

Proceedings

**29th IEEE Conference on Computer Vision
and Pattern Recognition**

CVPR 2016

Proceedings

29th IEEE Conference on Computer Vision and Pattern Recognition

26 June – 1 July 2016
Las Vegas, Nevada



Los Alamitos, California
Washington • Tokyo



All rights reserved.

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries may photocopy beyond the limits of US copyright law, for private use of patrons, those articles in this volume that carry a code at the bottom of the first page, provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Other copying, reprint, or republication requests should be addressed to: IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, P.O. Box 133, Piscataway, NJ 08855-1331.

The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Computer Society, or the Institute of Electrical and Electronics Engineers, Inc.

IEEE Computer Society Order Number E5869
BMS Part Number CFP16003-USB
ISBN 978-1-4673-8850-4

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: +1 800 272 6657
Fax: +1 714 821 4641
<http://computer.org/cspress>
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: +1 732 981 0060
Fax: +1 732 981 9667
[http://shop.ieee.org/store/
customer-service@ieee.org](http://shop.ieee.org/store/customer-service@ieee.org)

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: +81 3 3408 3118
Fax: +81 3 3408 3553
tokyo.ofc@computer.org

Individual paper REPRINTS may be ordered at: <reprints@computer.org>

Editorial production by Lisa O'Conner
Cover art production by Mark Bartosik



**IEEE Computer Society
Conference Publishing Services (CPS)**

<http://www.computer.org/cps>

2016 IEEE Conference on Computer Vision and Pattern Recognition

CVPR 2016

Table of Contents

Message from General Chair and Program Chairs.....	xlvi
Organizing Committee and Area Chairs.....	xlvi
Outstanding Reviewers.....	xlvi

Oral & Spotlight Session 1-1A

O1-1A: Image Captioning and Question Answering

Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data	1
<i>Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell</i>	
Generation and Comprehension of Unambiguous Object Descriptions	11
<i>Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy</i>	
Stacked Attention Networks for Image Question Answering	21
<i>Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola</i>	
Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction	30
<i>Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han</i>	
Neural Module Networks	39
<i>Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein</i>	

S1-1A: Language and Vision

Learning Deep Representations of Fine-Grained Visual Descriptions	49
<i>Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele</i>	
Multi-cue Zero-Shot Learning with Strong Supervision	59
<i>Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele</i>	
Latent Embeddings for Zero-Shot Classification	69
<i>Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele</i>	

One-Shot Learning of Scene Locations via Feature Trajectory Transfer	78
<i>Roland Kwitt, Sebastian Hegenbart, and Marc Niethammer</i>	
Learning Attributes Equals Multi-Source Domain Generalization	87
<i>Chuang Gan, Tianbao Yang, and Boqing Gong</i>	
Anticipating Visual Representations from Unlabeled Video	98
<i>Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba</i>	

Oral & Spotlight Session 1-1B

O1-1B: Matching and Alignment

Learning to Assign Orientations to Feature Points	107
<i>Kwang Moo Yi, Yannick Verdie, Pascal Fua, and Vincent Lepetit</i>	
Learning Dense Correspondence via 3D-Guided Cycle Consistency	117
<i>Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros</i>	
The Global Patch Collider	127
<i>Shenlong Wang, Sean Ryan Fanello, Christoph Rhemann, Shahram Izadi, and Pushmeet Kohli</i>	
Joint Probabilistic Matching Using m-Best Solutions	136
<i>Seyed Hamid RezaTofighi, Anton Milani, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid</i>	
Face Alignment Across Large Poses: A 3D Solution	146
<i>Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li</i>	

S1-1B: Segmentation and Contour Detection

Interactive Segmentation on RGBD Images via Cue Selection	156
<i>Jie Feng, Brian Price, Scott Cohen, and Shih-Fu Chang</i>	
Layered Scene Decomposition via the Occlusion-CRF	165
<i>Chen Liu, Pushmeet Kohli, and Yasutaka Furukawa</i>	
Affinity CNN: Learning Pixel-Centric Pairwise Relations for Figure/Ground Embedding	174
<i>Michael Maire, Takuya Narihira, and Stella X. Yu</i>	
Weakly Supervised Object Boundaries	183
<i>Anna Khoreva, Rodrigo Benenson, Mohamed Omran, Matthias Hein, and Bernt Schiele</i>	
Object Contour Detection with a Fully Convolutional Encoder-Decoder Network	193
<i>Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang</i>	

Poster Session P1-1

What Value Do Explicit High Level Concepts Have in Vision to Language Problems?	203
<i>Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel</i>	
Fast Detection of Curved Edges at Low SNR	213
<i>Nati Ofir, Meirav Galun, Boaz Nadler, and Ronen Basri</i>	

Object Skeleton Extraction in Natural Images by Fusing Scale-Associated Deep Side Outputs	222
<i>Wei Shen, Kai Zhao, Yuan Jiang, Yan Wang, Zhijiang Zhang, and Xiang Bai</i>	
Learning Relaxed Deep Supervision for Better Edge Detection	231
<i>Yu Liu and Michael S. Lew</i>	
Occlusion Boundary Detection via Deep Exploration of Context	241
<i>Huan Fu, Chaohui Wang, Dacheng Tao, and Michael J. Black</i>	
SemiContour: A Semi-Supervised Learning Approach for Contour Detection	251
<i>Zizhao Zhang, Fuyong Xing, Xiaoshuang Shi, and Lin Yang</i>	
Learning to Localize Little Landmarks	260
<i>Saurabh Singh, Derek Hoiem, and David Forsyth</i>	
InterActive: Inter-Layer Activeness Propagation	270
<i>Lingxi Xie, Liang Zheng, Jingdong Wang, Alan Yuille, and Qi Tian</i>	
Exploit Bounding Box Annotations for Multi-Label Object Recognition	280
<i>Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai</i>	
TI-POOLING: Transformation-Invariant Pooling for Feature Learning in Convolutional Neural Networks	289
<i>Dmitry Laptev, Nikolay Savinov, Joachim M. Buhmann, and Marc Pollefeys</i>	
Fashion Style in 128 Floats: Joint Ranking and Classification Using Weak Data for Feature Extraction	298
<i>Edgar Simo-Serra and Hiroshi Ishikawa</i>	
Equiangular Kernel Dictionary Learning with Applications to Dynamic Texture Analysis	308
<i>Yuhui Quan, Chenglong Bao, and Hui Ji</i>	
Compact Bilinear Pooling	317
<i>Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell</i>	
Accumulated Stability Voting: A Robust Descriptor from Descriptors of Multiple Scales	327
<i>Tsun-Yi Yang, Yen-Yu Lin, and Yung-Yu Chuang</i>	
CoMaL: Good Features to Match on Object Boundaries	336
<i>Swarna K. Ravindran and Anurag Mittal</i>	
Progressive Feature Matching with Alternate Descriptor Selection and Correspondence Enrichment	346
<i>Yuan-Ting Hu and Yen-Yu Lin</i>	
A New Finsler Minimal Path Model with Curvature Penalization for Image Segmentation and Closed Contour Detection	355
<i>Da Chen, Jean-Marie Mirebeau, and Laurent D. Cohen</i>	
Scale-Aware Alignment of Hierarchical Image Segmentation	364
<i>Yuhua Chen, Dengxin Dai, Jordi Pont-Tuset, and Luc Van Gool</i>	
Deep Interactive Object Selection	373
<i>Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang</i>	
Pull the Plug? Predicting If Computers or Humans Should Segment Images	382
<i>Danna Gurari, Suyog Dutt Jain, Margrit Betke, and Kristen Grauman</i>	

In the Shadows, Shape Priors Shine: Using Occlusion to Improve Multi-region Segmentation	392
<i>Yuka Kihara, Matvey Soloviev, and Tsuhan Chen</i>	
Convexity Shape Constraints for Image Segmentation	402
<i>Loic A. Royer, David L. Richmond, Carsten Rother, Bjoern Andres, and Dagmar Kainmueller</i>	
MCMC Shape Sampling for Image Segmentation with Nonparametric Shape Priors	411
<i>Ertunc Erdil, Sinan YILdiRiM, Müjdat Çetin, and Tolga Tasdizen</i>	
From Noise Modeling to Blind Image Denoising	420
<i>Fengyuan Zhu, Guangyong Chen, and Pheng Ann Heng</i>	
Efficient and Robust Color Consistency for Community Photo Collections	430
<i>Jaesik Park, Yu-Wing Tai, Sudipta N. Sinha, and In So Kweon</i>	
Needle-Match: Reliable Patch Matching under High Uncertainty	439
<i>Or Lotan and Michal Irani</i>	
ReconNet: Non-Iterative Reconstruction of Images from Compressively Sensed Measurements	449
<i>Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok</i>	
Soft-Segmentation Guided Object Motion Deblurring	459
<i>Jinshan Pan, Zhe Hu, Zhixun Su, Hsin-Ying Lee, and Ming-Hsuan Yang</i>	
Two Illuminant Estimation and User Correction Preference	469
<i>Dongliang Cheng, Abdelrahman Kamel, Brian Price, Scott Cohen, and Michael S. Brown</i>	
Deep Contrast Learning for Salient Object Detection	478
<i>Guanbin Li and Yizhou Yu</i>	
Multiview Image Completion with Space Structure Propagation	488
<i>Seung-Hwan Baek, Inchang Choi, and Min H. Kim</i>	
Composition-Preserving Deep Photo Aesthetics Assessment	497
<i>Long Mai, Hailin Jin, and Feng Liu</i>	
Automatic Image Cropping: A Computational Complexity Study	507
<i>Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li</i>	
A Deeper Look at Saliency: Feature Contrast, Semantics, and Beyond	516
<i>Neil D. B. Bruce, Christopher Catton, and Sasa Janjic</i>	
Spatially Binned ROC: A Comprehensive Saliency Metric	525
<i>Calden Wloka and John Tstotsos</i>	
GraB: Visual Saliency via Novel Graph Model and Background Priors	535
<i>Qiaosong Wang, Wen Zheng, and Robinson Piramuthu</i>	
Predicting When Saliency Maps are Accurate and Eye Fixations Consistent	544
<i>Anna Volokitin, Michael Gygli, and Xavier Boix</i>	
Split and Match: Example-Based Adaptive Patch Sampling for Unsupervised Style Transfer	553
<i>Oriel Frigo, Neus Sabater, Julie Delon, and Pierre Hellier</i>	

Detection and Accurate Localization of Circular Fiducials under Highly Challenging Conditions	562
<i>Lilian Calvet, Pierre Gurdjos, Carsten Griwodz, and Simone Gasparini</i>	
Scene Recognition with CNNs: Objects, Scales and Dataset Bias	571
<i>Luis Herranz, Shuqiang Jiang, and Xiangyang Li</i>	
Learning Action Maps of Large Environments via First-Person Vision	580
<i>Nicholas Rhinehart and Kris M. Kitani</i>	
Single-Image Crowd Counting via Multi-Column Convolutional Neural Network	589
<i>Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma</i>	
Shallow and Deep Convolutional Networks for Saliency Prediction	598
<i>Junting Pan, Elisa Sayrol, Xavier Giro-I-Nieto, Kevin McGuinness, and Noel E. O'Connor</i>	
Sample and Filter: Nonparametric Scene Parsing via Efficient Filtering	607
<i>Mohammad Najafi, Sarah Taghavi Namin, Mathieu Salzmann, and Lars Petersson</i>	
DeLay: Robust Spatial Layout Estimation for Cluttered Indoor Scenes	616
<i>Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese</i>	
A Text Detection System for Natural Scenes with Convolutional Feature Learning and Cascaded Classification	625
<i>Siyu Zhu and Richard Zanibbi</i>	
Reversible Recursive Instance-Level Object Segmentation	633
<i>Xiaodan Liang, Yunchao Wei, Xiaohui Shen, Zequn Jie, Jiashi Feng, Liang Lin, and Shuicheng Yan</i>	
Coherent Parametric Contours for Interactive Video Object Segmentation	642
<i>Yao Lu, Xue Bai, Linda Shapiro, and Jue Wang</i>	
Manifold SLIC: A Fast Method to Compute Content-Sensitive Superpixels	651
<i>Yong-Jin Liu, Cheng-Chi Yu, Min-Jing Yu, and Ying He</i>	
Deep Saliency with Encoded Low Level Distance Map and High Level Features	660
<i>Gayoung Lee, Yu-Wing Tai, and Junmo Kim</i>	
Instance-Level Segmentation for Autonomous Driving with Deep Densely Connected MRFs	669
<i>Ziyu Zhang, Sanja Fidler, and Raquel Urtasun</i>	
DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection	678
<i>Nian Liu and Junwei Han</i>	
Object Co-segmentation via Graph Optimized-Flexible Manifold Ranking	687
<i>Rong Quan, Junwei Han, Dingwen Zhang, and Feiping Nie</i>	
Primary Object Segmentation in Videos via Alternate Convex Optimization of Foreground and Background Distributions	696
<i>Won-Dong Jang, Chulwoo Lee, and Chang-Su Kim</i>	
Automatic Fence Segmentation in Videos of Dynamic Scenes	705
<i>Renjiao Yi, Jue Wang, and Ping Tan</i>	
Discovering the Physical Parts of an Articulated Object Class from Multiple Videos	714
<i>Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari</i>	

A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation	724
<i>F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung</i>	
Learning Temporal Regularity in Video Sequences	733
<i>Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis</i>	
Bilateral Space Video Segmentation	743
<i>Nicolas Märki, Federico Perazzi, Oliver Wang, and Alexander Sorkine-Hornung</i>	
ReD-SFA: Relation Discovery Based Slow Feature Analysis for Trajectory Clustering	752
<i>Zhang Zhang, Kaiqi Huang, Tieniu Tan, Peipei Yang, and Jun Li</i>	

Oral & Spotlight Session 1-2A

O1-2A: Object Recognition and Detection

Training Region-Based Object Detectors with Online Hard Example Mining	761
<i>Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick</i>	
Deep Residual Learning for Image Recognition	770
<i>Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun</i>	
You Only Look Once: Unified, Real-Time Object Detection	779
<i>Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi</i>	
LocNet: Improving Localization Accuracy for Object Detection	789
<i>Spyros Gidaris and Nikos Komodakis</i>	
Sketch Me That Shoe	799
<i>Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy</i>	

S1-2A: Object Detection 1

Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images	808
<i>Shuran Song and Jianxiong Xiao</i>	
Object Detection from Video Tubelets with Convolutional Neural Networks	817
<i>Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang</i>	
Learning with Side Information through Modality Hallucination	826
<i>Judy Hoffman, Saurabh Gupta, and Trevor Darrell</i>	
Object-Proposal Evaluation Protocol is 'Gameable'	835
<i>Neelima Chavali, Harsh Agrawal, Aroma Mahendru, and Dhruv Batra</i>	
HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection	845
<i>Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun</i>	
We Don't Need No Bounding-Boxes: Training Object Class Detectors Using Only Human Verification	854
<i>Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari</i>	
Factors in Finetuning Deep Model for Object Detection with Long-Tail Distribution	864
<i>Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang</i>	

