for improving the correlation, and a novel Feature Consistency Module for mitigating the feature inconsistency. Experiments show that our proposed Cas-RetinaNet achieves stable performance gains across different models and input scales. Specifically, our method improves RetinaNet from 39.1 AP to 41.1 AP on the challenging MS COCO dataset without any bells or whistles.

228 Deep Learning Fusion of RGB and Depth Images for Pedestrian Detection

Zhixin Guo (Ghent University), Wenzhi Liao (Ghent University), Yifan Xiao (Ghent University), Peter Veelaert (UGent), Wilfried Philips (IPI - Ghent University - imec)

In this paper, we propose an effective method based on the Faster-RCNN structure to combine RGB and depth images for pedestrian detection. Dur-

ing the training stage, we generate a semantic segmentation map from the depth image and use it to refine the convolutional features extracted from the RGB images. In addition, we acquire more accurate region proposals by exploring the perspective projection with the help of depth information. Experimental results demonstrate that our proposed method achieves the state-of-the-art RGBD pedestrian detection performance on KITTI dataset.

229 **BMNet: A Reconstructed Network for Lightweight Object Detection via Branch Merging**

Hefei Ling (Huazhong University of Science and Technology), Li Zhang (Huazhong University of Science and Technology), Yangyang Qin (Huazhong University of Science and Technology), Yuxuan Shi (Huazhong University of Science and Technology), Lei Wu (Huazhong University of Science and Technology), Jiazhong Chen (Huazhong University of Science and Technology), Baiyan Zhang (Huazhong University of Science and Technology)

Object detection has made great progress in recent years along with the rapid development of deep learning. However, most current object detection networks cannot be used in the devices with limited computation power and memory resource, such as electronic chips, mobile phones, etc. To achieve an object detection network for the resource-constrained scenario, this paper proposes a reconstructed Network for lightweight object detection via Branch Merging (BMNet). BMNet introduces an innovative and efficient architecture named 2-way Merging Lightweight Dense Block (2-way MLDB), which merges the duplicate parts of two branches in a dense block of the backbone network to obtain multi-receptive field features with fewer parameters and computations. In addition, to alleviate the decrease of accuracy caused by drastically reduced parameter size, BMNet builds an FPN-like SSD based on an Attention Prediction Block (APB) structure. Through extensive experiments on two classic benchmarks (PASCAL VOC 2007 and MS COCO), we demonstrate that BMNet is superior to the most advanced lightweight object detection solutions such as Tiny SSD, MobileNet-SSD, MobileNetV2-SSD and Pelee in terms of parameter size, FLOPs and accuracy. Concretely, BMNet achieves 73.48% of mAP on PASCAL VOC 2007 dataset with only 1.49 M parameters and 1.51 B FLOPs, which is the latest result with relatively low resource requirements and without pre-training to date.

230 **An Adaptive Supervision Framework for Active Learning in Object Detection**

Sai Vikas Desai (Indian Institute of Technology, Hyderabad), Akshay Chandra Lagandula (Indian Institute Of Technology, Hyderabad), Wei Guo (The University of Tokyo), Seishi Ninomiya (The University of Tokyo), Vineeth N Balasubramanian (Indian Institute of Technology, Hyderabad)

Active learning approaches in computer vision generally involve querying strong labels for data. However, previous works have shown that weak su-

pervision can be effective in training models for vision tasks while greatly reducing annotation costs. Using this knowledge, we propose an adaptive supervision framework for active learning and demonstrate its effectiveness on the task of object detection. Instead of directly querying bounding box annotations (strong labels) for the most informative samples, we first query weak labels and optimize the model. Using a switching condition, the required supervision level can be increased. Our framework requires little to no change in model architecture. Our extensive experiments show that the proposed framework can be used to train good generalizable models with much lesser annotation costs than the state of the art active learning approaches for object detection.

231 Adversarial Signboard against Object Detector

Yi Huang (Nanyang Technological University), Kwok-Yan Lam (Nanyang Technological University), Wai-Kin Adams Kong (Nanyang Technological University)

Object detector is an indispensable component in many computer vision and artificial intelligence systems, such as autonomous robot and image analyzer for profiling social media users. Analyzing its vulnerabilities is essential for detecting and preventing attacks and minimizing potential loss. Researchers have proposed a number of adversarial examples to evaluate the robustness of object detectors. All these adversarial examples change pixels inside target objects to carry out attacks but only some of them are suitable for physical attacks. According to the best knowledge of the authors, no published work successfully attacks object detector without changing pixels inside the target object. In an unpublished work, the authors designed an adversarial border which tightly surrounds target object and successfully misleads Faster R-CNN and YOLOv3 digitally and physically. Adversarial border does not change pixels inside target object but makes it look weird. In this paper, a new adversarial example named adversarial signboard, which looks like a signboard, is proposed. By putting it below a target object, it can mislead the state-of-the-art object detectors. Using stop sign as a target object, adversarial signboard is evaluated on 48 videos with totally 5416 frames. The experimental results show that adversarial signboard derived from Faster R-CNN with ResNet-101 as a backbone network can mislead Faster R-CNN with a different backbone network, Mask R-CNN, YOLOv3 and R-FCN digitally and physically.

232 Domain Adaptation for Object Detection via Style Consistency

Adrian Lopez Rodriguez (Imperial College London), Krystian Mikolajczyk (Imperial College London)

We propose a domain adaptation approach for object detection. We introduce a two-step method: the first step makes the detector robust to

low-level differences and the second step adapts the classifiers to changes in the high-level features. For the first step, we use a style transfer method for pixel-adaptation of source images to the target domain. We find that enforcing low distance in the high-level features of the object detector between the style transferred images and the source images improves the performance in the target domain. For the second step, we propose a robust pseudo labelling approach to reduce the noise in both positive and negative sampling. Experimental evaluation is performed using the detector SSD300 on PASCAL VOC extended with the dataset proposed in, where the target domain images are of different styles. Our approach significantly improves the state-of-the-art performance in this benchmark.

233 HydraPicker: Fully Automated Particle Picking in Cryo-EM by Utilizing Dataset Bias in Single Shot Detection

Abbas Masoumzadeh (York University), Marcus Brubaker (York University)

Particle picking in cryo-EM is a form of object detection for noisy, low contrast, and out-of-focus microscopy images, taken of different (unknown) structures. This paper presents a fully automated approach which, for the first time, explicitly considers training on multiple structures, while simultaneously learning both specialized models for each structure used for training and a generic model that can be applied to unseen structures. The presented architecture is fully convolutional and divided into two parts: (i) a portion which shares its weights across all structures and (ii) $N+1$ parallel sets of sub-architectures, N of which are specialized to the structures used for training and a generic model whose weights are tied to the layers for the specialized models. Experiments reveal improvements in multiple use cases over the-state-of-art and present additional possibilities to practitioners.

234 Rethinking Convolutional Feature Extraction for Small Object Detection

Burhan Mudassar (Georgia Institute of Technology), Saibal Mukhopadhyay (Georgia Institute of Technology)

Deep learning based object detection architectures have significantly advanced the state of the art. However, a study of recent detection methods shows a wide gap between small object performance and performance on medium and large objects. This gap is prevalent across architectures and across backbones. We show that this gap is primarily due to reduction in the feature map size as we traverse the backbone. Through simple modifications to the backbone structure, we show a marked improvement in performance for small objects. In addition, we propose a dual-path configuration with weight sharing for recovering large object performance.