Oral & Spotlight Session 1-2B

O1-2B: Vision with Alternative Sensors

Information-Driven Adaptive Structured-Light Scanners	874
<i>Guy Rosman, Daniela Rus, and John W. Fisher III</i>	
Simultaneous Optical Flow and Intensity Estimation from an Event Camera	884
<i>Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger</i>	
Macroscopic Interferometry: Rethinking Depth Estimation with Frequency-Domain Time-of-Flight	893
<i>Achuta Kadambi, Jamie Schiel, and Ramesh Raskar</i>	
ASP Vision: Optically Computing the First Layer of Convolutional Neural Networks Using Angle Sensitive Pixels	903
<i>Huaijin G. Chen, Suren Jayasuriya, Jiyue Yang, Judy Stephen, Sriram Sivaramakrishnan, Ashok Veeraraghavan, and Alyosha Molnar</i>	
Computational Imaging for VLBI Image Reconstruction	913
<i>Katherine L. Bouman, Michael D. Johnson, Daniel Zoran, Vincent L. Fish, Sheperd S. Doeleman, and William T. Freeman</i>	

S1-2B: Video Analysis 1

You Lead, We Exceed: Labor-Free Video Concept Learning by Jointly Exploiting Web Videos and Images	923
<i>Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei</i>	
Track and Segment: An Iterative Unsupervised Approach for Video Object Proposals	933
<i>Fanyi Xiao and Yong Jae Lee</i>	
Beyond Local Search: Tracking Objects Everywhere with Instance-Specific Proposals	943
<i>Gao Zhu, Fatih Porikli, and Hongdong Li</i>	
Groupwise Tracking of Crowded Similar-Appearance Targets from Low-Continuity Image Sequences	952
<i>Hongkai Yu, Youjie Zhou, Jeff Simmons, Craig P. Przybyla, Yuewei Lin, Xiaochuan Fan, Yang Mi, and Song Wang</i>	
Social LSTM: Human Trajectory Prediction in Crowded Spaces	961
<i>Alexandre Alahi, Kratharth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese</i>	
What Players do with the Ball: A Physically Constrained Interaction Modeling	972
<i>Andrii Maksai, Xinchao Wang, and Pascal Fua</i>	
Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization	982
<i>Ting Yao, Tao Mei, and Yong Rui</i>	

Poster Session P1-2

Direct Prediction of 3D Body Poses from Motion Compensated Sequences	991
<i>Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua</i>	
Video2GIF: Automatic Generation of Animated GIFs from Video	1001
<i>Michael Gygli, Yale Song, and Liangliang Cao</i>	
NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis	1010
<i>Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang</i>	
Progressively Parsing Interactional Objects for Fine Grained Action Detection	1020
<i>Bingbing Ni, Xiaokang Yang, and Shenghua Gao</i>	
Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning	1029
<i>Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang</i>	
From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection	1039
<i>Jingjing Meng, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan</i>	
Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs	1049
<i>Zheng Shou, Dongang Wang, and Shih-Fu Chang</i>	
Summary Transfer: Exemplar-Based Subset Selection for Video Summarization	1059
<i>Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman</i>	
POD: Discovering Primary Objects in Videos Based on Evolutionary Refinement of Object Recurrence, Background, and Primary Object Models	1068
<i>Yeong Jun Koh, Won-Dong Jang, and Chang-Su Kim</i>	
What If We Do Not have Multiple Videos of the Same Action? — Video Action Localization Using Web Images	1077
<i>Waqas Sultani and Mubarak Shah</i>	
Beyond F-Formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images	1086
<i>Lu Zhang and Hayley Hung</i>	
DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations	1096
<i>Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang</i>	
SketchNet: Sketch Classification with Web Images	1105
<i>Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao</i>	
Embedding Label Structures for Fine-Grained Feature Representation	1114
<i>Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang</i>	
Fine-Grained Image Classification by Exploring Bipartite-Graph Labels	1124
<i>Feng Zhou and Yuanqing Lin</i>	
Picking Deep Filter Responses for Fine-Grained Image Recognition	1134
<i>Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian</i>	

SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-Grained Recognition	1143
<i>Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas</i>	
Fine-Grained Categorization and Dataset Bootstrapping Using Deep Metric Learning with Humans in the Loop	1153
<i>Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie</i>	
Mining Discriminative Triplets of Patches for Fine-Grained Classification	1163
<i>Yaming Wang, Jonghyun Choi, Vlad I. Morariu, and Larr S. Davis</i>	
Part-Stacked CNN for Fine-Grained Visual Categorization	1173
<i>Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang</i>	
Learning Compact Binary Descriptors with Unsupervised Deep Neural Networks	1183
<i>Kevin Lin, Jiwen Lu, Chu-Song Chen, and Jie Zhou</i>	
Solving Small-Piece Jigsaw Puzzles by Growing Consensus	1193
<i>Kilho Son, Daniel Moreno, James Hays, and David B. Cooper</i>	
Pairwise Matching through Max-Weight Bipartite Belief Propagation	1202
<i>Zhen Zhang, Qinfeng Shi, Julian McAuley, Wei Wei, Yanning Zhang, and Anton van den Hengel</i>	
Structured Feature Similarity with Explicit Feature Map	1211
<i>Takumi Kobayashi</i>	
Temporal Epipolar Regions	1220
<i>Mor Dar and Yael Moses</i>	
Recurrent Attention Models for Depth-Based Person Identification	1229
<i>Albert Haque, Alexandre Alahi, and Li Fei-Fei</i>	
Learning a Discriminative Null Space for Person Re-identification	1239
<i>Li Zhang, Tao Xiang, and Shaogang Gong</i>	
Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification	1249
<i>Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang</i>	
How Far are We from Solving Pedestrian Detection?	1259
<i>Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele</i>	
Similarity Learning with Spatial Constraints for Person Re-identification	1268
<i>Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng</i>	
Sample-Specific SVM Learning for Person Re-identification	1278
<i>Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan</i>	
Joint Learning of Single-Image and Cross-Image Representations for Person Re-identification	1288
<i>Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang</i>	
A Multi-level Contextual Model for Person Recognition in Photo Albums	1297
<i>Haoxiang Li, Jonathan Brandt, Zhe Lin, Xiaohui Shen, and Gang Hua</i>	

Unsupervised Cross-Dataset Transfer Learning for Person Re-identification	1306
<i>Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian</i>	
Pedestrian Detection Inspired by Appearance Constancy and Shape Symmetry	1316
<i>Jiale Cao, Yanwei Pang, and Xuelong Li</i>	
Recurrent Convolutional Network for Video-Based Person Re-identification	1325
<i>Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller</i>	
Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function	1335
<i>De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng</i>	
Top-Push Video-Based Person Re-identification	1345
<i>Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng</i>	
Improving Person Re-identification via Pose-Aware Multi-shot Matching	1354
<i>Yeong-Jun Cho and Kuk-Jin Yoon</i>	
Hierarchical Gaussian Descriptor for Person Re-identification	1363
<i>Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato</i>	
STCT: Sequentially Training Convolutional Networks for Visual Tracking	1373
<i>Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu</i>	
Determining Occlusions from Space and Time Image Reconstructions	1382
<i>Juan-Manuel Pérez-Rúa, Tomas Crivelli, Patrick Bouthemy, and Patrick Pérez</i>	
Online Multi-object Tracking via Structural Constraint Event Aggregation	1392
<i>Ju Hong Yoon, Chang-Ryeol Lee, Ming-Hsuan Yang, and Kuk-Jin Yoon</i>	
Staple: Complementary Learners for Real-Time Tracking	1401
<i>Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip H. S. Torr</i>	
Robust Optical Flow Estimation of Double-Layer Images under Transparency or Reflection	1410
<i>Jiaolong Yang, Hongdong Li, Yuchao Dai, and Robby T. Tan</i>	
Siamese Instance Search for Tracking	1420
<i>Ran Tao, Efstratios Gavves, and Arnold W. M. Smeulders</i>	
Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking	1430
<i>Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg</i>	
3D Part-Based Sparse Tracker with Automatic Synchronization and Registration	1439
<i>Adel Bibi, Tianzhu Zhang, and Bernard Ghanem</i>	
Recurrently Target-Attending Tracking	1449
<i>Zhen Cui, Shengtao Xiao, Jiashi Feng, and Shuicheng Yan</i>	
Structured Regression Gradient Boosting	1459
<i>Ferran Diego and Fred A. Hamprecht</i>	
Loss Functions for Top-k Error: Analysis and Insights	1468
<i>Maksim Lapin, Matthias Hein, and Bernt Schiele</i>	

Metric Learning as Convex Combinations of Local Models with Generalization Guarantees	1478
<i>Valentina Zantedeschi, Rémi Emonet, and Marc Sebban</i>	
Efficient Training of Very Deep Neural Networks for Supervised Hashing	1487
<i>Ziming Zhang, Yuting Chen, and Venkatesh Saligrama</i>	
Information Bottleneck Learning Using Privileged Information for Visual Recognition	1496
<i>Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto</i>	

Oral & Spotlight Session 2-1A

O2-1A: Recognition and Parsing in 3D

3D Action Recognition from Novel Viewpoints	1506
<i>Hossein Rahmani and Ajmal Mian</i>	
3D Shape Attributes	1516
<i>David F. Fouhey, Abhinav Gupta, and Andrew Zisserman</i>	
Three-Dimensional Object Detection and Layout Prediction Using Clouds of Oriented Gradients	1525
<i>Zhile Ren and Erik B. Sudderth</i>	
3D Semantic Parsing of Large-Scale Indoor Spaces	1534
<i>Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese</i>	
Dense Human Body Correspondences Using Convolutional Networks	1544
<i>Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li</i>	

S2-1A: Recognition Beyond Objects

Geometry-Informed Material Recognition	1554
<i>Joseph Degol, Mani Golparvar-Fard, and Derek Hoiem</i>	
Towards Open Set Deep Networks	1563
<i>Abhijit Bendale and Terrance E. Boult</i>	
What's Wrong with That Object? Identifying Images of Unusual Objects by Modelling the Detection Score Distribution	1573
<i>Peng Wang, Lingqiao Liu, Chunhua Shen, Zi Huang, Anton van den Hengel, and Heng Tao Shen</i>	
Large-Scale Location Recognition and the Geometric Burstiness Problem	1582
<i>Torsten Sattler, Michal Havlena, Konrad Schindler, and Marc Pollefeys</i>	
Regularity-Driven Building Facade Matching between Aerial and Street Views	1591
<i>Mark Wolff, Robert T. Collins, and Yanxi Liu</i>	
Do Computational Models Differ Systematically from Human Object Perception?	1601
<i>R. T. Pramod and S. P. Arun</i>	

Oral & Spotlight Session 2-1B

O2-1B: Image Processing and Restoration

Contour Detection in Unstructured 3D Point Clouds	1610
<i>Timo Hackel, Jan D. Wegner, and Konrad Schindler</i>	
Unsupervised Learning of Edges	1619
<i>Yin Li, Manohar Paluri, James M. Rehg, and Piotr Dollár</i>	
Blind Image Deblurring Using Dark Channel Prior	1628
<i>Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang</i>	
Deeply-Recursive Convolutional Network for Image Super-Resolution	1637
<i>Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee</i>	
Accurate Image Super-Resolution Using Very Deep Convolutional Networks	1646
<i>Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee</i>	

S2-1B: Image Processing and Restoration

RAW Image Reconstruction Using a Self-Contained sRGB-JPEG Image with Only 64 KB Overhead	1655
<i>Rang M. H. Nguyen and Michael S. Brown</i>	
Group MAD Competition? A New Methodology to Compare Objective Image Quality Models	1664
<i>Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang</i>	
Non-local Image Dehazing	1674
<i>Dana Berman, Tali Treibitz, and Shai Avidan</i>	
A Holistic Approach to Cross-Channel Image Noise Modeling and Its Application to Image Denoising	1683
<i>Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim</i>	
Multispectral Images Denoising by Intrinsic Tensor Sparsity Regularization	1692
<i>Qi Xie, Qian Zhao, Deyu Meng, Zongben Xu, Shuhang Gu, Wangmeng Zuo, and Lei Zhang</i>	
A Comparative Study for Single Image Blind Deblurring	1701
<i>Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang</i>	

Poster Session P2-1

Spatiotemporal Bundle Adjustment for Dynamic 3D Reconstruction	1710
<i>Minh Vo, Srinivasa G. Narasimhan, and Yaser Sheikh</i>	
Inextensible Non-Rigid Shape-from-Motion by Second-Order Cone Programming	1719
<i>Ajad Chhatkuli, Daniel Pizarro, Toby Collins, and Adrien Bartoli</i>	
Optimal Relative Pose with Unknown Correspondences	1728
<i>Johan Fredriksson, Viktor Larsson, Carl Olsson, and Fredrik Kahl</i>	

Homography Estimation from the Common Self-Polar Triangle of Separate Ellipses	1737
<i>Haifei Huang, Hui Zhang, and Yiu-Ming Cheung</i>	
Heterogeneous Light Fields	1745
<i>Maximilian Diebold, Bernd Jähne, and Alexander Gatto</i>	
A Consensus-Based Framework for Distributed Bundle Adjustment	1754
<i>Anders Eriksson, John Bastian, Tat-Jun Chin, and Mats Isaksson</i>	
Globally Optimal Manhattan Frame Estimation in Real-Time	1763
<i>Kyungdon Joo, Tae-Hyun Oh, Junsik Kim, and In So Kweon</i>	
Mirror Surface Reconstruction under an Uncalibrated Camera	1772
<i>Kai Han, Kwan-Yee K. Wong, Dirk Schnieders, and Miaomiao Liu</i>	
A Hole Filling Approach Based on Background Reconstruction for View Synthesis in 3D Video	1781
<i>Guibo Luo, Yuesheng Zhu, Zhaotian Li, and Liming Zhang</i>	
A Direct Least-Squares Solution to the PnP Problem with Unknown Focal Length	1790
<i>Yinqiang Zheng and Laurent Kneip</i>	
Efficient Intersection of Three Quadrics and Applications in Computer Vision	1799
<i>Zuzana Kukelova, Jan Heller, and Andrew Fitzgibbon</i>	
Using Spatial Order to Boost the Elimination of Incorrect Feature Matches	1809
<i>Lior Talker, Yael Moses, and Ilan Shimshoni</i>	
A Probabilistic Framework for Color-Based Point Set Registration	1818
<i>Martin Danelljan, Giulia Meneghetti, Fahad Shahbaz Khan, and Michael Felsberg</i>	
Blind Image Deconvolution by Automatic Gradient Activation	1827
<i>Dong Gong, Mingkui Tan, Yanning Zhang, Anton van den Hengel, and Qinfeng Shi</i>	
PSyCo: Manifold Span Reduction for Super Resolution	1837
<i>Eduardo Pérez-Pellitero, Jordi Salvador, Javier Ruiz-Hidalgo, and Bodo Rosenhahn</i>	
Parametric Object Motion from Blur	1846
<i>Jochen Gast, Anita Sellent, and Stefan Roth</i>	
Image Deblurring Using Smartphone Inertial Sensors	1855
<i>Zhe Hu, Lu Yuan, Stephen Lin, and Ming-Hsuan Yang</i>	
Seven Ways to Improve Example-Based Single Image Super Resolution	1865
<i>Radu Timofte, Rasmus Rothe, and Luc Van Gool</i>	
Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network	1874
<i>Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang</i>	
They are Not Equally Reliable: Semantic Event Search Using Differentiated Concept Classifiers	1884
<i>Xiaojun Chang, Yao-Liang Yu, Yi Yang, and Eric P. Xing</i>	
Going Deeper into First-Person Activity Recognition	1894
<i>Minghuang Ma, Haoqi Fan, and Kris M. Kitani</i>	

Cascaded Interactional Targeting Network for Egocentric Video Analysis	1904
<i>Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, and Qi Tian</i>	
Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos	1914
<i>Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem</i>	
Discriminative Hierarchical Rank Pooling for Activity Recognition	1924
<i>Basura Fernando, Peter Anderson, Marcus Hutter, and Stephen Gould</i>	
Convolutional Two-Stream Network Fusion for Video Action Recognition	1933
<i>Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman</i>	
Learning Activity Progression in LSTMs for Activity Detection and Early Detection	1942
<i>Shugao Ma, Leonid Sigal, and Stan Sclaroff</i>	
VLAD3: Encoding Dynamics of Deep Features for Action Recognition	1951
<i>Yingwei Li, Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos</i>	
A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection	1961
<i>Bharat Singh, Tim K. Marks, Michael Jones, Oncel Tuzel, and Ming Shao</i>	
A Hierarchical Deep Temporal Model for Group Activity Recognition	1971
<i>Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori</i>	
A Hierarchical Pose-Based Approach to Complex Action Understanding Using Dictionaries of Actionlets and Motion Poselets	1981
<i>Ivan Lillo, Juan Carlos Niebles, and Alvaro Soto</i>	
A Key Volume Mining Deep Framework for Action Recognition	1991
<i>Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao</i>	
Improved Hamming Distance Search Using Variable Length Hashing	2000
<i>Eng-Jon Ong and Mirosław Bober</i>	
Shortlist Selection with Residual-Aware Distance Estimator for K-Nearest Neighbor Search	2009
<i>Jae-Pil Heo, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Sung-Eui Yoon</i>	
Supervised Quantization for Similarity Search	2018
<i>Xiaojuan Wang, Ting Zhang, Guo-Jun Qi, Jinhui Tang, and Jingdong Wang</i>	
Efficient Large-Scale Approximate Nearest Neighbor Search on the GPU	2027
<i>Patrick Wieschollek, Oliver Wang, Alexander Sorkine-Hornung, and Hendrik P. A. Lensch</i>	
Collaborative Quantization for Cross-Modal Similarity Search	2036
<i>Ting Zhang and Jingdong Wang</i>	
Aggregating Image and Text Quantized Correlated Components	2046
<i>Thi Quynh Nhi Tran, Hervé Le Borgne, and Michel Crucianu</i>	
Efficient Indexing of Billion-Scale Datasets of Deep Descriptors	2055
<i>Artem Babenko Yandex and Victor Lempitsky</i>	
Deep Supervised Hashing for Fast Image Retrieval	2064
<i>Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen</i>	

Efficient Large-Scale Similarity Search Using Matrix Factorization	2073
<i>Ahmet Iscen, Michael Rabbat, and Teddy Furon</i>	
Incremental Object Discovery in Time-Varying Image Collections	2082
<i>Theodora Kontogianni, Markus Mathias, and Bastian Leibe</i>	
Detecting Migrating Birds at Night	2091
<i>Jia-Bin Huang, Rich Caruana, Andrew Farnsworth, Steve Kelling, and Narendra Ahuja</i>	
When Naïve Bayes Nearest Neighbors Meet Convolutional Neural Networks	2100
<i>Ilija Kuzborskij, Fabio Maria Carlucci, and Barbara Caputo</i>	
Traffic-Sign Detection and Classification in the Wild	2110
<i>Zhe Zhu, Dun Liang, Songhai Zhang, Xiaolei Huang, Baoli Li, and Shimin Hu</i>	
Large Scale Semi-Supervised Object Detection Using Visual and Semantic Knowledge Transfer	2119
<i>Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen</i>	
Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers	2129
<i>Fan Yang, Wongun Choi, and Yuanqing Lin</i>	
Dictionary Pair Classifier Driven Convolutional Neural Networks for Object Detection	2138
<i>Keze Wang, Liang Lin, Wangmeng Zuo, Shuhang Gu, and Lei Zhang</i>	
Monocular 3D Object Detection for Autonomous Driving	2147
<i>Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun</i>	
How Hard Can It Be? Estimating the Difficulty of Visual Search in an Image	2157
<i>Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P. Papadopoulos, and Vittorio Ferrari</i>	
Deep Relative Distance Learning: Tell the Difference between Similar Vehicles	2167
<i>Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang</i>	
Eye Tracking for Everyone	2176
<i>Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba</i>	
Efficient Globally Optimal 2D-to-3D Deformable Shape Matching	2185
<i>Zorah Löhner, Emanuele Rodolà, Frank R. Schmidt, Michael M. Bronstein, and Daniel Cremers</i>	
Ambiguity Helps: Classification with Disagreements in Crowdsourced Annotations	2194
<i>Viktoriia Sharmanska, Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Novi Quadrianto</i>	
A Task-Oriented Approach for Cost-Sensitive Recognition	2203
<i>Roozbeh Mottaghi, Hannaneh Hajishirzi, and Ali Farhadi</i>	
Refining Architectures of Deep Convolutional Neural Networks	2212
<i>Sukrit Shankar, Duncan Robertson, Yani Ioannou, Antonio Criminisi, and Roberto Cipolla</i>	
iLab-20M: A Large-Scale Controlled Object Dataset to Investigate Deep Learning	2221
<i>Ali Borji, Saeed Izadi, and Laurent Itti</i>	

Recursive Recurrent Nets with Attention Modeling for OCR in the Wild	2231
<i>Chen-Yu Lee and Simon Osindero</i>	
Deep Decision Network for Multi-class Image Classification	2240
<i>Venkatesh N. Murthy, Vivek Singh, Terrence Chen, R. Manmatha, and Dorin Comaniciu</i>	
Less is More: Zero-Shot Learning from Online Textual Documents with Noise Suppression	2249
<i>Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel</i>	
Fast Algorithms for Linear and Kernel SVM+	2258
<i>Wen Li, Dengxin Dai, Mingkui Tan, Dong Xu, and Luc Van Gool</i>	

Oral & Spotlight Session 2-2A

O2-2A: Recognition and Labeling

Hierarchically Gated Deep Networks for Semantic Segmentation	2267
<i>Guo-Jun Qi</i>	
Deep Structured Scene Parsing by Learning with Image Descriptions	2276
<i>Liang Lin, Guangrun Wang, Rui Zhang, Ruimao Zhang, Xiaodan Liang, and Wangmeng Zuo</i>	
CNN-RNN: A Unified Framework for Multi-label Image Classification	2285
<i>Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu</i>	
Walk and Learn: Facial Attribute Representation Learning from Egocentric Video and Contextual Data	2295
<i>Jing Wang, Yu Cheng, and Rogerio Schmidt Feris</i>	
CNN-N-Gram for Handwriting Word Recognition	2305
<i>Arik Poznanski and Lior Wolf</i>	

2A: Object Detection 2

Synthetic Data for Text Localisation in Natural Images	2315
<i>Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman</i>	
End-to-End People Detection in Crowded Scenes	2325
<i>Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng</i>	
Real-Time Salient Object Detection with a Minimum Spanning Tree	2334
<i>Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien</i>	
Local Background Enclosure for RGB-D Salient Object Detection	2343
<i>David Feng, Nick Barnes, Shaodi You, and Chris McCarthy</i>	
Adaptive Object Detection Using Adjacency and Zoom Prediction	2351
<i>Yongxi Lu, Tara Javidi, and Svetlana Lazebnik</i>	
Semantic Channels for Fast Pedestrian Detection	2360
<i>Arthur Daniel Costea and Sergiu Nedevschi</i>	
G-CNN: An Iterative Grid Based Object Detector	2369
<i>Mahyar Najibi, Mohammad Rastegari, and Larry S. Davis</i>	

Oral & Spotlight Session 2-2B

O2-2B: Computational Photography and Faces

Recurrent Face Aging	2378
<i>Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe</i>	
Face2Face: Real-Time Face Capture and Reenactment of RGB Videos	2387
<i>Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner</i>	
Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions	2396
<i>Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F. Cohn, and Nicu Sebe</i>	
Visually Indicated Sounds	2405
<i>Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman</i>	
Image Style Transfer Using Convolutional Neural Networks	2414
<i>Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge</i>	

S2-2B: Computational Photography and Biomedical Applications

Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification	2424
<i>Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz</i>	
Hedgehog Shape Priors for Multi-Object Segmentation	2434
<i>Hossam Isack, Olga Veksler, Milan Sonka, and Yuri Boykov</i>	
Latent Variable Graphical Model Selection Using Harmonic Analysis: Applications to the Human Connectome Project (HCP)	2443
<i>Won Hwa Kim, Hyunwoo J. Kim, Nagesh Adluru, and Vikas Singh</i>	
Simultaneous Estimation of Near IR BRDF and Fine-Scale Surface Geometry	2452
<i>Gyeongmin Choe, Srinivasa G. Narasimhan, and In So Kweon</i>	
Do It Yourself Hyperspectral Imaging with Everyday Digital Cameras	2461
<i>Seoung Wug Oh, Michael S. Brown, Marc Pollefeys, and Seon Joo Kim</i>	
Automatic Content-Aware Color and Tone Stylization	2470
<i>Joon-Young Lee, Kalyan Sunkavalli, Zhe Lin, Xiaohui Shen, and In So Kweon</i>	
Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis	2479
<i>Chuan Li and Michael Wand</i>	

Poster Session P2-2

DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation	2487
<i>Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng</i>	

Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation	2497
<i>Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers</i>	
Conformal Surface Alignment with Optimal Möbius Search	2507
<i>Huu Le, Tat-Jun Chin, and David Suter</i>	
Coupled Harmonic Bases for Longitudinal Characterization of Brain Networks	2517
<i>Seong Jae Hwang, Nagesh Adluru, Maxwell D. Collins, Sathya N. Ravi, Barbara B. Bendlin, Sterling C. Johnson, and Vikas Singh</i>	
Automating Carotid Intima-Media Thickness Video Interpretation with Convolutional Neural Networks	2526
<i>Jae Y. Shin, Nima Tajbakhsh, R. Todd Hurst, Christopher B. Kendall, and Jianming Liang</i>	
Context Encoders: Feature Learning by Inpainting	2536
<i>Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros</i>	
Comparative Deep Learning of Hybrid Representations for Image Recommendations	2545
<i>Chenyi Lei, Dong Liu, Weiping Li, Zheng-Jun Zha, and Houqiang Li</i>	
Fast ConvNets Using Group-Wise Brain Damage	2554
<i>Vadim Lebedev and Victor Lempitsky</i>	
Learning to Co-Generate Object Proposals with a Deep Structured Network	2565
<i>Zeeshan Hayder, Xuming He, and Mathieu Salzmann</i>	
DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks	2574
<i>Seyed-Mohsen, Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard</i>	
Blockout: Dynamic Model Selection for Hierarchical Deep Networks	2583
<i>Calvin Murdock, Zhen Li, Howard Zhou, and Tom Duerig</i>	
FireCaffe: Near-Linear Acceleration of Deep Neural Network Training on Compute Clusters	2592
<i>Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, and Kurt Keutzer</i>	
MDL-CW: A Multimodal Deep Learning Framework with CrossWeights	2601
<i>Sarah Rastegar, Mahdieh Soleymani Baghshah, Hamid R. Rabiee, and Seyed Mohsen Shojaei</i>	
Structured Receptive Fields in CNNs	2610
<i>Jörn-Henrik Jacobsen, Jan van Gemert, Zhongyou Lou, and Arnold W. M. Smeulders</i>	
First Person Action Recognition Using Deep Learned Descriptors	2620
<i>Suriya Singh, Chetan Arora, and C. V. Jawahar</i>	
Recognizing Micro-Actions and Reactions from Paired Egocentric Videos	2629
<i>Ryo Yonetani, Kris M. Kitani, and Yoichi Sato</i>	
Mining 3D Key-Pose-Motifs for Action Recognition	2639
<i>Chunyu Wang, Yizhou Wang, and Alan L. Yuille</i>	
Predicting the Where and What of Actors and Actions through Online Action Localization	2648
<i>Khurram Soomro, Haroon Idrees, and Mubarak Shah</i>	