Compared to state of the art methods that rely on multi-scale training and network partitioning we show competitive performance without any bells and whistles on the MS COCO dataset. We show state of the art small object performance with a mobile object detector SSD Mobilenet v1.

17 **Guided Zoom: Questioning Network Evidence for Fine-grained Classification**

Sarah Bargal (Boston University), Andrea Zunino (Istituto Italiano di Tecnologia), Vitali Petsiuk (Boston University), Jianming Zhang (Adobe Research), Kate Saenko (Boston University), Vittorio Murino (Istituto Italiano di Tecnologia), Stan Sclaroff (Boston University)

We propose Guided Zoom, an approach that utilizes spatial grounding of a model’s decision to make more informed predictions. It does so by making sure the model has “the right reasons” for a prediction, defined as reasons that are coherent with those used to make similar correct decisions at training time. The reason/evidence upon which a deep convolutional neural network makes a prediction is defined to be the spatial grounding, in the pixel space, for a specific class conditional probability in the model output. Guided Zoom examines how reasonable such evidence is for each of the top-k predicted classes, rather than solely trusting the top-1 prediction. We show that Guided Zoom improves the classification accuracy of a deep convolutional neural network model and obtains state-of-the-art results on three fine-grained classification benchmark datasets.

236 **Geometry-Aware Video Object Detection for Static Cameras**

Dan Xu (University of Oxford), Weidi Xie (University of Oxford), Andrew Zisserman (University of Oxford)

In this paper we propose a geometry-aware model for video object detection. Specifically, we consider the setting that cameras can be well approximated as static, e.g. in video surveillance scenarios, and scene pseudo depth maps can therefore be inferred easily from the object scale on the image plane. We make the following contributions: First, we extend the recent anchor-free detector (CornerNet) to video object detections. In order to exploit the spatial-temporal information while maintaining high efficiency, the proposed model accepts video clips as input, and only makes predictions for the starting and the ending frames, i.e. heatmaps of object bounding box corners and the corresponding embeddings for grouping. Second, to tackle the challenge from scale variations in object detection, scene geometry information, e.g. a depth map, is explicitly incorporated into deep networks for multi-scale feature selection and for the network prediction. Third, we validate the proposed architectures on an autonomous driving dataset generated from the Carla simulator, and on a real dataset for human detection (DukeMTMC dataset). When compar-

ing with the existing competitive single-stage or two-stage detectors, the proposed geometry-aware spatio-temporal network achieves significantly better results.

138 PCAS: Pruning Channels with Attention Statistics for 237 Deep Network Compression

Kohei Yamamoto (Oki Electric Industry Co., Ltd.), Kurato Maeno (Oki Electric Industry Co., Ltd.)

Compression techniques for deep neural networks are important for implementing them on small embedded devices. In particular, channel-pruning is a useful technique for realizing compact networks. However, many conventional methods require manual setting of compression ratios in each layer. It is difficult to analyze the relationships between all layers, especially for deeper models. To address these issues, we propose a simple channel-pruning technique based on attention statistics that enables to evaluate the importance of channels. We improved the method by means of a criterion for automatic channel selection, using a single compression ratio for the entire model in place of per-layer model analysis. The proposed approach achieved superior performance over conventional methods with respect to accuracy and the computational costs for various models and datasets. We provide analysis results for behavior of the proposed criterion on different datasets to demonstrate its favorable properties for channel pruning.

139 Large Margin In Softmax Cross-Entropy Loss 238

Takumi Kobayashi (National Institute of Advanced Industrial Science and Technology)

Deep convolutional neural networks (CNNs) are trained mostly based on the softmax cross-entropy loss to produce promising performance on various image classification tasks. While much research effort has been made to improve the building blocks of CNNs, the classifier margin in the loss attracts less attention for optimizing CNNs in contrast to the kernel-based methods, such as SVM. In this paper, we propose a novel method to induce a large-margin CNN for improving the classification performance. By analyzing the formulation of the softmax loss, we clarify the margin embedded in the loss as well as its connection to the distribution of softmax logits. Based on this analysis, the proposed method is formulated as regularization imposed on the logits to induce a large-margin classifier in a compatible form with the softmax loss. The experimental results on image classification using various CNNs demonstrate that the proposed method favorably improves performance compared to the other large-margin losses.

143 **Group Based Deep Shared Feature Learning for Fine-grained Image Classification**

Xuelu Li (The Pennsylvania State University), Vishal Monga (Pennsylvania State University)

Fine-grained image classification has emerged as a significant challenge because objects in such images have small inter-class visual differences but with large variations in pose, lighting, and viewpoints, etc. Most existing work focuses on highly customized feature extraction via deep network architectures which have been shown to deliver state of the art performance. Given that images from distinct classes in fine-grained classification share significant features of interest, we present a new deep network architecture that explicitly models shared features and removes their effect to achieve enhanced classification results. Our modeling of shared features is based on a new group based learning wherein existing classes are divided into groups and multiple shared feature patterns are discovered (learned). We call this framework Group based deep Shared Feature Learning (GSFL) and the resulting learned network as GSFL-Net. Specifically, the proposed GSFL-Net develops a specially designed autoencoder which is constrained by a newly proposed Feature Expression Loss to decompose a set of features into their constituent shared and discriminative components. During inference, only the discriminative feature component is used to accomplish the classification task. A key benefit of our specialized autoencoder is that it is versatile and can be combined with state-of-the-art fine-grained feature extraction models and trained together with them to improve their performance directly. Experiments on benchmark datasets show that GSFL-Net can enhance classification accuracy over the state of the art with a more interpretable architecture.

240 **A Top-Down Unified Framework for Instance-level Human Parsing**

Haifang Qin (Peking University), Weixiang Hong (National University of Singapore), Wei-Chih Hung (University of California, Merced), Yi-Hsuan Tsai (NEC Labs America), Ming-Hsuan Yang (University of California, Merced)

Instance-level human parsing is one of the essential tasks for human-centric analysis which aims to segment various body parts and associate each part with the corresponding human instance simultaneously. Most state-of-the-art methods group instances upon multi-human parsing results, but they tend to miss instances and fail in grouping under the crowded scene. To address this problem, we propose a top-down unified framework to simultaneously detect human instance and parse every part within that instance. To better parse the single human, we also design an attention module, which is aggregated to our parsing network. As a result, our approach is capable of obtaining fine-grained parsing results and the corresponding human mask in a single forward pass. Experiments show that the proposed

algorithm performs favorably against state-of-the-art methods on the CIHP and PASCAL-Person-Part datasets.

241 Rethinking Classification and Localization for Cascade R-CNN

Ang Li (Nanjing University of Science and Technology), Xue Yang (Shanghai Jiao Tong University), Chongyang Zhang (Nanjing University of Science and Technology)

We extend the state-of-the-art Cascade R-CNN with a simple feature sharing mechanism. Our approach focuses on the performance increases on high IoU but decreases on low IoU thresholds—a key problem this detector suffers from. Feature sharing is extremely helpful, our results show that given this mechanism embedded into all stages, we can easily narrow the gap between the last stage and preceding stages on low IoU thresholds without resorting to the commonly used testing ensemble but the network itself. We also observe obvious improvements on all IoU thresholds benefited from feature sharing, and the resulting cascade structure can easily match or exceed its counterparts, only with negligible extra parameters introduced. To push the envelope, we demonstrate 43.2 AP on COCO object detection without any bells and whistles including testing ensemble, surpassing previous Cascade R-CNN by a large margin. Our framework is easy to implement and we hope it can serve as a general and strong baseline for future research.

146 Image Captioning with Unseen Objects 242

Berkan Demirel (HAVELSAN Inc. & METU), Ramazan Gokberk Cinbis (METU), Nazli Iklizler-Cinbis (Hacettepe University)

Image caption generation is a long standing and challenging problem at the intersection of computer vision and natural language processing. A number of recently proposed approaches utilize a fully supervised object recognition model within the captioning approach. Such models, however, tend to generate sentences which only consist of objects predicted by the recognition models, excluding instances of the classes without labelled training examples. In this paper, we propose a new challenging scenario that targets the image captioning problem in a fully zero-shot learning setting, where the goal is to be able to generate captions of test images containing objects that are not seen during training. The proposed approach jointly uses a novel zero-shot object detection model and a template-based sentence generator. Our experiments show promising results on the COCO dataset.

243 Spatially and Temporally Efficient Non-local Attention Network for Video-based Person Re-Identification

Chih-Ting Liu (National Taiwan University), Chih-Wei Wu (National Taiwan University), Yu-Chiang Frank Wang (National Taiwan University), Shao-Yi Chien (National Taiwan University)

Video-based person re-identification (Re-ID) aims at matching video sequences of pedestrians across non-overlapping cameras. It is a practical yet challenging task of how to embed spatial and temporal information of a video into its feature representation. While most existing methods learn the video characteristics by aggregating image-wise features and designing attention mechanisms in Neural Networks, they only explore the correlation between frames at high-level features. In this work, we target at refining the intermediate features as well as high-level features with non-local attention operations and make two contributions. (i) We propose a Non-local Video Attention Network (NVAN) to incorporate video characteristics into the representation at multiple feature levels. (ii) We further introduce a Spatially and Temporally Efficient Non-local Video Attention Network (STE-NVAN) to reduce the computation complexity by exploring spatial and temporal redundancy presented in pedestrian videos. Extensive experiments show that our NVAN outperforms state-of-the-arts by 3.8% in rank-1 accuracy on MARS dataset and confirms our STE-NVAN displays a much superior computation footprint compared to existing methods.

244 Global Aggregation then Local Distribution in Fully Convolutional Networks

Xiangtai Li (Peking University), Li Zhang (University of Oxford), Ansheng You (Peking University), Maoke Yang (DeepMotion), Yunhai Tong (Peking University), Kuiyuan Yang (DeepMotion)

It has been widely proven that modelling long-range dependencies in fully convolutional networks (FCNs) via global aggregation modules is critical for complex scene understanding tasks such as semantic segmentation and object detection. However, global aggregation is often dominated by features of large patterns and tends to oversmooth regions that contain small patterns (e.g., boundaries and small objects). To resolve this problem, we propose to first use *Global Aggregation* and then *Local Distribution*, which is called GALD, where long-range dependencies are more confidently used inside large pattern regions and vice versa. The size of each pattern at each position is estimated in the network as a per-channel mask map. GALD is end-to-end trainable and can be easily plugged into existing FCNs with various global aggregation modules for a wide range of vision tasks, and consistently improves the performance of state-of-the-art object detection and instance segmentation approaches. In particular, GALD used in semantic segmentation achieves new state-of-the-art per-

formance on Cityscapes test set with mIoU 83.3%. Code is available at: <https://github.com/lxtGH/GALD-Net>.

245 **ClueNet : A Deep Framework for Occluded Pedestrian Pose Estimation**

Perla Sai Raj Kishore (Institute of Engineering & Management), Sudip Das (Indian Statistical Institute), Partha Sarathi Mukherjee (Indian Statistical Institute), Ujjwal Bhattacharya (ISI Kolkata)

Pose estimation of a pedestrian helps to gather information about the current activity or the instant behaviour of the subject. Such information is useful for autonomous vehicles, augmented reality, video surveillance, etc. Although a large volume of pedestrian detection studies are available in the literature, detection of the same in situations of significant occlusions still remains a challenging task. In this work, we take a step further to propose a novel deep learning framework, called ClueNet, to detect as well as estimate the entire pose of occluded pedestrians in an unsupervised manner. ClueNet is a two stage framework where the first stage generates visual clues for the second stage to accurately estimate the pose of occluded pedestrians. The first stage employs a multi-task network to segment the visible parts and predict a bounding box enclosing the visible and occluded regions for each pedestrian. The second stage uses these predictions from the first stage for pose estimation. Here we propose a novel strategy, called Mask and Predict, to train our ClueNet to estimate the pose even for occluded regions. Additionally, we make use of various other training strategies to further improve our results. The proposed work is first of its kind and the experimental results on CityPersons and MS COCO datasets show the superior performance of our approach over existing methods.

152 **AutoCorrect: Deep Inductive Alignment of Noisy Geometric Annotations**

Honglie Chen (University of Oxford), Weidi Xie (University of Oxford), Andrea Vedaldi (University of Oxford), Andrew Zisserman (University of Oxford)

We propose AutoCorrect, a method to automatically learn object-annotation alignments from a dataset with annotations affected by geometric noise. The method is based on a consistency loss that enables deep neural networks to be trained, given only noisy annotations as input, to correct the annotations. When some noise-free annotations are available, we show that the consistency loss reduces to a stricter self-supervised loss. We also show that the method can implicitly leverage object symmetries to reduce the ambiguity arising in correcting noisy annotations. When multiple objects-annotation pairs are present in an image, we introduce a spatial memory map that allows the network to correct annotations sequentially,

one at a time, while accounting for all other annotations in the image and corrections performed so far. Through ablation, we show the benefit of these contributions, demonstrating excellent result on geo-spatial imagery. Specifically, we show result using a new dataset of Railway tracks as well as the public Inria Building benchmarks, achieving new state-of-the-art results for the latter.

247 Improving Object Detection from Scratch via Gated Feature Reuse

Zhiqiang Shen (Carnegie Mellon University), Honghui Shi (IBM, UIUC), Jiahui Yu (UIUC), Hai Phan (Carnegie Mellon University), Rogerio Feris (IBM Research AI, MIT-IBM Watson AI Lab), Liangliang Cao (HelloVera), Ding Liu (UIUC), Xinchao Wang (Stevens Institute of Technology), Thomas Huang (UIUC), Marios Savvides (Carnegie Mellon University)

In this paper, we present a simple and parameter-efficient drop-in module for one-stage object detectors like SSD when learning from scratch (i.e., without pre-trained models). We call our module GFR (Gated Feature Reuse), which exhibits two main advantages. First, we introduce a novel gate-controlled prediction strategy enabled by Squeeze-and-Excitation to adaptively enhance or attenuate supervision at different scales based on the input object size. As a result, our model is more effective in detecting diverse sizes of objects. Second, we propose a feature-pyramids structure to squeeze rich spatial and semantic features into a single prediction layer, which strengthens feature representation and reduces the number of parameters to learn. We apply the proposed structure on DSOD and SSD detection frameworks, and evaluate the performance on PASCAL VOC 2007, 2012 and COCO datasets. With fewer model parameters, GFR-DSOD outperforms the baseline DSOD by 1.4%, 1.1%, 1.7% and 0.7%, respectively. GFR-SSD also outperforms the original SSD and SSD with dense prediction by 3.6% and 2.8% on VOC 2007 dataset.