Actions ~ Transformations	2658
<i>Xiaolong Wang, Ali Farhadi, and Abhinav Gupta</i>	
Visual Path Prediction in Complex Scenes with Crowded Moving Objects	2668
<i>Youngjoon Yoo, Kimin Yun, Sangdoo Yun, Jonghee Hong, Hawook Jeong, and Jin Young Choi</i>	
End-to-End Learning of Action Detection from Frame Glimpses in Videos	2678
<i>Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei</i>	
Action Recognition in Video Using Sparse Coding and Relative Features	2688
<i>Anali Alfaro, Domingo Mery, and Alvaro Soto</i>	
Improving Human Action Recognition by Non-action Classification	2698
<i>Yang Wang and Minh Hoai</i>	
Actionness Estimation Using Hybrid Fully Convolutional Networks	2708
<i>Limin Wang, Yu Qiao, Xiaoou Tang, and Luc Van Gool</i>	
Real-Time Action Recognition with Enhanced Motion Vector CNNs	2718
<i>Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang</i>	
Laplacian Patch-Based Image Synthesis	2727
<i>Joo Ho Lee, Inchang Choi, and Min H. Kim</i>	
Rain Streak Removal Using Layer Priors	2736
<i>Yu Li, Robby T. Tan, Xiaojie Guo, Jiangbo Lu, and Michael S. Brown</i>	
Gradient-Domain Image Reconstruction Framework with Intensity-Range and Base-Structure Constraints	2745
<i>Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi</i>	
Removing Clouds and Recovering Ground Observations in Satellite Image Sequences via Temporally Contiguous Robust Matrix Completion	2754
<i>Jialei Wang, Peder A. Olsen, Andrew R. Conn, and Aurélie C. Lozano</i>	
D3: Deep Dual-Domain Based Fast Restoration of JPEG-Compressed Images	2764
<i>Zhangyang Wang, Ding Liu, Shiyu Chang, Qing Ling, Yingzhen Yang, and Thomas S. Huang</i>	
From Bows to Arrows: Rolling Shutter Rectification of Urban Scenes	2773
<i>Vijay Rengarajan, A. N. Rajagopalan, and R. Aravind</i>	
A Weighted Variational Model for Simultaneous Reflectance and Illumination Estimation	2782
<i>Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding</i>	
Visualizing and Understanding Deep Texture Representations	2791
<i>Tsung-Yu Lin and Subhransu Maji</i>	
Robust Kernel Estimation with Outliers Handling for Image Deblurring	2800
<i>Jinshan Pan, Zhouchen Lin, Zhixun Su, and Ming-Hsuan Yang</i>	
Online Collaborative Learning for Open-Vocabulary Visual Classifiers	2809
<i>Hanwang Zhang, Xindi Shang, Wenzhuo Yang, Huan Xu, Huanbo Luan, and Tat-Seng Chua</i>	
Rethinking the Inception Architecture for Computer Vision	2818
<i>Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna</i>	

Cross Modal Distillation for Supervision Transfer	2827
<i>Saurabh Gupta, Judy Hoffman, and Jitendra Malik</i>	
Efficient Point Process Inference for Large-Scale Object Detection	2837
<i>Trung T. Pham, Seyed Hamid RezaTofighi, Ian Reid, and Tat-Jun Chin</i>	
Weakly Supervised Deep Detection Networks	2846
<i>Hakan Bilen and Andrea Vedaldi</i>	
BORDER: An Oriented Rectangles Approach to Texture-Less Object Recognition	2855
<i>Jacob Chan, Jimmy Addison Lee, and Qian Kemao</i>	
Active Image Segmentation Propagation	2864
<i>Suyog Dutt Jain and Kristen Grauman</i>	
Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks	2874
<i>Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick</i>	
RIFD-CNN: Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection	2884
<i>Gong Cheng, Peicheng Zhou, and Junwei Han</i>	
Reinforcement Learning for Visual Object Detection	2894
<i>Stefan Mathe, Aleksis Pirinen, and Cristian Sminchisescu</i>	
Detecting Repeating Objects Using Patch Correlation Analysis	2903
<i>Inbar Huberman and Raanan Fattal</i>	
Analyzing Classifiers: Fisher Vectors and Deep Neural Networks	2912
<i>Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek</i>	
Learning Deep Features for Discriminative Localization	2921
<i>Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba</i>	
Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels	2930
<i>Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick</i>	
Learning Aligned Cross-Modal Representations from Weakly Aligned Data	2940
<i>Lluís Castrejón, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba</i>	
A Probabilistic Collaborative Representation Based Approach for Pattern Classification	2950
<i>Sijia Cai, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng</i>	
Learning Structured Inference Neural Networks with Label Relations	2960
<i>Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori</i>	
Discriminative Multi-modal Feature Fusion for RGBD Indoor Scene Recognition	2969
<i>Hongyuan Zhu, Jean-Baptiste Weibel, and Shijian Lu</i>	
Conditional Graphical Lasso for Multi-label Image Classification	2977
<i>Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao</i>	
Region Ranking SVM for Image Classification	2987
<i>Zijun Wei and Minh Hoai</i>	

Predicting Motivations of Actions by Leveraging Text	2997
<i>Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba</i>	
BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition	3006
<i>Jakub Sochor, Adam Herout, and Jiri Havel</i>	
Highway Vehicle Counting in Compressed Domain	3016
<i>Xu Liu, Zilei Wang, Jiashi Feng, and Hongsheng Xi</i>	
Camera Calibration from Periodic Motion of a Pedestrian	3025
<i>Shiyao Huang, Xianghua Ying, Jiangpeng Rong, Zeyu Shang, and Hongbin Zha</i>	

Oral & Spotlight Session 3-1A

O3-1A: Actions and Human Pose

Dynamic Image Networks for Action Recognition	3034
<i>Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould</i>	
Detecting Events and Key Actors in Multi-person Videos	3043
<i>Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Hajja, Alexander Gorban, Kevin Murphy, and Li Fei-Fei</i>	
Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition	3054
<i>Behrooz Mahasseni and Sinisa Todorovic</i>	
Personalizing Human Video Pose Estimation	3063
<i>James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman</i>	
End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation	3073
<i>Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang</i>	

S3-1A: Activity Recognition

Actor-Action Semantic Segmentation with Grouping Process Models	3083
<i>Chenliang Xu and Jason J. Corso</i>	
Temporal Action Localization with Pyramid of Score Distribution Features	3093
<i>Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf A. Kassim</i>	
Recognizing Activities of Daily Living with a Wrist-Mounted Camera	3103
<i>Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, and Tatsuya Harada</i>	
Harnessing Object and Scene Semantics for Large-Scale Video Understanding	3112
<i>Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, and Leonid Sigal</i>	
Video-Story Composition via Plot Analysis	3122
<i>Jinsoo Choi, Tae-Hyun Oh, and In So Kweon</i>	
Temporal Action Detection Using a Statistical Language Model	3131
<i>Alexander Richard and Juergen Gall</i>	

Oral & Spotlight Session 3-1B

O3-1B: Semantic Segmentation

Multi-scale Patch Aggregation (MPA) for Simultaneous Detection and Segmentation	3141
<i>Shu Liu, Xiaojuan Qi, Jianping Shi, Hong Zhang, and Jiaya Jia</i>	
Instance-Aware Semantic Segmentation via Multi-task Network Cascades	3150
<i>Jifeng Dai, Kaiming He, and Jian Sun</i>	
ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation	3159
<i>Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun</i>	
Feature Space Optimization for Semantic Video Segmentation	3168
<i>Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun</i>	
Large-Scale Semantic 3D Reconstruction: An Adaptive Multi-resolution Model for Multi-class Volumetric Labeling	3176
<i>Maroš Bláha, Christoph Vogel, Audrey Richard, Jan D. Wegner, Thomas Pock, and Konrad Schindler</i>	

S3-1B: Semantic Parsing and Segmentation

Semantic Object Parsing with Local-Global Long Short-Term Memory	3185
<i>Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan</i>	
Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation	3194
<i>Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian Reid</i>	
Learning Transferrable Knowledge for Semantic Segmentation with Deep Convolutional Neural Network	3204
<i>Seunghoon Hong, Junhyuk Oh, Honglak Lee, and Bohyung Han</i>	
The Cityscapes Dataset for Semantic Urban Scene Understanding	3213
<i>Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele</i>	
Gaussian Conditional Random Field Network for Semantic Segmentation	3224
<i>Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, and Rama Chellappa</i>	
The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes	3234
<i>German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez</i>	

Poster Session P3-1

Progressive Prioritized Multi-view Stereo	3244
<i>Alex Locher, Michal Perdoch, and Luc Van Gool</i>	
WarpNet: Weakly Supervised Matching for Single-View Reconstruction	3253
<i>Angjoo Kanazawa, David W. Jacobs, and Manmohan Chandraker</i>	
What Sparse Light Field Coding Reveals about Scene Structure	3262
<i>Ole Johannsen, Antonin Sulc, and Bastian Goldluecke</i>	

Online Reconstruction of Indoor Scenes from RGB-D Streams	3271
<i>Hao Wang, Jun Wang, and Liang Wang</i>	
Patches, Planes and Probabilities: A Non-Local Prior for Volumetric 3D Reconstruction	3280
<i>Ali Osman Ulusoy, Michael J. Black, and Andreas Geiger</i>	
Single Image Camera Calibration with Lenticular Arrays for Augmented Reality	3290
<i>Ian Schillebeeckx and Robert Pless</i>	
Augmented Blendshapes for Real-Time Simultaneous 3D Head Modeling and Facial Motion Capture	3299
<i>Diego Thomas and Rin-Ichiro Taniguchi</i>	
Learned Binary Spectral Shape Descriptor for 3D Shape Correspondence	3309
<i>Jin Xie, Meng Wang, and Yi Fang</i>	
Multiple Models Fitting as a Set Coverage Problem	3318
<i>Luca Magri and Andrea Fusiello</i>	
Piecewise-Planar 3D Approximation from Wide-Baseline Stereo	3327
<i>C. Verleysen and C. De Vleeschouwer</i>	
Sparse to Dense 3D Reconstruction from Rolling Shutter Images	3337
<i>Olivier Saurer, Marc Pollefeys, and Gim Hee Lee</i>	
Consistency of Silhouettes and Their Duals	3346
<i>Matthew Trager, Martial Hebert, and Jean Ponce</i>	
Rolling Shutter Absolute Pose Problem with Known Vertical Direction	3355
<i>Cenek Albl, Zuzana Kukelova, and Tomas Pajdla</i>	
Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image	3364
<i>Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother</i>	
Multicamera Calibration from Visible and Mirrored Epipoles	3373
<i>Andrey Bushnevskiy, Lorenzo Sorgi, and Bodo Rosenhahn</i>	
Joint Unsupervised Deformable Spatio-Temporal Alignment of Sequences	3382
<i>Lazaros Zafeiriou, Epameinondas Antonakos, Stefanos Zafeiriou, and Maja Pantic</i>	
Deep Region and Multi-label Learning for Facial Action Unit Detection	3391
<i>Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang</i>	
Constrained Joint Cascade Regression Framework for Simultaneous Facial Action Unit Recognition and Facial Landmark Detection	3400
<i>Yue Wu and Qiang Ji</i>	
Unconstrained Face Alignment via Cascaded Compositional Learning	3409
<i>Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang</i>	
Automated 3D Face Reconstruction from Multiple Images Using Quality Measures	3418
<i>Marcel Pietraschke and Volker Blanz</i>	
Occlusion-Free Face Alignment: Deep Regression Networks Coupled with De-Corrupt AutoEncoders	3428
<i>Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen</i>	

Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis	3438
<i>Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin</i>	
Learning Reconstruction-Based Remote Gaze Estimation	3447
<i>Pei Yu, Jiahuan Zhou, and Ying Wu</i>	
Joint Training of Cascaded CNN for Face Detection	3456
<i>Hongwei Qin, Junjie Yan, Xiu Li, and Xiaolin Hu</i>	
Facial Expression Intensity Estimation Using Ordinal Information	3466
<i>Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji</i>	
Proposal Flow	3475
<i>Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce</i>	
ProNet: Learning to Propose Object-Specific Boxes for Cascaded Neural Networks	3485
<i>Chen Sun, Manohar Paluri, Ronan Collobert, Ram Nevatia, and Lubomir Bourdev</i>	
Seeing Behind the Camera: Identifying the Authorship of a Photograph	3494
<i>Christopher Thomas and Adriana Kovashka</i>	
Material Classification Using Raw Time-of-Flight Measurements	3503
<i>Shuochen Su, Felix Heide, Robin Swanson, Jonathan Klein, Clara Callenberg, Matthias Hullin, and Wolfgang Heidrich</i>	
Weakly Supervised Object Localization with Progressive Domain Adaptation	3512
<i>Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang</i>	
Newtonian Image Understanding: Unfolding the Dynamics of Objects in Static Images	3521
<i>Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi</i>	
Identifying Good Training Data for Self-Supervised Free Space Estimation	3530
<i>Ali Harakeh, Daniel Asmar, and Elie Shammas</i>	
Learning to Match Aerial Images with Deep Attentive Architectures	3539
<i>Hani Altwaijry, Eduard Trulls, James Hays, Pascal Fua, and Serge Belongie</i>	
Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection	3548
<i>Krishna Kumar Singh, Fanyi Xiao, and Yong Jae Lee</i>	
DeepCAMP: Deep Convolutional Action & Attribute Mid-Level Patterns	3557
<i>Ali Diba, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool</i>	
Canny Text Detector: Fast and Robust Scene Text Localization Algorithm	3566
<i>Hojin Cho, Myungchul Sung, and Bongjin Jun</i>	
Temporal Multimodal Learning in Audiovisual Speech Recognition	3574
<i>Di Hu, Xuelong Li, and Xiaoqiang Lu</i>	
Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd	3583
<i>Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim</i>	
Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs	3593
<i>Liuhaog Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann</i>	

Semantic Segmentation with Boundary Neural Fields	3602
<i>Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani</i>	
HD Maps: Fine-Grained Road Segmentation by Parsing Ground and Aerial Images	3611
<i>Gellért Mátyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun</i>	
DAG-Recurrent Neural Networks for Scene Labeling	3620
<i>Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang</i>	
Saliency Guided Dictionary Learning for Weakly-Supervised Image Parsing	3630
<i>Baisheng Lai and Xiaojin Gong</i>	
Attention to Scale: Scale-Aware Semantic Image Segmentation	3640
<i>Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille</i>	
Scene Labeling Using Sparse Precision Matrix	3650
<i>Nasim Souly and Mubarak Shah</i>	
Iterative Instance Segmentation	3659
<i>Ke Li, Bharath Hariharan, and Jitendra Malik</i>	
Recurrent Attentional Networks for Saliency Detection	3668
<i>Jason Kuen, Zhenhua Wang, and Gang Wang</i>	
Instance-Level Video Segmentation from Object Tracks	3678
<i>Guillaume Seguin, Piotr Bojanowski, Rémi Lajugie, and Ivan Laptev</i>	
Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer	3688
<i>Jun Xie, Martin Kiefel, Ming-Ting Sun, and Andreas Geiger</i>	
Amplitude Modulated Video Camera - Light Separation in Dynamic Scenes	3698
<i>Amir Kolaman, Maxim Lvov, Rami Hagege, and Hugo Guterman</i>	
A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo	3707
<i>Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan</i>	
Depth from Semi-Calibrated Stereo and Defocus	3717
<i>Ting-Chun Wang, Manohar Srikanth, and Ravi Ramamoorthi</i>	
Exploiting Spectral-Spatial Correlation for Coded Hyperspectral Image Restoration	3727
<i>Ying Fu, Yinqiang Zheng, Imari Sato, and Yoichi Sato</i>	
Variable Aperture Light Field Photography: Overcoming the Diffraction-Limited Spatio-Angular Resolution Tradeoff	3737
<i>Julie Chang, Isaac Kauvar, Xuemei Hu, and Gordon Wetzstein</i>	
Convolutional Networks for Shape from Light Field	3746
<i>Stefan Heber and Thomas Pock</i>	
Panoramic Stereo Videos with a Single Camera	3755
<i>Rajat Aggarwal, Amrisha Vohra, and Anoop M. Namboodiri</i>	
The Next Best Underwater View	3764
<i>Mark Sheinin and Yoav Y. Schechner</i>	
Reconstructing Shapes and Appearances of Thin Film Objects Using RGB Images	3774
<i>Yoshie Kobayashi, Tetsuro Morimoto, Imari Sato, Yasuhiro Mukaigawa, Takao Tomono, and Katsushi Ikeuchi</i>	

Noisy Label Recovery for Shadow Detection in Unfamiliar Domains	3783
<i>Tomás F. Yago Vicente, Minh Hoai, and Dimitris Samaras</i>	

Oral & Spotlight Session 3-2A

O3-2A: Video Understanding

Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled	3793
<i>Oscar Koller, Hermann Ney, and Richard Bowden</i>	
Recognizing Car Fluents from Video	3803
<i>Bo Li, Tianfu Wu, Caiming Xiong, and Song-Chun Zhu</i>	
Pairwise Decomposition of Image Sequences for Active Multi-view Recognition	3813
<i>Edward Johns, Stefan Leutenegger, and Andrew J. Davison</i>	
Inferring Forces and Learning Human Utilities from Videos	3823
<i>Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu</i>	
Force from Motion: Decoding Physical Sensation in a First Person Video	3834
<i>Hyun Soo Park, Jyh-Jing Hwang, and Jianbo Shi</i>	

S3-2A: Video Analysis 2

Robust Multi-Body Feature Tracker: A Segmentation-Free Approach	3843
<i>Pan Ji, Hongdong Li, Mathieu Salzmann, and Yiran Zhong</i>	
Slow and Steady Feature Analysis: Higher Order Temporal Coherence in Video	3852
<i>Dinesh Jayaraman and Kristen Grauman</i>	
Volumetric 3D Tracking by Detection	3862
<i>Chun-Hao Huang, Benjamin Allain, Jean-Sébastien Franco, Nassir Navab, Slobodan Ilic, and Edmond Boyer</i>	
The Solution Path Algorithm for Identity-Aware Multi-object Tracking	3871
<i>Shouou-I Yu, Deyu Meng, Wangmeng Zuo, and Alexander Hauptmann</i>	
In Defense of Sparse Tracking: Circulant Sparse Tracker	3880
<i>Tianzhu Zhang, Adel Bibi, and Bernard Ghanem</i>	
Optical Flow with Semantic Segmentation and Localized Layers	3889
<i>Laura Sevilla-Lara, Deqing Sun, Varun Jampani, and Michael J. Black</i>	
Video Segmentation via Object Flow	3899
<i>Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J. Black</i>	

Oral & Spotlight Session 3-2B

O3-2B: Grouping and Optimization Methods

Closed-Form Training of Mahalanobis Distance for Supervised Clustering	3909
<i>Marc T. Law, Yaoliang Yu, Matthieu Cord, and Eric P. Xing</i>	

Scalable Sparse Subspace Clustering by Orthogonal Matching Pursuit	3918
<i>Chong You, Daniel P. Robinson, and René Vidal</i>	
Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering	3928
<i>Chong You, Chun-Guang Li, Daniel P. Robinson, and René Vidal</i>	
Sparse Coding and Dictionary Learning with Linear Dynamical Systems	3938
<i>Wenbing Huang, Fuchun Sun, Lele Cao, Deli Zhao, Huaping Liu, and Mehrtash Harandi</i>	
Sublabel–Accurate Relaxation of Nonconvex Energies	3948
<i>Thomas Möllenhoff, Emanuel Laude, Michael Moeller, Jan Lellmann, and Daniel Cremers</i>	

S3-2B: Statistical Methods and Transfer Learning

The Multiverse Loss for Robust Transfer Learning	3957
<i>Etai Littwin and Lior Wolf</i>	
Learning from the Mistakes of Others: Matching Errors in Cross-Dataset Learning	3967
<i>Viktoriia Sharmanska and Novi Quadrianto</i>	
An Efficient Exact-PGA Algorithm for Constant Curvature Manifolds	3976
<i>Rudrasis Chakraborty, Dohyung Seo, and Baba C. Vemuri</i>	
Online Learning with Bayesian Classification Trees	3985
<i>Samuel Rota Bulò and Peter Kotschieder</i>	
Cross-Stitch Networks for Multi-task Learning	3994
<i>Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert</i>	
Deep Metric Learning via Lifted Structured Feature Embedding	4004
<i>Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese</i>	
Fast Algorithms for Convolutional Neural Networks	4013
<i>Andrew Lavin and Scott Gray</i>	

Poster Session P3-2

Coordinating Multiple Disparity Proposals for Stereo Computation	4022
<i>Ang Li, Dapeng Chen, Yuanliu Liu, and Zejian Yuan</i>	
Joint Multiview Segmentation and Localization of RGB-D Images Using Depth-Induced Silhouette Consistency	4031
<i>Chi Zhang, Zhiwei Li, Rui Cai, Hongyang Chao, and Yong Rui</i>	
A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation	4040
<i>Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox</i>	
6D Dynamic Camera Relocalization from Single Reference Image	4049
<i>Wei Feng, Fei-Peng Tian, Qian Zhang, and Jizhou Sun</i>	
Dense Monocular Depth Estimation in Complex Dynamic Scenes	4058
<i>René Ranftl, Vibhav Vineet, Qifeng Chen, and Vladlen Koltun</i>	
Using Self-Contradiction to Learn Confidence Measures in Stereo Vision	4067
<i>Christian Mostegel, Markus Rumpel, Friedrich Fraundorfer, and Horst Bischof</i>	

Understanding RealWorld Indoor Scenes with Synthetic Data	4077
<i>Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla</i>	
Stereo Matching with Color and Monochrome Cameras in Low-Light Conditions	4086
<i>Hae-Gon Jeon, Joon-Young Lee, Sunghoon Im, Hyowon Ha, and In So Kweon</i>	
Camera Calibration from Dynamic Silhouettes Using Motion Barcodes	4095
<i>Gil Ben-Artzi, Yoni Kasten, Shmuel Peleg, and Michael Werman</i>	
Structure-from-Motion Revisited	4104
<i>Johannes L. Schönberger and Jan-Michael Frahm</i>	
Constructing Canonical Regions for Fast and Effective View Selection	4114
<i>Wencheng Wang and Tianhao Gao</i>	
Prior-Less Compressible Structure from Motion	4123
<i>Chen Kong and Simon Lucey</i>	
Rolling Shutter Camera Relative Pose: Generalized Epipolar Geometry	4132
<i>Yuchao Dai, Hongdong Li, and Laurent Kneip</i>	
Structure from Motion with Objects	4141
<i>Marco Crocco, Cosimo Rubino, and Alessio Del Bue</i>	
DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features	4150
<i>Ayan Sinha, Chiho Choi, and Karthik Ramani</i>	
Multi-oriented Text Detection with Fully Convolutional Networks	4159
<i>Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai</i>	
Robust Scene Text Recognition with Automatic Rectification	4168
<i>Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai</i>	
Mnemonic Descent Method: A Recurrent Process Applied for End-to-End Face Alignment	4177
<i>George Trigeorgis, Patrick Snape, Mihalís A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou</i>	
Large-Pose Face Alignment via CNN-Based Dense 3D Model Fitting	4188
<i>Amin Jourabloo and Xiaoming Liu</i>	
Adaptive 3D Face Reconstruction from Unconstrained Photo Collections	4197
<i>Joseph Roth, Yiyi Tong, and Xiaoming Liu</i>	
Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks	4207
<i>Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz</i>	
Kinematic Structure Correspondences via Hypergraph Matching	4216
<i>Hyung Jin Chang, Tobias Fischer, Maxime Petit, Martina Zambelli, and Yiannis Demiris</i>	
CP-mtML: Coupled Projection Multi-Task Metric Learning for Large Scale Face Retrieval	4226
<i>Binod Bhattarai, Gaurav Sharma, and Frederic Jurie</i>	

PatchBatch: A Batch Augmented Loss for Optical Flow	4236
<i>David (Dedi) Gadot and Lior Wolf</i>	
Joint Recovery of Dense Correspondence and Cosegmentation in Two Images	4246
<i>Tatsunori Taniai, Sudipta N. Sinha, and Yoichi Sato</i>	
Multi-view People Tracking via Hierarchical Trajectory Composition	4256
<i>Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu</i>	
Object Tracking via Dual Linear Structured SVM and Explicit Feature Map	4266
<i>Jifeng Ning, Jimei Yang, Shaojie Jiang, Lei Zhang, and Ming-Hsuan Yang</i>	
Robust, Real-Time 3D Tracking of Multiple Objects with Similar Appearances	4275
<i>Taiki Sekii</i>	
An Egocentric Look at Video Photographer Identity	4284
<i>Yedid Hoshen and Shmuel Peleg</i>	
Learning Multi-domain Convolutional Neural Networks for Visual Tracking	4293
<i>Hyeonseob Nam and Bohyung Han</i>	
Hedged Deep Tracking	4303
<i>Yuankai Qi, Shengping Zhang, Lei Qin, Hongxun Yao, Qingming Huang, Jongwoo Lim, and Ming-Hsuan Yang</i>	
Structural Correlation Filter for Robust Visual Tracking	4312
<i>Si Liu, Tianzhu Zhang, Xiaochun Cao, and Changsheng Xu</i>	
Visual Tracking Using Attention-Modulated Disintegration and Integration	4321
<i>Jongwon Choi, Hyung Jin Chang, Jiyeoup Jeong, Yiannis Demiris, and Jin Young Choi</i>	
A Continuous Occlusion Model for Road Scene Understanding	4331
<i>Vikas Dhiman, Quoc-Huy Tran, Jason J. Corso, and Manmohan Chandraker</i>	
VirtualWorlds as Proxy for Multi-object Tracking Analysis	4340
<i>Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig</i>	
Uncalibrated Photometric Stereo by Stepwise Optimization Using Principal Components of Isotropic BRDFs	4350
<i>Keisuke Midorikawa, Toshihiko Yamasaki, and Kiyoharu Aizawa</i>	
Unbiased Photometric Stereo for Colored Surfaces: A Variational Approach	4359
<i>Yvain Quéau, Roberto Mecca, and Jean-Denis Durou</i>	
3D Reconstruction of Transparent Objects with Position-Normal Consistency	4369
<i>Yiming Qian, Minglun Gong, and Yee-Hong Yang</i>	
Real-Time Depth Refinement for Specular Objects	4378
<i>Roy Or-El, Rom Hershkovitz, Aaron Wetzler, Guy Rosman, Alfred M. Bruckstein, and Ron Kimmel</i>	
Recovering Transparent Shape from Time-of-Flight Distortion	4387
<i>Kenichiro Tanaka, Yasuhiro Mukaigawa, Hiroyuki Kubo, Yasuyuki Matsushita, and Yasushi Yagi</i>	
Robust Light Field Depth Estimation for Noisy Scene with Occlusion	4396
<i>Williem and In Kyu Park</i>	

Rotational Crossed-Slit Light Fields	4405
<i>Nianyi Li, Haiting Lin, Bilin Sun, Mingyuan Zhou, and Jingyi Yu</i>	
Single Image Object Modeling Based on BRDF and r-Surfaces Learning	4414
<i>Fabrizio Natola, Valsamis Ntouskos, Fiora Pirri, and Marta Sanzari</i>	
A Nonlinear Regression Technique for Manifold Valued Data with Applications to Medical Image Analysis	4424
<i>Monami Banerjee, Rudrasis Chakraborty, Edward Ofori, Michael S. Okun, David E. Vaillancourt, and Baba C. Vemuri</i>	
RAID-G: Robust Estimation of Approximate Infinite Dimensional Gaussian with Application to Material Recognition	4433
<i>Qilong Wang, Peihua Li, Wangmeng Zuo, and Lei Zhang</i>	
An Empirical Evaluation of Current Convolutional Architectures' Ability to Manage Nuisance Location and Scale Variability	4442
<i>Nikolaos Karianakis, Jingming Dong, and Stefano Soatto</i>	
Learning Sparse High Dimensional Filters: Image Filtering, Dense CRFs and Bilateral Neural Networks	4452
<i>Varun Jampani, Martin Kiefel, and Peter V. Gehler</i>	
Mixture of Bilateral-Projection Two-Dimensional Probabilistic Principal Component Analysis	4462
<i>Fujiao Ju, Yanfeng Sun, Junbin Gao, Simeng Liu, Yongli Hu, and Baocai Yin</i>	
Rolling Rotations for Recognizing Human Actions from 3D Skeletal Data	4471
<i>Raviteja Vemulapalli and Rama Chellappa</i>	
Improving the Robustness of Deep Neural Networks via Stability Training	4480
<i>Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow</i>	
Logistic Boosting Regression for Label Distribution Learning	4489
<i>Chao Xing, Xin Geng, and Hui Xue</i>	
Efficient Temporal Sequence Comparison and Classification Using Gram Matrix Embeddings on a Riemannian Manifold	4498
<i>Xikang Zhang, Yin Wang, Mengran Gou, Mario Sznajder, and Octavia Camps</i>	
Deep Reflectance Maps	4508
<i>Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars</i>	
Semantic Filtering	4517
<i>Qingxiang Yang</i>	
UAVSensor Fusion with Latent-Dynamic Conditional Random Fields in Coronal Plane Estimation	4527
<i>Amir M. Rahimi, Raphael Ruschell, and B. S. Manjunath</i>	
Robust Visual Place Recognition with Graph Kernels	4535
<i>Elena Stumm, Christopher Mei, Simon Lacroix, Juan Nieto, Marco Hutter, and Roland Siegwart</i>	

Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform	4545
<i>Liang-Chieh Chen, Jonathan T. Barron, George Papandreou, Kevin Murphy, and Alan L. Yuille</i>	

Oral & Spotlight Session 4-1A

O4-1A: Image & Video Captioning and Descriptions

Natural Language Object Retrieval	4555
<i>Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell</i>	
DenseCap: Fully Convolutional Localization Networks for Dense Captioning	4565
<i>Justin Johnson, Andrej Karpathy, and Li Fei-Fei</i>	
Unsupervised Learning from Narrated Instruction Videos	4575
<i>Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, and Josef Sivic</i>	
Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks	4584
<i>Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu</i>	
Jointly Modeling Embedding and Translation to Bridge Video and Language	4594
<i>Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui</i>	