158 Mining Discriminative Food Regions for Accurate Food Recognition

Jianing Qiu (Imperial College London), Po Wen Lo (Imperial College London), Yingnan Sun (Imperial College London), Siyao Wang (Imperial College London), Benny Lo (Imperial College London)

Automatic food recognition is the very first step towards the passive dietary monitoring. In this paper, we address the problem of food recognition by mining discriminative food regions. Taking inspiration from Adversarial Erasing, a strategy that progressively discovers discriminative object regions for the weakly supervised semantic segmentation, we propose a novel network architecture in which a primary network maintains the base accuracy of classifying an input image, an auxiliary network adversarially mines discriminative food regions, and a region network classifies the resulting mined regions. The global (original input image) and the local

(mined region) representations are then integrated for the final prediction. The proposed architecture denoted as PAR-Net is end-to-end trainable, and highlights discriminative regions in an online fashion. In addition, we introduce a new fine-grained food dataset named as Sushi-50, which consists of 50 different sushi categories. Extensive experiments have been conducted to evaluate the proposed approach. On three food datasets chosen (Food-101, Vireo-172, and Sushi-50), our method performs consistently and achieves state-of-the-art results (top-1 testing accuracy of 90.4%, 90.2%, 92.0%, respectively) compared with other existing approaches.

249 Meta Learning for Unsupervised Clustering

Han-Ul Kim (Korea University), Yeong Jun Koh (Chungnam National University), Chang-Su Kim (Korea University)

Learning an embedding space is essential in clustering. Deep learning has been used recently for this purpose, yielding impressive clustering results. However, it remains challenging to discover clusters in small data, which are insufficient to train deep networks. To address this challenge, we adopt the meta learning strategy, which learns to learn new tasks efficiently. We propose a novel meta learner, called MLC-Net, which mimics numerous clustering tasks during the training to learn an effective embedding space for new clustering tasks. MLC-Net has three building blocks: encoder, centroid, and prediction blocks. The encoder block transforms input patterns into an embedding space, while the centroid block estimates a representative feature for each cluster, called pseudo-centroid. It makes the embedding space more effective and more reliable, by learning an embedding space and a pseudo-centroid estimator jointly. Extensive experimental results on the Omniglot, MNIST, and Mini-ImageNet datasets demonstrate that MLC-Net achieves the state-of-the-art unsupervised clustering, as well as few-shot classification, performances.

250 Enhanced 3D convolutional networks for crowd counting

Zhikang Zou (Huazhong University of Science and Technology), Huiliang Shao (Huazhong University of Science and Technology), Xiaoye Qu (Huazhong University of Science and Technology), Wei Wei (Huazhong University of Science and Technology), Pan Zhou (Huazhong University of Science and Technology)

Recently, convolutional neural networks (CNNs) are the leading defacto method for crowd counting. However, when dealing with video datasets, CNN-based methods still process each video frame independently, thus ignoring the powerful temporal information between consecutive frames. In this work, we propose a novel architecture termed as “temporal channel-aware” (TCA) block, which achieves the capability of exploiting the temporal interdependencies among video sequences. Specifically, we in-

corporate 3D convolution kernels to encode local spatio-temporal features. Furthermore, the global contextual information is encoded into modulation weights which adaptively recalibrate channel-aware feature responses. With the local and global context combined, the proposed block enhances the discriminative ability of the feature representations and contributes to more precise results in diverse scenes. By stacking TCA blocks together, we obtain the deep trainable architecture called enhanced 3D convolutional networks (E3D). The experiments on three benchmark datasets show that the proposed method delivers state-of-the-art performance. To verify the generality, an extended experiment is conducted on a vehicle dataset TRANCOS and our approach beats previous methods by large margins.

251 Image Classification with Hierarchical Multigraph Networks

Boris Knyazev (University of Guelph), Xiao Lin (SRI International), Mohamed Amer (RobustAI), Graham Taylor (University of Guelph)

Graph Convolutional Networks (GCNs) are a class of general models that can learn from graph structured data. Despite being general, GCNs are admittedly inferior to convolutional neural networks (CNNs) when applied to vision tasks, mainly due to the lack of domain knowledge that is hardcoded into CNNs, such as spatially oriented translation invariant filters. However, a great advantage of GCNs is the ability to work on irregular inputs, such as superpixels of images. This could significantly reduce the computational cost of image reasoning tasks. Another key advantage inherent to GCNs is the natural ability to model multirelational data. Building upon these two promising properties, in this work, we show best practices for designing GCNs for image classification; in some cases even outperforming CNNs on the MNIST, CIFAR-10 and PASCAL image datasets.

Segmentation and Grouping

252 Towards Weakly Supervised Semantic Segmentation in 284 3D Graph-Structured Point Clouds of Wild Scenes

Haiyan Wang (City University of New York), Xuejian Rong (City University of New York), Liang Yang (City University of New York), YingLi Tian (City University of New York)

The deficiency of 3D segmentation labels is one of the main obstacles to effective point cloud segmentation, especially for wild scenes with varieties of different objects. To alleviate this issue, we propose a novel graph convolutional deep framework for large-scale semantic scene segmentation in point clouds with solely 2D supervision. Different with

numerous preceding multi-view supervised approaches focusing on single object point clouds, we argue that 2D supervision is also capable of providing enough guidance information for training 3D semantic segmentation model of natural scene point clouds while not explicitly capturing their inherent structures, even with only single view per sample. Specifically, a Graph-based Pyramid Feature Network (GPFN) is designed to implicitly infer both global and local features of point sets, and a perspective rendering and semantic fusion module are proposed to provide refined 2D supervision signals for training along with a 2D-3D joint optimization strategy. Extensive experimental results demonstrate the effectiveness of our 2D supervised framework which achieves comparable results with the state-of-the-art approaches trained with full 3D labels for semantic point cloud segmentation on the popular S3DIS benchmark.

253 Fast-SCNN: Fast Semantic Segmentation Network **289**

Rudra Poudel (Toshiba Research Europe, Ltd.), Stephan Liwicki (Toshiba Research Europe, Ltd.), Roberto Cipolla (University of Cambridge)

The encoder-decoder framework is state-of-the-art for offline semantic image segmentation. Since the rise in autonomous systems, real-time computation is increasingly desirable. In this paper, we introduce fast segmentation convolutional neural network (Fast-SCNN), an above real-time semantic segmentation model on high resolution image data (1024x2048px) suited to efficient computation on embedded devices with low memory and power. We introduce a ‘learning to downsample’ module which computes low-level features for multiple resolution branches simultaneously. Our network combines spatial detail at high resolution with deep features extracted at lower resolution, yielding an accuracy of 68.0% mean intersection over union at 123.5 frames per second on full scale image of Cityscapes. We also show that large scale pre-training is unnecessary. We thoroughly validate our experiments with ImageNet pre-training and the coarse labeled data of Cityscapes. Finally, we show even faster computation with competitive results on subsampled inputs, without any network modifications.