S4-1A: High Level Semantics

We are Humor Beings: Understanding and Predicting Visual Humor	4603
<i>Arjun Chandrasekaran, Ashwin K. Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh</i>	
Where to Look: Focus Regions for Visual Question Answering	4613
<i>Kevin J. Shih, Saurabh Singh, and Derek Hoiem</i>	
Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources	4622
<i>Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel</i>	
MovieQA: Understanding Stories in Movies through Question-Answering	4631
<i>Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler</i>	
TGIF: A New Dataset and Benchmark on Animated GIF Description	4641
<i>Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo</i>	
Image Captioning with Semantic Attention	4651
<i>Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo</i>	

Oral & Spotlight Session 4-1B

O4-1B: Non-rigid Reconstruction and Motion Analysis

Temporally Coherent 4D Reconstruction of Complex Dynamic Scenes	4660
<i>Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton</i>	
Consensus of Non-rigid Reconstructions	4670
<i>Minsik Lee, Jungchan Cho, and Songhwai Oh</i>	
Isometric Non-rigid Shape-from-Motion in Linear Time	4679
<i>Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli</i>	
Learning Online Smooth Predictors for Realtime Camera Planning Using Recurrent Decision Trees	4688
<i>Jianhui Chen, Hoang M. Le, Peter Carr, Yisong Yue, and James J. Little</i>	
Egocentric Future Localization	4697
<i>Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi</i>	
Full Flow: Optical Flow Estimation By Global Optimization over Regular Grids	4706
<i>Qifeng Chen and Vladlen Koltun</i>	

S4-1B: Human Pose Estimation

Structured Feature Learning for Pose Estimation	4715
<i>Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang</i>	
Convolutional Pose Machines	4724
<i>Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh</i>	
Human Pose Estimation with Iterative Error Feedback	4733
<i>João Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik</i>	

Poster Session P4-1

WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks	4743
<i>Thibaut Durand, Nicolas Thome, and Matthieu Cord</i>	
DisturbLabel: Regularizing CNN on the Loss Layer	4753
<i>Lingxi Xie, Jingdong Wang, Zhen Wei, Meng Wang, and Qi Tian</i>	
Gradual DropIn of Layers to Train Very Deep Neural Networks	4763
<i>Leslie N. Smith, Emily M. Hand, and Timothy Doster</i>	
Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition	4772
<i>Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori</i>	
Deep SimNets	4782
<i>Nadav Cohen, Or Sharir, and Amnon Shashua</i>	
Studying Very Low Resolution Recognition Using Deep Networks	4792
<i>Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S. Huang</i>	
Deep Gaussian Conditional Random Field Network: A Model-Based Deep Network for Discriminative Denoising	4801
<i>Raviteja Vemulapalli, Oncel Tuzel, and Ming-Yu Liu</i>	

Event-Specific Image Importance	4810
<i>Yufei Wang, Zhe Lin, Xiaohui Shen, Radomír Mech, Gavin Miller, and Garrison W. Cottrell</i>	
Quantized Convolutional Neural Networks for Mobile Devices	4820
<i>Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng</i>	
Inverting Visual Representations with Convolutional Networks	4829
<i>Alexey Dosovitskiy and Thomas Brox</i>	
Pose-Aware Face Recognition in the Wild	4838
<i>Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan</i>	
Multi-view Deep Network for Cross-View Classification	4847
<i>Meina Kan, Shiguang Shan, and Xilin Chen</i>	
Sparsifying Neural Network Connections for Face Recognition	4856
<i>Yi Sun, Xiaogang Wang, and Xiaoou Tang</i>	
Pairwise Linear Regression Classification for Image Set Retrieval	4865
<i>Qingxiang Feng, Yicong Zhou, and Rushi Lan</i>	
The MegaFace Benchmark: 1 Million Faces for Recognition at Scale	4873
<i>Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard</i>	
Learnt Quasi-Transitive Similarity for Retrieval from Large Collections of Faces	4883
<i>Ognjen Arandjelovic</i>	
Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition	4893
<i>Yandong Wen, Zhifeng Li, and Yu Qiao</i>	
Copula Ordinal Regression for Joint Estimation of Facial Action Unit Intensity	4902
<i>Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic</i>	
A Robust Multilinear Model Learning Framework for 3D Faces	4911
<i>Timo Bolkart and Stefanie Wuhrer</i>	
Ordinal Regression with Multiple Output CNN for Age Estimation	4920
<i>Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua</i>	
DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation	4929
<i>Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele</i>	
Thin-Slicing for Pose: Learning to Understand Pose without Explicit Pose Estimation	4938
<i>Suha Kwak, Minsu Cho, and Ivan Laptev</i>	
A Dual-Source Approach for 3D Pose Estimation from a Single Image	4948
<i>Hashim Yasin, Umar Iqbal, Björn Krüger, Andreas Weber, and Juergen Gall</i>	
Efficiently Creating 3D Training Data for Fine Hand Pose Estimation	4957
<i>Markus Oberweger, Gernot Riegler, Paul Wohlhart, and Vincent Lepetit</i>	
Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video	4966
<i>Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis</i>	
Answer-Type Prediction for Visual Question Answering	4976
<i>Kushal Kafle and Christopher Kanan</i>	

VisualWord2Vec (Vis-W2V): Learning Visually Grounded Word Embeddings Using Abstract Scenes	4985
<i>Satwik Kottur, Ramakrishna Vedantam, José M. F. Moura, and Devi Parikh</i>	
Visual7W: Grounded Question Answering in Images	4995
<i>Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei</i>	
Learning Deep Structure-Preserving Image-Text Embeddings	5005
<i>Liwei Wang, Yin Li, and Svetlana Lazebnik</i>	
Yin and Yang: Balancing and Answering Binary Visual Questions	5014
<i>Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh</i>	
GIFT: A Real-Time and Scalable 3D Shape Search Engine	5023
<i>Song Bai, Xiang Bai, Zhichao Zhou, Zhaoxiang Zhang, and Longin Jan Latecki</i>	
Functional Faces: Groupwise Dense Correspondence Using Functional Maps	5033
<i>Chao Zhang, William A. P. Smith, Arnaud Dessein, Nick Pears, and Hang Dai</i>	
Similarity Metric for Curved Shapes in Euclidean Space	5042
<i>Girum G. Demisse, Djamila Aouada, and Björn Ottersten</i>	
Shape Analysis with Hyperbolic Wasserstein Distance	5051
<i>Jie Shi, Wen Zhang, and Yalin Wang</i>	
Tensor Power Iteration for Multi-graph Matching	5062
<i>Xinchu Shi, Haibin Ling, Weiming Hu, Junliang Xing, and Yanning Zhang</i>	
Multivariate Regression on the Grassmannian for Predicting Novel Domains	5071
<i>Yongxin Yang and Timothy M. Hospedales</i>	
Learning Cross-Domain Landmarks for Heterogeneous Domain Adaptation	5081
<i>Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang</i>	
Geospatial Correspondences for Multimodal Registration	5091
<i>Diego Marcos, Raffay Hamid, and Devis Tuia</i>	
Constrained Deep Transfer Feature Learning and Its Applications	5101
<i>Yue Wu and Qiang Ji</i>	
Deep Canonical Time Warping	5110
<i>George Trigeorgis, Mihalis A. Nicolaou, Stefanos Zafeiriou, and Björn W. Schuller</i>	
Multilinear Hyperplane Hashing	5119
<i>Xianglong Liu, Xinjie Fan, Cheng Deng, Zhujin Li, Hao Su, and Dacheng Tao</i>	
Large Scale Hard Sample Mining with Monte Carlo Tree Search	5128
<i>Olivier Canévet and François Fleuret</i>	
Multi-label Ranking from Positive and Unlabeled Data	5138
<i>Atsushi Kanehira and Tatsuya Harada</i>	
Joint Unsupervised Learning of Deep Representations and Image Clusters	5147
<i>Jianwei Yang, Devi Parikh, and Dhruv Batra</i>	
Kernel Sparse Subspace Clustering on Symmetric Positive Definite Manifolds	5157
<i>Ming Yin, Yi Guo, Junbin Gao, Zhaoshui He, and Shengli Xie</i>	

Symmetry reCAPTCHA	5165
<i>Christopher Funk and Yanxi Liu</i>	
Unsupervised Learning of Discriminative Attributes and Visual Representations	5175
<i>Chen Huang, Chen Change Loy, and Xiaoou Tang</i>	
When VLAD Met Hilbert	5185
<i>Mehrtash Harandi, Mathieu Salzmann, and Fatih Porikli</i>	
Approximate Log-Hilbert-Schmidt Distances between Covariance Operators for Image Classification	5195
<i>Hà Quang Minh, Marco San Biagio, Loris Bazzani, and Vittorio Murino</i>	
Subspace Clustering with Priors via Sparse Quadratically Constrained Quadratic Programming	5204
<i>Yongfang Cheng, Yin Wang, Mario Sznaier, and Octavia Camps</i>	
Robust Tensor Factorization with Unknown Noise	5213
<i>Xi'ai Chen, Zhi Han, Yao Wang, Qian Zhao, Deyu Meng, and Yandong Tang</i>	
Kernel Approximation via Empirical Orthogonal Decomposition for Unsupervised Feature Learning	5222
<i>Yusuke Mukuta and Tatsuya Harada</i>	
Active Learning for Delineation of Curvilinear Structures	5231
<i>Agata Mosinska-Domanska, Raphael Sznitman, Przemyslaw Glowacki, and Pascal Fua</i>	
Recognizing Emotions from Abstract Paintings Using Non-Linear Matrix Completion	5240
<i>Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe</i>	
Tensor Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Tensors via Convex Optimization	5249
<i>Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan</i>	
Sliced Wasserstein Kernels for Probability Distributions	5258
<i>Soheil Kolouri, Yang Zou, and Gustavo K. Rohde</i>	
Trace Quotient Meets Sparsity: A Method for Learning Low Dimensional Image Representations	5268
<i>Xian Wei, Hao Shen, and Martin Kleinstenber</i>	
Backtracking ScSPM Image Classifier for Weakly Supervised Top-Down Saliency	5278
<i>Hisham Cholakkal, Jubin Johnson, and Deepu Rajan</i>	
MSR-VTT: A Large Video Description Dataset for Bridging Video and Language	5288
<i>Jun Xu, Tao Mei, Ting Yao, and Yong Rui</i>	

Oral & Spotlight Session 4-2A

04-2A: Learning and CNN Architectures

NetVLAD: CNN Architecture for Weakly Supervised Place Recognition	5297
<i>Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic</i>	
Structural-RNN: Deep Learning on Spatio-Temporal Graphs	5308
<i>Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena</i>	

Learning to Select Pre-Trained Deep Representations with Bayesian Evidence Framework	5318
<i>Yong-Deok Kim, Taewoong Jang, Bohyung Han, and Seungjin Choi</i>	
Synthesized Classifiers for Zero-Shot Learning	5327
<i>Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha</i>	
Semi-supervised Vocabulary-Informed Learning	5337
<i>Yanwei Fu and Leonid Sigal</i>	

S4-2A: Learning and Optimization

Simultaneous Clustering and Model Selection for Tensor Affinities	5347
<i>Zhuwen Li, Shuoguang Yang, Loong-Fah Cheong, and Kim-Chuan Toh</i>	
Discriminatively Embedded K-Means for Multi-view Clustering	5356
<i>Jinglin Xu, Junwei Han, and Feiping Nie</i>	
Min Norm Point Algorithm for Higher Order MRF-MAP Inference	5365
<i>Ishant Shanu, Chetan Arora, and Parag Singla</i>	
Learning Deep Representation for Imbalanced Classification	5375
<i>Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang</i>	
Learning Local Image Descriptors with Deep Siamese and Triplet Convolutional Networks by Minimizing Global Loss Functions	5385
<i>Vijay Kumar B G, Gustavo Carneiro, and Ian Reid</i>	
Sparse Coding for Third-Order Super-Symmetric Tensor Descriptors with Application to Texture Recognition	5395
<i>Piotr Koniusz and Anoop Cherian</i>	
Random Features for Sparse Signal Classification	5404
<i>Jen-Hao Rick Chang, Aswin C. Sankaranarayanan, and B. V. K. Vijaya Kumar</i>	

Oral & Spotlight Session 4-2B

O4-2B: 3D Shape Reconstruction

High-Quality Depth from Uncalibrated Small Motion Clip	5413
<i>Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, and In So Kweon</i>	
Efficient 3D Room Shape Recovery from a Single Panorama	5422
<i>Hao Yang and Hui Zhang</i>	
Structured Prediction of Unobserved Voxels from a Single Depth Image	5431
<i>Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J. Brostow</i>	
HyperDepth: Learning Depth from Structured Light without Matching	5441
<i>Sean Ryan Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, and Shahram Izadi</i>	
SVBRDF-Invariant Shape and Reflectance Estimation from Light-Field Cameras	5451
<i>Ting-Chun Wang, Manmohan Chandraker, Alexei A. Efros, and Ravi Ramamoorthi</i>	

S4-2B: 3D Reconstruction

Semantic 3D Reconstruction with Continuous Regularization and Ray Potentials Using a Visibility Consistency Constraint	5460
<i>Nikolay Savinov, Christian Häne, L'ubor Ladický, and Marc Pollefeys</i>	
Theory and Practice of Structure-From-Motion Using Affine Correspondences	5470
<i>Carolina Raposo and João P. Barreto</i>	
Just Look at the Image: Viewpoint-Specific Surface Normal Prediction for Improved Multi-View Reconstruction	5479
<i>Silvano Galliani and Konrad Schindler</i>	
From Dusk Till Dawn: Modeling in the Dark	5488
<i>Filip Radenovic, Johannes L. Schönberger, Dinghuang Ji, Jan-Michael Frahm, Ondrej Chum, and Jiri Matas</i>	
Accelerated Generative Models for 3D Point Cloud Data	5497
<i>Ben Eckart, Kihwan Kim, Alejandro Troccoli, Alonzo Kelly, and Jan Kautz</i>	
Monocular Depth Estimation Using Neural Regression Forest	5506
<i>Anirban Roy and Sinisa Todorovic</i>	
Deep Stereo: Learning to Predict New Views from the World's Imagery	5515
<i>John Flynn, Ivan Neulander, James Philbin, and Noah Snavely</i>	

Oral & Spotlight Session 4-3A

O4-3A: Face, Gesture, & Situation Recognition: Algorithms and Datasets

WIDER FACE: A Face Detection Benchmark	5525
<i>Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang</i>	
Situation Recognition: Visual Semantic Role Labeling for Image Understanding	5534
<i>Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi</i>	