254 Dual Graph Convolutional Network for Semantic Segmentation

Li Zhang (University of Oxford), Xiangtai Li (Peking University), Anurag Arnab (University of Oxford), Kuiyuan Yang (DeepMotion), Yunhai Tong (Peking University), Philip Torr (University of Oxford)

Exploiting long-range contextual information is key for pixel-wise prediction tasks such as semantic segmentation. In contrast to previous work that uses multi-scale feature fusion or dilated convolutions, we propose a

novel graph-convolutional network (GCN) to address this problem. Our Dual Graph Convolutional Network (DGCNet) models the global context of the input feature by modelling two orthogonal graphs in a single framework. The first component models spatial relationships between pixels in the image, whilst the second models interdependencies along the channel dimensions of the network’s feature map. This is done efficiently by projecting the feature into a new, lower-dimensional space where all pairwise interactions can be modelled, before reprojecting into the original space. Our simple method provides substantial benefits over a strong baseline and achieves state-of-the-art results on both Cityscapes (82.0% mean IoU) and Pascal Context (53.7% mean IoU) datasets. Our code is available at: <https://github.com/lzrobots/DGCNet>.

255 Where are the Masks: Instance Segmentation with Image-level Supervision

Issam Hadj Laradji (University of British Columbia), David Vazquez (Element AI), Mark Schmidt (University of British Columbia)

A major obstacle in instance segmentation is that existing methods often need many per-pixel labels in order to be effective. These labels require large human effort and for certain applications, such labels are not readily available. To address this limitation, we propose a novel framework that can effectively train with image-level labels, which are significantly cheaper to acquire. For instance, one can do an internet search for the term “car” and obtain many images where a car is present with minimal effort. Our framework consists of two stages: (1) train a classifier to generate pseudo masks for the objects of interest; (2) train a fully supervised Mask R-CNN on these pseudo masks. Our two main contribution are proposing a pipeline that is simple to implement and is amenable to different segmentation methods; and achieves new state-of-the-art results for this problem setup. Our results are based on evaluating our method on PASCAL VOC 2012, a standard dataset for weakly supervised methods, where we demonstrate major performance gains compared to existing methods with respect to mean average precision.

256 Geometry-Aware End-to-End Skeleton Detection

Weijian Xu (University of California, San Diego), Gaurav Parmar (University of California, San Diego), Zhuowen Tu (University of California, San Diego)

In this paper, we propose a new skeleton detection method that is geometry-aware and can be learned in an end-to-end fashion. Recent approaches in this area are based primarily on the holistically-nested edge detector (HED) that is learned in a fundamentally bottom-up fashion by minimizing a pixel-wise cross-entropy loss. Here, we introduce a new objective function inspired by the Hausdorff distance that carries both global and

local shape information and is made differentiable through an end-to-end neural network framework. When compared with the existing approaches on several widely adopted skeleton benchmarks, our method achieves state-of-the-art results under the standard F-measure. This sheds some light towards directly incorporating shape and geometric constraints in an end-to-end fashion for image segmentation and detection problems — a viewpoint that has been mostly neglected in the past.

153 Adaptive Compression-based Lifelong Learning **257**

Shivangi Srivastava (Wageningen University and Research), Maxim Berman (KU Leuven), Matthew Blaschko (KU Leuven), Devis Tuia (Wageningen University and Research)

The problem of a deep learning model losing performance on a previously learned task when fine-tuned to a new one is a widespread phenomenon, known as Catastrophic forgetting. There are two major ways to mitigate this problem: either preserving activations of the initial network during training with a new task; or restricting the new network activations to remain close to the initial ones. The latter approach falls under the denomination of lifelong learning, where the model is updated in a way that it performs well on both old and new tasks, without having access to the old task's training samples anymore.

Recently, approaches like pruning networks for freeing network capacity during sequential learning of tasks have been gaining in popularity. Such approaches allow learning small networks while making redundant parameters available for the next tasks. The common problem encountered with these approaches is that the pruning percentage is hard-coded, irrespective of the number of samples, of the complexity of the learning task and of the number of classes in the dataset. We propose a method based on Bayesian optimization to perform adaptive compression/pruning of the network and show its effectiveness in lifelong learning. Our method learns to perform heavy pruning for small and/or simple datasets while using milder compression rates for large and/or complex data. Experiments on classification and semantic segmentation demonstrate the applicability of learning network compression, where we are able to effectively preserve performances along sequences of tasks of varying complexity.

258 Working Hands: A Hand-Tool Assembly Dataset for Image Segmentation and Activity Mining

Roy Shilkrot (Stony Brook University), Supreeth Narasimhaswamy (Stony Brook University), Saif Vazir (Stony Brook University), Minh Hoai Nguyen (Stony Brook University)

Computer vision in manufacturing is a decades long effort into automatic inspection and verification of the work pieces, while visual recognition focusing on the human operators is becoming ever prominent. Semantic

segmentation is an exemplary vision task that is key to enabling crucial assembly applications such as completion time tracking and manual process verification. However, focus on segmentation of human hands while performing complex tasks such as manual assembly is still lacking. Segmenting hands from tools, work pieces, background and other body parts is difficult because of self-occlusions and intricate hand grips and poses. In this paper we introduce Working Hands, a dataset of pixel-level annotated images of hands performing 13 different tool-based assembly tasks, from both real-world captures and virtual-world renderings, with RGB+D images from a high-resolution range camera and ray casting engine. Moreover, using the dataset, we can learn a generic Hand-Task Descriptor that is useful for retrieving hand images and video performing similar operations across different non-annotated datasets.

259 DABNet: Depth-wise Asymmetric Bottleneck for Real-time Semantic Segmentation

Gen Li (Sungkyunkwan University), Joongkyu Kim (Sungkyunkwan University)

As a pixel-level prediction task, semantic segmentation needs large computational cost with enormous parameters to obtain high performance. Recently, due to the increasing demand for autonomous systems and robots, it is significant to make a trade-off between accuracy and inference speed. In this paper, we propose a novel Depth-wise Asymmetric Bottleneck (DAB) module to address this dilemma, which efficiently adopts depth-wise asymmetric convolution and dilated convolution to build a bottleneck structure. Based on the DAB module, we design a Depth-wise Asymmetric Bottleneck Network (DABNet) especially for real-time semantic segmentation, which creates sufficient receptive field and densely utilizes the contextual information. Experiments on Cityscapes and CamVid datasets demonstrate that the proposed DABNet achieves a balance between speed and precision. Specifically, without any pretrained model and post-processing, it achieves 70.1% Mean IoU on the Cityscapes test dataset with only 0.76 million parameters and a speed of 104 FPS on a single GTX 1080Ti card.

260 Feature Pyramid Encoding Network for Real-time Semantic Segmentation

Mengyu Liu (University of Manchester), Hujun Yin (University of Manchester)

Although current deep learning methods have achieved impressive results for semantic segmentation, they incur high computational costs and have a huge number of parameters. For real-time applications, inference speed and memory usage are two important factors. To address the challenge, we propose a lightweight feature pyramid encoding network (FPENet) to

make a good trade-off between accuracy and speed. Specifically, we use a feature pyramid encoding block to encode multi-scale contextual features with depthwise dilated convolutions in all stages of the encoder. A mutual embedding upsample module is introduced in the decoder to aggregate the high-level semantic features and low-level spatial details efficiently. The proposed network outperforms existing real-time methods with fewer parameters and improved inference speed on the Cityscapes and CamVid benchmark datasets. Specifically, FPENet achieves 68.0% mean IoU on the Cityscapes test set with only 0.4M parameters and 102 FPS speed on an NVIDIA TITAN V GPU.

142 Convolutional CRFs for Semantic Segmentation 261

Marvin Teichmann (University of Cambridge), Roberto Cipolla (University of Cambridge)

For the challenging semantic image segmentation task the best performing models have traditionally combined the structured modelling capabilities of Conditional Random Fields (CRFs) with the feature extraction power of CNNs. In more recent works however, CRF post-processing has fallen out of favour. We argue that this is mainly due to the slow training and inference speeds of CRFs, as well as the difficulty of learning the internal CRF parameters. To overcome both issues we propose to add the assumption of conditional independence to the framework of fully-connected CRFs. This allows us to reformulate the inference in terms of convolutions, which can be implemented highly efficiently on GPUs. Doing so speeds up inference and training by two orders of magnitude. All parameters of the convolutional CRFs can easily be optimized using backpropagation. Towards the goal of facilitating further CRF research we have made our implementations publicly available.

262 Visuomotor Understanding for Representation Learning of Driving Scenes

Seokju Lee (Korea Advanced Institute of Science and Technology), Junsik Kim (Korea Advanced Institute of Science and Technology), Tae-Hyun Oh (MIT CSAIL), Yongseop Jeong (Korea Advanced Institute of Science and Technology), Donggeun Yoo (Lunit), Stephen Lin (Microsoft Research), In So Kweon (Korea Advanced Institute of Science and Technology)

Dashboard cameras capture a tremendous amount of driving scene video each day. These videos are purposefully coupled with vehicle sensing data, such as from the speedometer and inertial sensors, providing an additional sensing modality for free. In this work, we leverage the large-scale unlabeled yet naturally paired data for visual representation learning in the driving scenario. A representation is learned in an end-to-end self-supervised framework for predicting dense optical flow from a single frame with paired sensing data. We postulate that success on this task

requires the network to learn semantic and geometric knowledge in the ego-centric view. For example, forecasting a future view to be seen from a moving vehicle requires an understanding of scene depth, scale, and movement of objects. We demonstrate that our learned representation can benefit other tasks that require detailed scene understanding and outperforms competing unsupervised representations on semantic segmentation.

263 Referring Expression Object Segmentation with Caption-Aware Consistency

Yi-Wen Chen (Academia Sinica), Yi-Hsuan Tsai (NEC Labs America), Tiantian Wang (University of California at Merced), Yen-Yu Lin (Academia Sinica), Ming-Hsuan Yang (University of California at Merced)

Referring expressions are natural language descriptions that identify a particular object within a scene and are widely used in our daily conversations. In this work, we focus on segmenting the object in an image specified by a referring expression. To this end, we propose an end-to-end trainable comprehension network that consists of the language and visual encoders to extract feature representations from both domains. We introduce the spatial-aware dynamic filters to transfer knowledge from the language domain to the visual one, and can effectively capture the spatial information of the specified object. To further make useful communication between the language and visual modules, we employ a caption generation network that takes features shared across both domains as input, and improves both representations via a consistency that enforces the generated sentence to be similar to the original query. We evaluate the proposed framework on three referring expression datasets and show that our method performs favorably against the state-of-the-art algorithms.

264 Dispersion based Clustering for Unsupervised Person Re-identification

Guodong Ding (Nanjing University of Science and Technology), Salman Khan (Australian National University (ANU)), Zhenmin Tang (Nanjing University of Science and Technology)

The cumbersome acquisition of large-scale annotations for person re-identification task makes its deployment difficult in real-world scenarios. It is necessary to teach models to learn without explicit supervision. This paper proposes a simple but effective clustering approach for unsupervised person re-identification. We explore a basic concept in statistics, namely *dispersion*, to achieve a robust clustering criterion. Dispersion reflects the compactness of a cluster when assessed within and reveals the separation when measured at the inter-cluster level. Based on this insight, we propose a Dispersion based Clustering (DBC) approach which performs better at discovering the underlying data patterns. The approach can automatically prioritize standalone data points and prevents poor clustering. Our

extensive experimental results demonstrate that the proposed methodology outperforms the state-of-the-art unsupervised methods on person re-identification.

265 Oval Shape Constraint based Optic Disc and Cup Segmentation in Fundus Photographs

Jun Wu (Northwestern Polytechnical University), Kaiwei Wang (Northwestern Polytechnical University), Zongjiang Shang (Northwestern Polytechnical University), Jie Xu (Beijing Tongren Hospital), Dayong Ding (Vistel Inc.), Xirong Li (Renmin University of China), Gang Yang (Renmin University of China)

In a fundus photograph, morphological changes of the optic disc and cup are crucial for diagnosing optic neuropathy. To achieve an accurate pixel-wise segmentation of the optic disc and cup, the domain-specific knowledge such as the oval shape constraint has not been sufficiently explored in most of the existing methods, leading to unacceptable geometric distortions in many cases. Few attempts try to consider the general convexity constraint or specific building geometric properties, but they are still not suitable for the typical oval shape segmentation. In this paper, an oval shape constraint based loss function (OS-loss) is proposed to improve the existing deep learning network for segmenting optic disc and cup. A penalty point set is proposed to represent unreasonable contour points of a target object using the oval shape constraint. These points will be penalized and integrated into the training loss function of the baseline network. Further, an oval-friendly metric called shape error (SE) is proposed to better reflect the fitness of two oval contours. Experiments on the public RIM-ONE-r3 dataset with 159 fundus photographs and a private W10K dataset with 9,879 fundus photographs prove the effectiveness of the proposed OS-loss function. Compared to the original CE-net, the mean error of the Cup to Disc Ratio (CDR) of the proposed OS-loss method in the RIM-ONE-r3 dataset decreases 1.98%. In the W10K dataset, the mean CDR error decreases by 1.03% for the ResU-net and decreases by 2.1% for the CE-net.

266 An end-to-end deep learning approach for simultaneous background modeling and subtraction

Víctor Mondéjar-Guerra (UDC), Jorge Novo (University of A Coruña), José Rouco (University of A Coruña), Marcos Ortega (University of A Coruña)

Background subtraction is an active research topic due to its great utility on many video analysis applications. In this work, a new approach for background subtraction employing an end-to-end deep learning architecture is proposed. The proposed architecture consists in two nested networks that are trained together. The first one extracts the background model features of the scene from a small group of frames. The second performs the

subtraction operation given the previous features and a target frame. In contrast to most of the recent deep learning proposals, our trained model can be used on any scene without the need of being retrained. The method has been trained and evaluated using the public CDnet2014 database following a scene-wise cross-validation approach. The obtained results show a competitive performance of the proposed method on background subtraction, proving its ability to extrapolate to unseen scenes.

Video Analysis

267 **Spatio-temporal Relational Reasoning for Video Question Answering** 295

Gursimran Singh (University of British Columbia), Leonid Sigal (University of British Columbia), Jim Little (University of British Columbia)

Video question answering is the task of automatically answering questions about videos. Among query types which include identification, localization, and counting, the most challenging questions enquire about relationships among different entities. Answering such questions, and many others, require modeling relationships between entities in the spatial domain and evolution of those relationships in the temporal domain. We argue that current approaches have limited capacity to model such long-range spatial and temporal dependencies. To address these challenges, we present a novel spatio-temporal reasoning neural module which enables modeling complex multi-entity relationships in space and long-term ordered dependencies in time. We evaluate our module on two benchmark datasets which require spatio-temporal reasoning: TGIF-QA and SVQA. We achieve state-of-the-art performance on both datasets. More significantly, we achieve substantial improvements on some of the most challenging question types, like counting, which demonstrate the effectiveness of our proposed spatio-temporal relational module.