S4-3A: People and Faces

A 3D Morphable Model Learnt from 10,000 Faces	5543
<i>James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway</i>	
Some Like It Hot — Visual Guidance for Preference Prediction	5553
<i>Rasmus Rothe, Radu Timofte, and Luc Van Gool</i>	
EmotionNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild	5562
<i>C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez</i>	
ForgetMeNot: Memory-Aware Forensic Facial Sketch Matching	5571
<i>Shuxin Ouyang, Timothy M. Hospedales, Yi-Zhe Song, and Xueming Li</i>	
LOMo: Latent Ordinal Model for Facial Analysis in Videos	5580
<i>Karan Sikka, Gaurav Sharma, and Marian Bartlett</i>	

Discriminative Invariant Kernel Features: A Bells-and-Whistles-Free Approach to Unsupervised Face Recognition and Pose Estimation	5590
<i>Dipan K. Pal, Felix Juefei-Xu, and Marios Savvides</i>	
Bottom-Up and Top-Down Reasoning with Hierarchical Rectified Gaussians	5600
<i>Peiyun Hu and Deva Ramanan</i>	
Fits Like a Glove: Rapid and Reliable Hand Shape Personalization	5610
<i>David Joseph Tan, Thomas Cashman, Jonathan Taylor, Andrew Fitzgibbon, Daniel Tarlow, Sameh Khamis, Shahram Izadi, and Jamie Shotton</i>	
Slicing Convolutional Neural Network for Crowd Video Understanding	5620
<i>Jing Shao, Chen Change Loy, Kai Kang, and Xiaogang Wang</i>	

Spotlight Session 4-3B

S4-3B: 3D, Stereo, Matching, and Saliency Estimation

Linear Shape Deformation Models with Local Support Using Graph-Based Structured Matrix Factorisation	5629
<i>Florian Bernard, Peter Gemmar, Frank Hertel, Jorge Goncalves, and Johan Thunberg</i>	
Motion from Structure (MfS): Searching for 3D Objects in Cluttered Point Trajectories	5639
<i>Jayakorn Vongkulbhisal, Ricardo Cabral, Fernando De la Torre, and João P. Costeira</i>	
Volumetric and Multi-view CNNs for Object Classification on 3D Data	5648
<i>Charles R. Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J. Guibas</i>	
Detecting Vanishing Points Using Global Image Context in a Non-ManhattanWorld	5657
<i>Menghua Zhai, Scott Workman, and Nathan Jacobs</i>	
Learning Weight Uncertainty with Stochastic Gradient MCMC for Shape Classification	5666
<i>Chunyuan Li, Andrew Stevens, Changyou Chen, Yunchen Pu, Zhe Gan, and Lawrence Carin</i>	
A Field Model for Repairing 3D Shapes	5676
<i>Duc Thanh Nguyen, Binh-Son Hua, Minh-Khoi Tran, Quang-Hieu Pham, and Sai-Kit Yeung</i>	
GOGMA: Globally-Optimal Gaussian Mixture Alignment	5685
<i>Dylan Campbell and Lars Petersson</i>	
Efficient Deep Learning for Stereo Matching	5695
<i>Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun</i>	
Efficient Coarse-to-Fine Patch Match for Large Displacement Optical Flow	5704
<i>Yinlin Hu, Rui Song, and Yunsong Li</i>	
FANNG: Fast Approximate Nearest Neighbour Graphs	5713
<i>Ben Harwood and Tom Drummond</i>	
Exemplar-Driven Top-Down Saliency Detection via Deep Association	5723
<i>Shengfeng He and Rynson W. H. Lau</i>	
Unconstrained Salient Object Detection via Proposal Subset Optimization	5733
<i>Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Mech</i>	

Recombinator Networks: Learning Coarse-to-Fine Feature Aggregation	5743
<i>Sina Honari, Jason Yosinski, Pascal Vincent, and Christopher Pal</i>	
End-to-End Saliency Mapping via Probability Distribution Prediction	5753
<i>Saumya Jetley, Naila Murray, and Eleonora Vig</i>	

Poster Session P4-2

A Paradigm for Building Generalized Models of Human Image Perception through Data Fusion	5762
<i>Shaojing Fan, Tian-Tsong Ng, Bryan L. Koenig, Ming Jiang, and Qi Zhao</i>	
Longitudinal Face Modeling via Temporal Deep Restricted Boltzmann Machines	5772
<i>Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui</i>	
Saliency Unified: A Deep Architecture for simultaneous Eye Fixation Prediction and Salient Object Segmentation	5781
<i>Srinivas S. S. Kruthiventi, Vennela Gudisa, Jaley H. Dholakiya, and R. Venkatesh Babu</i>	
Estimating Correspondences of Deformable Objects “In-the-Wild”	5791
<i>Yuxiang Zhou, Epameinondas Antonakos, Joan Alabort-I-Medina, Anastasios Roussos, and Stefanos Zafeiriou</i>	
Gravitational Approach for Point Set Registration	5802
<i>Vladislav Golyanik, Sk Aziz Ali, and Didier Stricker</i>	
Context-Aware Gaussian Fields for Non-rigid Point Set Registration	5811
<i>Gang Wang, Zhicheng Wang, Yufei Chen, Qiangqiang Zhou, and Weidong Zhao</i>	
Trust No One: Low Rank Matrix Factorization Using Hierarchical RANSAC	5820
<i>Magnus Oskarsson, Kenneth Batstone, and Kalle Åström</i>	
Relaxation-Based Preprocessing Techniques for Markov Random Field Inference	5830
<i>Chen Wang and Ramin Zabih</i>	
Sparse Coding for Classification via Discrimination Ensemble	5839
<i>Yuhui Quan, Yong Xu, Yuping Sun, Yan Huang, and Hui Ji</i>	
Principled Parallel Mean-Field Inference for Discrete Random Fields	5848
<i>Pierre Baqué, Timur Bagautdinov, François Fleuret, and Pascal Fua</i>	
Guaranteed Outlier Removal with Mixed Integer Linear Programs	5858
<i>Tat-Jun Chin, Yang Heng Kee, Anders Eriksson, and Frank Neumann</i>	
Memory Efficient Max Flow for Multi-label Submodular MRFs	5867
<i>Thalaiyasingam Ajanthan, Richard Hartley, and Mathieu Salzmann</i>	
Proximal Riemannian Pursuit for Large-Scale Trace-Norm Minimization	5877
<i>Mingkui Tan, Shijie Xiao, Junbin Gao, Dong Xu, Anton van den Hengel, and Qinfeng Shi</i>	
Minimizing the Maximal Rank	5887
<i>Erik Bylow, Carl Olsson, Fredrik Kahl, and Mikael Nilsson</i>	
Solving Temporal Puzzles	5896
<i>Caglayan Dicle, Burak Yilmaz, Octavia Camps, and Mario Szaier</i>	

Estimating Sparse Signals with Smooth Support via Convex Programming and Block Sparsity	5906
<i>Sohil Shah, Tom Goldstein, and Christoph Studer</i>	
TenSR: Multi-dimensional Tensor Sparse Representation	5916
<i>Na Qi, Yunhui Shi, Xiaoyan Sun, and Baocai Yin</i>	
Moral Lineage Tracing	5926
<i>Florian Jug, Evgeny Levinkov, Corinna Blasse, Eugene W. Myers, and Bjoern Andres</i>	
Globally Optimal Rigid Intensity Based Registration: A Fast Fourier Domain Approach	5936
<i>Behrooz Nasihatkon, Frida Fejne, and Fredrik Kahl</i>	
On Benefits of Selection Diversity via Bilevel Exclusive Sparsity	5945
<i>Haichuan Yang, Yijun Huang, Lam Tran, Ji Liu, and Shuai Huang</i>	
Fast Training of Triplet-Based Deep Binary Embedding Networks	5955
<i>Bohan Zhuang, Guosheng Lin, Chunhua Shen, and Ian Reid</i>	
Marr Revisited: 2D-3D Alignment via Surface Normal Prediction	5965
<i>Aayush Bansal, Bryan Russell, and Abhinav Gupta</i>	
Recovering the Missing Link: Predicting Class-Attribute Associations for Unsupervised Zero-Shot Learning	5975
<i>Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen</i>	
Fast Zero-Shot Image Tagging	5985
<i>Yang Zhang, Boqing Gong, and Mubarak Shah</i>	
Modality and Component Aware Feature Fusion for RGB-D Scene Classification	5995
<i>Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham</i>	
PPP: Joint Pointwise and Pairwise Image Label Prediction	6005
<i>Yilin Wang, Suhang Wang, Jiliang Tang, Huan Liu, and Baoxin Li</i>	
Cataloging Public Objects Using Aerial and Street-Level Images — Urban Trees	6014
<i>Jan D. Wegner, Steve Branson, David Hall, Konrad Schindler, and Pietro Perona</i>	
Deep Exemplar 2D-3D Detection by Adapting from Real to Rendered Views	6024
<i>Francisco Massa, Bryan C. Russell, and Mathieu Aubry</i>	
Zero-Shot Learning via Joint Latent Similarity Embedding	6034
<i>Ziming Zhang and Venkatesh Saligrama</i>	
CRAFT Objects from Images	6043
<i>Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li</i>	
Author Index	6052

Message from the General and Program Chairs

Welcome to Las Vegas and the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). In addition to the main four day program of presentations, interactive sessions, plenary talks, demos, exhibitions, and social functions, CVPR 2016 has a number of colocated events, including 29 workshops and 22 tutorials. As the field of artificial intelligence has become a major player in the technology world, this year's CVPR has made history in a number of exciting ways.

First and foremost, we received a record 2145 valid submissions to the main conference, of which 1865 were fully reviewed (the others were administratively rejected for technical or ethical reasons or were withdrawn before review). To select papers from these submissions, we invited 72 researchers to act as Area Chairs (ACs). ACs were selected to provide a broad range of expertise, to balance junior and senior members, and to represent a variety of geographical locations. Additionally, we recruited a team of 1141 experienced reviewers from the broader computer vision and pattern recognition community. The original list of reviewers was augmented with reviewers recommended by the ACs to add expertise for papers where appropriate reviewers were not initially available.

The reviewing process accepted 643 papers (29.9% of valid submissions). 83 of these were accepted as oral presentations (3.9% of valid submissions) and 123 were accepted as spotlight presentations (for a total of 9.7% of valid submissions with live presentations). All papers will also appear in the interactive poster sessions.

CVPR 2016 introduces a new form of presentation — "Spotlights". Spotlights are short oral presentations aimed at highlighting the main contributions, novelties and results of papers. Our rationale behind adding spotlight presentations to CVPR 2016 was to increase the number of works that get presented to a large audience thus gaining maximum exposure and visibility. We believe that this new presentation format (oral/spotlight/poster) will make CVPR 2016 a more exciting conference.

The review process was similar to previous years. The Program Chairs made an effort to assign papers on similar topics to the same set of Area Chairs (ACs), so that related papers could be compared directly. To achieve this, between 70 and 130 papers were pre-assigned to each AC with the help of the Toronto Paper Matching System while accounting for author provided suggestions. ACs then went over the paper titles and abstracts and bid on those they would like to handle. Using all of this information as input, ACs received a final assignment of around 30 papers for which they provided ranked lists of candidate reviewers. ACs were asked to suggest candidate reviewers who were experts in the topic, had a broad view of the field, were reliable, and could provide high quality reviews. Their choices were additionally guided by the reviewers' bids (the process followed for reviewer bidding was similar to that used for ACs). For 99.9% of the papers all 3 of the assigned reviewers were amongst those listed by ACs as candidate reviewers. For 70% of the papers at least 2 of the assigned reviewers were ranked by ACs amongst their top 3 candidates for that paper, and for 95% of the papers at least 1 reviewer was ranked amongst the top 3 candidates. Each paper was reviewed by at least three reviewers and considered by at least three ACs before a decision was made. Borderline papers and candidate orals and spotlights were discussed in groups of three non-conflicted ACs with common areas of expertise. Oral and spotlight recommendations were made by panels of 12 ACs after extensive discussion.

The Program Chairs did not submit any papers to CVPR 2016, allowing them to avoid direct conflicts throughout the review process. This year, General Chairs were allowed to submit papers, and therefore did not have any software access to the CMT system beyond that of an author. Additionally, CVPR 2016 is the first CVPR to use Researcher.cc to identify conflicts between reviewers and authors. Reviewers and ACs were excluded from any access to papers from their research groups, affiliated institutions, or collaborators. The double-blind nature of the CVPR review process was thus strictly maintained throughout.

Message from the General and Program Chairs

Since its inception in the 1960s, Computer Vision has grown into one of the most important areas of research and development in modern artificial intelligence and computer science. As we observe the blossom of computer vision technologies applied in industry, we take this opportunity at CVPR 2016 to celebrate the coming of age of computer vision, and to facilitate an even closer tie between our academic world and the industry world. As a result, for the first time in CVPR's history, we are providing an exciting, "trade-show" like atmosphere to foster maximal visibility and exposure for each on-site exhibitor from promising startups and creative standouts to the biggest industry leaders. The technologies developed by these companies range from software to hardware, and demonstrate the great impact of computer vision on different industry areas. As of now, more than 70 companies have signed up to participate in our Expo. In addition to on site booths, we also offer a unique web-presence video this year, enabling recruiting opportunities and technology highlights. So far, over 20 companies are committed to participate in this special format of presence. Furthermore, companies in the computer vision industry are showing strong interest in sponsoring the CVPR community. Together they have contributed a historical amount of sponsorship funding. Additionally, more than 10 companies showed interest in sponsoring promotional items, resulting in a significant amount of additional fundings for the conference.

CVPR 2016 has also made history in its diverse organizing committee. A record number of women are involved, from General Chairs to Area Chairs. The field of computer vision has historically been male-dominated. We are pleased to see a growing number of women and under-represented minority researchers in our community. CVPR 2016 has made an explicit statement to celebrate their contributions to our field. We hope that at the conference, you get to know some of the women leaders of our field, and encourage more women and under-represented minority students to study and work in this exciting field. We truly hope that this is only the beginning!

Last but not least, we wish to thank all members of the Organizing Committee, the 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 CVPR 2016 a first rate conference. Our thanks also go to the organizers of previous CVPRs, in particular Rene Vidal and Andrew Zisserman, for their helpful advice and support, and to Ari Kobren for a huge effort maintaining Researcher.cc. A few people working behind the scenes deserve special thanks. Yipin Zhou and Arsalan Mousavian provided critical technical assistance with CMT and managing the largest CVPR ever. It could not have been a success without them. Nicole Finn and Liz Ryan provided critical support to the general conference organization and financial well being of CVPR. Tali Treibitz and Dana Berman have provided high quality and rapid turnaround of the CVPR web pages. Eric Mortensen and Kate Saenko have again lent expertise developed over many years to shepherd the papers through the publication process. A special thank-you should also go to the Corporate Relations team - Jia Li, Kristin Dana, Brendan Morris and Andrea Frome. They have worked tirelessly to put together the inaugural CVPR Expo and to solicit industry support. Greg Mori worked with a number of teams and contributed tremendously to local arrangements, the AC meeting, CVPR website, and industry Expo. Last but not the least, this CVPR would not be possible without the record amount of support from our corporate sponsors and participants.