268 **Mutual Suppression Network for Video Prediction using Disentangled Features** 296

Jungbeom Lee (Seoul National University), Jangho Lee (Seoul National University), Sungmin Lee (Seoul National University), Sungroh Yoon (Seoul National University)

Video prediction has been considered a difficult problem because the video contains not only high-dimensional spatial information but also complex temporal information. Video prediction can be performed by finding features in recent frames, and using them to generate approximations to upcoming frames. We approach this problem by disentangling spatial and

temporal features in videos. We introduce a mutual suppression network (MSnet) which are trained in an adversarial manner and then produces spatial features which are free of motion information, and motion features with no spatial information. MSnet then uses motion-guided connection within an encoder-decoder-based architecture to transform spatial features from a previous frame to the time of an upcoming frame. We show how MSnet can be used for video prediction using disentangled representations. We also carry out experiments to assess the effectiveness of our method to disentangle features. MSnet obtains better results than other recent video prediction methods even though it has simpler encoders.

269 Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading

Xinshuo Weng (Carnegie Mellon University), Kris Kitani (Carnegie Mellon University)

We focus on the word-level visual lipreading, which requires recognizing the word being spoken, given only the video but not the audio. State-of-the-art methods explore the use of end-to-end neural networks, including a shallow (up to three layers) 3D convolutional neural network (CNN) + a deep 2D CNN (e.g., ResNet) as the front-end to extract visual features, and a recurrent neural network (e.g., bidirectional LSTM) as the back-end for classification. In this work, we propose to replace the shallow 3D CNNs + deep 2D CNNs front-end with recent successful deep 3D CNNs — two-stream (i.e., grayscale video and optical flow streams) I3D. We evaluate different combinations of front-end and back-end modules with the grayscale video and optical flow inputs on the LRW dataset. The experiments show that, compared to the shallow 3D CNNs + deep 2D CNNs front-end, the deep 3D CNNs front-end with pre-training on the large-scale image and video datasets (e.g., ImageNet and Kinetics) can improve the classification accuracy. Also, we demonstrate that using the optical flow input alone can achieve comparable performance as using the grayscale video as input. Moreover, the two-stream network using both the grayscale video and optical flow inputs can further improve the performance. Overall, our two-stream I3D front-end with a Bi-LSTM back-end results in an absolute improvement of 5.3% over the previous art on the LRW dataset.

270 Motion-Aware Feature for Improved Video Anomaly Detection

Yi Zhu (University of California, Merced), Shawn Newsam (University of California, Merced)

Motivated by our observation that motion information is the key to good anomaly detection performance in video, we propose a temporal augmented network to learn a motion-aware feature. This feature alone can achieve competitive performance with previous state-of-the-art meth-

ods, and when combined with them, can achieve significant performance improvements. Furthermore, we incorporate temporal context into the Multiple Instance Learning (MIL) ranking model by using an attention block. The learned attention weights can help to differentiate between anomalous and normal video segments better. With the proposed motion-aware feature and the temporal MIL ranking model, we outperform previous approaches by a large margin on both anomaly detection and anomalous action recognition tasks in the UCF Crime dataset.

271 Attention-based Facial Behavior Analytics in Social Communication

Lezi Wang (Rutgers University), Chongyang Bai (Dartmouth College), Maksim Bolonkin (Dartmouth College), VS Subrahmanian (Dartmouth College), Judee Burgoon (University of Arizona), Norah Dunbar (University of California, Santa Barbara), Dimitris Metaxas (Rutgers University)

In this study, we address a cross-domain problem of applying computer vision approaches to reason about human facial behavior when people play *The Resistance* game. To capture the facial behaviors, we first collect several hours of video where the participants playing *The Resistance* game assume the roles of deceivers (spies) vs truth-tellers (villagers). We develop a novel attention-based neural network (NN) that advances the state of the art in understanding how a NN predicts the players' roles. This is accomplished by discovering through learning those pixels and related frames which are discriminative and contributed the most to the NN's inference. We demonstrate the effectiveness of our attention-based approach in discovering the frames and facial Action Units (AUs) that contributed to the NN's class decision. Our results are consistent with the current communication theory on deception.

272 Searching for Ambiguous Objects in Videos using Relational Referring Expressions

Hazan Anayurt (Middle East Technical University), Sezai Artun Ozyegin (Middle East Technical University), Ulfet Cetin (Middle East Technical University), Utku Aktas (Middle East Technical University), Sinan Kalkan (Middle East Technical University)

Humans frequently use referring (identifying) expressions to refer to objects. Especially in ambiguous settings, humans prefer expressions (called relational referring expressions) that describe an object with respect to a distinguishing, unique object. Unlike studies on video object search using referring expressions, in this paper, our focus is on (i) relational referring expressions in highly ambiguous settings, and (ii) methods that can both generate and comprehend a referring expression. For this goal, we first introduce a new dataset for video object search with referring expressions that includes numerous copies of the objects, making it difficult to use non-relational expressions. Moreover, we train two baseline

deep networks on this dataset, which show promising results. Finally, we propose a deep attention network that significantly outperforms the baselines on our dataset. The dataset and the codes are available at <https://github.com/hazananayurt/viref>.

156 Hybrid Deep Network for Anomaly Detection 273

Trong Nguyen Nguyen (University of Montreal), Jean Meunier (University of Montreal)

In this paper, we propose a deep convolutional neural network (CNN) for anomaly detection in surveillance videos. The model is adapted from a typical auto-encoder working on video patches under the perspective of sparse combination learning. Our CNN focuses on (unsupervisedly) learning common characteristics of normal events with the emphasis of their spatial locations (by supervised losses). To our knowledge, this is the first work that directly adapts the patch position as the target of a classification sub-network. The model is capable to provide a score of anomaly assessment for each video frame. Our experiments were performed on 4 benchmark datasets with various anomalous events and the obtained results were competitive with state-of-the-art studies.

274 VStegNET: Video Steganography Network using Spatio-Temporal features and Micro-Bottleneck

Suraj Kumar (Aligarh Muslim University), Aayush Mishra (IIT Mandi), Saiful Islam (Aligarh Muslim University), Aditya Nigam (IIT Mandi)

Steganography is the practice of hiding a secret message in a cover message such that the cover stays indiscernible after hiding and only the intended recipients can extract the secret from it. Traditional image steganography techniques hide the secret image into high-frequency regions of the cover images. These techniques typically result in lower embedding ratios and easy detection. In this paper, we propose VStegNET, a video steganography network that extracts spatio-temporal features using 3D-CNN and micro-bottleneck (Hourglass) which is the first of its kind in the literature of video steganography. The proposed network hides $M \times N$ (RGB) secret video frames into same sized cover video frames. We have trained our model on UCF 101 action recognition video dataset and evaluated its performance using various quantitative metrics (APD, PSNR, and SSIM) and compared it with previous the state-of-the-art. Furthermore, we have also presented a detailed analysis, supporting the proposal's superiority over image steganography models. Finally, several standard steganalysis tools like StegExpose, SRNET, etc. have been used to justify the steganographic capabilities of VStegNET.

275 Order Matters: Shuffling Sequence Generation for Video Prediction

Junyan Wang (Newcastle University), BingZhang Hu (Newcastle University), Yang Long (Newcastle University), Yu Guan (Newcastle University)

Predicting future frames in natural video sequences is a new challenge that is receiving increasing attention in the computer vision community. However, existing models suffer from severe loss of temporal information when the predicted sequence is long. Compared to previous methods focusing on generating more realistic contents, this paper extensively studies the importance of sequential order information for video generation. A novel Shuffling sEquence gEneration network (SEE-Net) is proposed that can learn to discriminate between natural and unnatural sequential orders by shuffling the video frames and comparing them to the real video sequences. Systematic experiments on three datasets with both synthetic and real-world videos manifest the effectiveness of shuffling sequence generation for video prediction in our proposed model and demonstrate state-of-the-art performance by both qualitative and quantitative evaluations. The source code is available at <https://github.com/andrewjywang/SEENet>.

276 Multi-Grained Spatio-temporal Modeling for Lip-reading

Chenhao Wang (Institute of Computing Technology, Chinese Academy of Sciences)

Lip-reading aims to recognize speech content from videos via visual analysis of speakers' lip movements. This is a challenging task due to the existence of – words which involve identical or highly similar lip movements, as well as diverse lip appearances and motion patterns among the speakers. To address these challenges, we propose a novel lip-reading model which captures not only the nuance between words but also styles of different speakers, by a multi-grained spatio-temporall modeling of the speaking process. Specifically, we first extract both frame-level fine-grained features and short-term medium-grained features by the visual front-end, which are then combined to obtain discriminative representations for words with similar phonemes. Next, a bidirectional ConvLSTM augmented with temporal attention aggregates spatio-temporal information in the entire input sequence, which is expected to be able to capture the coarse-gained patterns of each word and robust to various conditions in speaker identity, lighting conditions, and so on. By making full use of the information from different levels in a unified framework, the model is not only able to distinguish words with similar pronunciations, but also becomes robust to appearance changes. We evaluate our method on two challenging word-level lip-reading benchmarks and show the effectiveness of the proposed method, which also demonstrate the above claims.

277 Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks

Zitong Yu (CMVS, University of Oulu), Xiaobai Li (University of Oulu), Guoying Zhao (University of Oulu)

Recent studies demonstrated that the average heart rate (HR) can be measured from facial videos based on non-contact remote photoplethysmography (rPPG). However, for many medical applications (e.g., atrial fibrillation (AF) detection) knowing only the average HR is not sufficient, and measuring precise rPPG signals from face for heart rate variability (HRV) analysis is needed. Here we propose an rPPG measurement method, which is the first work to use deep spatio-temporal networks for reconstructing precise rPPG signals from raw facial videos. With the constraint of trend-consistency with ground truth pulse curves, our method is able to recover rPPG signals with accurate pulse peaks. Comprehensive experiments are conducted on two benchmark datasets, and results demonstrate that our method can achieve superior performance on both HR and HRV levels comparing to the state-of-the-art methods. We also achieve promising results of using reconstructed rPPG signals for AF detection and emotion recognition.

278 Spatio-Temporal Associative Representation for Video Person Re-Identification

Guile Wu (Queen Mary University of London), Xiatian Zhu (Samsung AI Centre, Cambridge), Shaogang Gong (Queen Mary University of London)

Learning discriminative spatio-temporal representation is the key for solving video re-identification (re-id) challenges. Most existing methods focus on learning appearance features and/or selecting image frames, but ignore optimising the compatibility and interaction of appearance and motion attentive information. To address this limitation, we propose a novel model to learning Spatio-Temporal Associative Representation (STAR). We design local frame-level spatio-temporal association to learn discriminative attentive appearance and short-term motion features, and global video-level spatio-temporal association to form compact and discriminative holistic video representation. We further introduce a pyramid ranking regulariser for facilitating end-to-end model optimisation. Extensive experiments demonstrate the superiority of STAR against state-of-the-art methods on four video re-id benchmarks, including MARS, DukeMTMC-VideoReID, iLIDS-VID and PRID-2011.

279 Use What You Have: Video retrieval using representations from collaborative experts

Yang Liu (University of Oxford), Samuel Albanie (University of Oxford), Arsha Nagrani (University of Oxford), Andrew Zisserman (University of Oxford)

The rapid growth of video on the internet has made searching for video content using natural language queries a significant challenge. Human generated queries for video datasets ‘in the wild’ vary a lot in terms of degree of specificity, with some queries describing ‘specific details’ such as the names of famous identities, content from speech, or text available on the screen. Our goal is to condense the multi-modal, extremely high dimensional information from videos into a single, compact video representation for the task of video retrieval using free-form text queries, where the degree of specificity is open-ended. For this we exploit existing knowledge in the form of pretrained semantic embeddings which include ‘general’ features such as motion, appearance, and scene features from visual content, and more ‘specific’ cues from ASR and OCR which may not always be available, but allow for more fine-grained disambiguation when present. We propose a collaborative experts model to aggregate information effectively from these different pretrained experts. The effectiveness of our approach is demonstrated empirically, setting new state-of-the-art performances on five retrieval benchmarks: MSR-VTT, LSMDC, MSVD, DiDeMo, and ActivityNet, while simultaneously reducing the number of parameters used by prior work. Code and data can be found at www.robots.ox.ac.uk/~vgg/research/collaborative-experts/.

280 VideoNavQA: Bridging the Gap between Visual and Embodied Question Answering

Catalina Cangea (University of Cambridge), Eugene Belilovsky (Mila), Aaron Courville (Université de Montréal)

Embodied Question Answering (EQA) is a recently proposed task, where an agent is placed in a rich 3D environment and must act based solely on its egocentric input to answer a given question. The desired outcome is that the agent learns to combine capabilities such as scene understanding, navigation and language understanding in order to perform complex reasoning in the visual world. However, initial advancements combining standard vision and language methods with imitation and reinforcement learning algorithms have shown EQA might be too complex and challenging for these techniques. In order to investigate the feasibility of EQA-type tasks, we build the VideoNavQA dataset that contains pairs of questions and videos generated in the House3D environment. The goal of this dataset is to assess question-answering performance from nearly-ideal navigation paths, while considering a much more complete variety of questions than current instantiations of the EQA task. We investigate several models, adapted from popular VQA methods, on this new benchmark. This establishes an initial understanding of how well VQA-style methods can perform within

this novel EQA paradigm.

281 MLGCN: Multi-Laplacian Graph Convolutional Networks for Human Action Recognition

Ahmed Mazari (Sorbonne Université), Hichem Sahbi (Sorbonne University)

Convolutional neural networks are nowadays witnessing a major success in different pattern recognition problems. These learning models were basically designed to handle vectorial data such as images but their extension to non-vectorial and semi-structured data (namely graphs with variable sizes, topology, etc.) remains a major challenge, though a few interesting solutions are currently emerging. In this paper, we introduce MLGCN; a novel spectral Multi-Laplacian Graph Convolutional Network. The main contribution of this method resides in a new design principle that learns graph-laplacians as convex combinations of other elementary laplacians—each one dedicated to a particular topology of the input graphs. We also introduce a novel pooling operator, on graphs, that proceeds in two steps: context-dependent node expansion is achieved, followed by a global average pooling; the strength of this two-step process resides in its ability to preserve the discrimination power of nodes while achieving permutation invariance. Experiments conducted on SBU and UCF-101 datasets, show the validity of our method for the challenging task of action recognition.

This book was typeset with \LaTeX , using the `memoir` class. The text was prepared entirely in GNU Emacs with `AUCTeX` package.

Typesetting by Kirill Sidorov.
Logo design by Aylwyn Bowen.