Finally, we wish all the attendees a highly stimulating, informative, and enjoyable conference.

Program Chairs: **Lourdes Agapito, Tamara Berg,
Jana Kosecka, Lihi Zelnik-Manor**

General Chairs: **Tinne Tuytelaars, Fei-Fei Li,
Ruzena Bajcsy**

CVPR 2016 Organizing Committee

General Chairs:	Ruzena Bajcsy Fei-Fei Li	Program Coordination Chairs:	Ira Kemelmacher-Shlizerman Greg Mori
Program Chairs:	Tinne Tuytelaars Lourdes Agapito Tamara Berg Jana Kosecka Lihi Zelnik-Manor	Corporate Relations Chairs:	Andrea Frome Jia Li Kristin Dana Brendan Morris Eric Mortensen Kate Saenko
Workshops Chair:	Svetlana Lazebnik	Publications Chairs:	Andrew Davison Adriana Kovashka Terry Boulton
Tutorials Chair:	Sanja Fidler Raquel Urtasun	Demos Chair:	Brendan Morris
Finance Chairs:	Octavia Camps Bryan Morse	Doctoral Consortium Chair:	Olga Russakovsky
Website Chairs:	Tali Treibitz Dana Berman	Technology Chair:	Tatiana Tommasi
Posters Chair:	Devi Parikh	Local Arrangements Chair:	Ginger Boulton
		Publicity & Press Chair:	
		Student Volunteers Chair:	
		Logistics Advisor:	

CVPR 2016 Area Chairs

Tal Arbel	Alyosha Efros	Frederic Jurie	Chris Pal
Ronen Basri	Irfan Essa	Ira Kemelmacher	Sylvain Paris
Dhruv Batra	Ryan Farrell	Hedvig Kjellström	Florent Perronnin
Serge Belongie	Paolo Favaro	Iasonas Kokkinos	Thomas Pock
Alex Berg	Rogério Feris	Vladlen Koltun	Marc Ranzato
Michal Black	Sanja Fidler	Nikos Komodakis	Bryan Russell
Lubomir Bourdev	Andrew Fitzgibbon	Pawan Kumar	Kate Saenko
Edmond Boyer	Charles Fowlkes	Kyros Kutulakos	Yoav Schechner
Yuri Boykov	Dieter Fox	Ivan Laptev	Greg Shakhnarovich
Gabriel Brostow	Jan-Michael Frahm	Bastian Leibe	Jianbo Shi
Octavia Camps	Pascal Fua	Jongwoo Lim	Richard Socher
Manmohan Chandraker	Ross Girshick	Steve Lin	Jian Sun
Rama Chellapa	Stephen Gould	Zhouchen Lin	Sabine Susstrunk
Dorin Comaniciu	Tal Hassner	Haibin Ling	Ayellet Tal
Rita Cucchiara	Kaiming He	Yi Ma	Jianxiong Xiao
Kostas Daniilidis	Adrian Hilton	Julien Mairal	Todd Zickler
Fernando de la Torre	Michal Irani	Michael Maire	Harpreet Sawhney
Jia Deng	Jiayia Jia	Subhransu Maji	Nuno Vasconcelos

CVPR 2016 Outstanding Reviewers

We are pleased to recognize the following researchers as "CVPR 2016 Outstanding Reviewers". These reviewers were identified by one or more of the CVPR Area Chairs

for their hard work in providing high quality and detailed reviews for their assigned papers.

Zeynep Akata
Relja Arandjelovic
Antonis Argyros
Tamar Avraham
Steve Branson
Jongmoo Choi
Andrew Davison
Alexander Fix
Andreas Geiger
Ioannis Gkioulekas
Kristen Grauman

Abhinav Gupta
C.V. Jawahar
Dinesh Jayaraman
Philipp Krahenbuhl
Josip Krapac
Jonathan Long
Philippos Mordohai
Alison Noble
Peter Ochs
Matthew O'Toole
Emanuele Rodola

Marcus Rohrbach
Javier Romero
Miki Rubinstein
Olga Russakovsky
Bogdan Savchynskyy
Stan Sclaroff
Joseph Sivic
Cees Snoek
Deqing Sun
Akihiko Torii
Tinne Tuytelaars

Anton van den Hengel
Jan van Gemert
Andrea Vedaldi
René Vidal
David Wipf
Lior Wolf
Christian Wolf
John Wright
Tao Xiang
Stefanos Zafeiriou
Maria Zontak

You Only Look Once: Unified, Real-Time Object Detection

Joseph Redmon*, Santosh Divvala*[†], Ross Girshick[‡], Ali Farhadi*[†]

University of Washington*, Allen Institute for AI[†], Facebook AI Research[‡]

<http://pjreddie.com/yolo/>

Abstract

We present YOLO, a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

Our unified architecture is extremely fast. Our base YOLO model processes images in real-time at 45 frames per second. A smaller version of the network, Fast YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on background. Finally, YOLO learns very general representations of objects. It outperforms other detection methods, including DPM and R-CNN, when generalizing from natural images to other domains like artwork.

1. Introduction

Humans glance at an image and instantly know what objects are in the image, where they are, and how they interact. The human visual system is fast and accurate, allowing us to perform complex tasks like driving with little conscious thought. Fast, accurate algorithms for object detection would allow computers to drive cars without specialized sensors, enable assistive devices to convey real-time scene information to human users, and unlock the potential for general purpose, responsive robotic systems.

Current detection systems repurpose classifiers to perform detection. To detect an object, these systems take a classifier for that object and evaluate it at various locations and scales in a test image. Systems like deformable parts models (DPM) use a sliding window approach where the classifier is run at evenly spaced locations over the entire image [10].

More recent approaches like R-CNN use region proposal

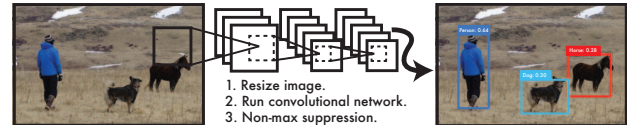


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448×448 , (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

methods to first generate potential bounding boxes in an image and then run a classifier on these proposed boxes. After classification, post-processing is used to refine the bounding boxes, eliminate duplicate detections, and rescore the boxes based on other objects in the scene [13]. These complex pipelines are slow and hard to optimize because each individual component must be trained separately.

We reframe object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. Using our system, you only look once (YOLO) at an image to predict what objects are present and where they are.

YOLO is refreshingly simple: see Figure 1. A single convolutional network simultaneously predicts multiple bounding boxes and class probabilities for those boxes. YOLO trains on full images and directly optimizes detection performance. This unified model has several benefits over traditional methods of object detection.

First, YOLO is extremely fast. Since we frame detection as a regression problem we don't need a complex pipeline. We simply run our neural network on a new image at test time to predict detections. Our base network runs at 45 frames per second with no batch processing on a Titan X GPU and a fast version runs at more than 150 fps. This means we can process streaming video in real-time with less than 25 milliseconds of latency. Furthermore, YOLO achieves more than twice the mean average precision of other real-time systems. For a demo of our system running in real-time on a webcam please see our project webpage: <http://pjreddie.com/yolo/>.

Second, YOLO reasons globally about the image when

making predictions. Unlike sliding window and region proposal-based techniques, YOLO sees the entire image during training and test time so it implicitly encodes contextual information about classes as well as their appearance. Fast R-CNN, a top detection method [14], mistakes background patches in an image for objects because it can't see the larger context. YOLO makes less than half the number of background errors compared to Fast R-CNN.

Third, YOLO learns generalizable representations of objects. When trained on natural images and tested on artwork, YOLO outperforms top detection methods like DPM and R-CNN by a wide margin. Since YOLO is highly generalizable it is less likely to break down when applied to new domains or unexpected inputs.

YOLO still lags behind state-of-the-art detection systems in accuracy. While it can quickly identify objects in images it struggles to precisely localize some objects, especially small ones. We examine these tradeoffs further in our experiments.

All of our training and testing code is open source. A variety of pretrained models are also available to download.

2. Unified Detection

We unify the separate components of object detection into a single neural network. Our network uses features from the entire image to predict each bounding box. It also predicts all bounding boxes across all classes for an image simultaneously. This means our network reasons globally about the full image and all the objects in the image. The YOLO design enables end-to-end training and real-time speeds while maintaining high average precision.

Our system divides the input image into an $S \times S$ grid. If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object.

Each grid cell predicts B bounding boxes and confidence scores for those boxes. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts. Formally we define confidence as $\Pr(\text{Object}) * \text{IOU}_{\text{pred}}^{\text{truth}}$. If no object exists in that cell, the confidence scores should be zero. Otherwise we want the confidence score to equal the intersection over union (IOU) between the predicted box and the ground truth.

Each bounding box consists of 5 predictions: x, y, w, h , and confidence. The (x, y) coordinates represent the center of the box relative to the bounds of the grid cell. The width and height are predicted relative to the whole image. Finally the confidence prediction represents the IOU between the predicted box and any ground truth box.

Each grid cell also predicts C conditional class probabilities, $\Pr(\text{Class}_i | \text{Object})$. These probabilities are conditioned on the grid cell containing an object. We only predict

one set of class probabilities per grid cell, regardless of the number of boxes B .

At test time we multiply the conditional class probabilities and the individual box confidence predictions,

$$\Pr(\text{Class}_i | \text{Object}) * \Pr(\text{Object}) * \text{IOU}_{\text{pred}}^{\text{truth}} = \Pr(\text{Class}_i) * \text{IOU}_{\text{pred}}^{\text{truth}} \quad (1)$$

which gives us class-specific confidence scores for each box. These scores encode both the probability of that class appearing in the box and how well the predicted box fits the object.

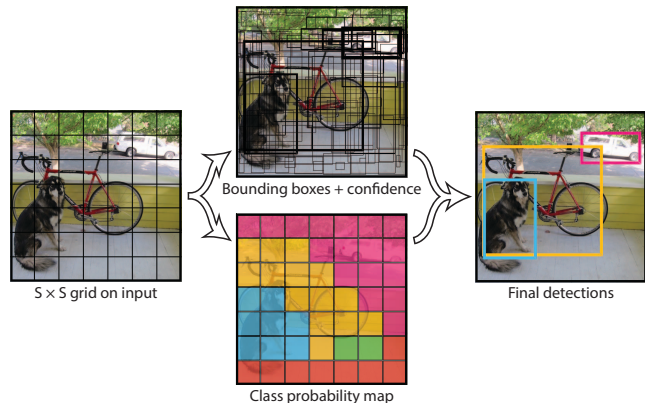


Figure 2: The Model. Our system models detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.

For evaluating YOLO on PASCAL VOC, we use $S = 7$, $B = 2$. PASCAL VOC has 20 labelled classes so $C = 20$. Our final prediction is a $7 \times 7 \times 30$ tensor.

2.1. Network Design

We implement this model as a convolutional neural network and evaluate it on the PASCAL VOC detection dataset [9]. The initial convolutional layers of the network extract features from the image while the fully connected layers predict the output probabilities and coordinates.

Our network architecture is inspired by the GoogLeNet model for image classification [33]. Our network has 24 convolutional layers followed by 2 fully connected layers. Instead of the inception modules used by GoogLeNet, we simply use 1×1 reduction layers followed by 3×3 convolutional layers, similar to Lin et al [22]. The full network is shown in Figure 3.

We also train a fast version of YOLO designed to push the boundaries of fast object detection. Fast YOLO uses a neural network with fewer convolutional layers (9 instead of 24) and fewer filters in those layers. Other than the size of the network, all training and testing parameters are the same between YOLO and Fast YOLO.

loss function:

$$\begin{aligned}
& \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
& + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\
& + \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} (C_i - \hat{C}_i)^2 \\
& + \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{noobj}} (C_i - \hat{C}_i)^2 \\
& + \sum_{i=0}^{S^2} \mathbb{1}_i^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \quad (3)
\end{aligned}$$

where $\mathbb{1}_i^{\text{obj}}$ denotes if object appears in cell i and $\mathbb{1}_{ij}^{\text{obj}}$ denotes that the j th bounding box predictor in cell i is “responsible” for that prediction.

Note that the loss function only penalizes classification error if an object is present in that grid cell (hence the conditional class probability discussed earlier). It also only penalizes bounding box coordinate error if that predictor is “responsible” for the ground truth box (i.e. has the highest IOU of any predictor in that grid cell).

We train the network for about 135 epochs on the training and validation data sets from PASCAL VOC 2007 and 2012. When testing on 2012 we also include the VOC 2007 test data for training. Throughout training we use a batch size of 64, a momentum of 0.9 and a decay of 0.0005.

Our learning rate schedule is as follows: For the first epochs we slowly raise the learning rate from 10^{-3} to 10^{-2} . If we start at a high learning rate our model often diverges due to unstable gradients. We continue training with 10^{-2} for 75 epochs, then 10^{-3} for 30 epochs, and finally 10^{-4} for 30 epochs.

To avoid overfitting we use dropout and extensive data augmentation. A dropout layer with rate = .5 after the first connected layer prevents co-adaptation between layers [18]. For data augmentation we introduce random scaling and translations of up to 20% of the original image size. We also randomly adjust the exposure and saturation of the image by up to a factor of 1.5 in the HSV color space.

2.3. Inference

Just like in training, predicting detections for a test image only requires one network evaluation. On PASCAL VOC the network predicts 98 bounding boxes per image and class probabilities for each box. YOLO is extremely fast at test time since it only requires a single network evaluation, unlike classifier-based methods.

The grid design enforces spatial diversity in the bounding box predictions. Often it is clear which grid cell an object falls in to and the network only predicts one box for each object. However, some large objects or objects near

the border of multiple cells can be well localized by multiple cells. Non-maximal suppression can be used to fix these multiple detections. While not critical to performance as it is for R-CNN or DPM, non-maximal suppression adds 2-3% in mAP.

2.4. Limitations of YOLO

YOLO imposes strong spatial constraints on bounding box predictions since each grid cell only predicts two boxes and can only have one class. This spatial constraint limits the number of nearby objects that our model can predict. Our model struggles with small objects that appear in groups, such as flocks of birds.

Since our model learns to predict bounding boxes from data, it struggles to generalize to objects in new or unusual aspect ratios or configurations. Our model also uses relatively coarse features for predicting bounding boxes since our architecture has multiple downsampling layers from the input image.

Finally, while we train on a loss function that approximates detection performance, our loss function treats errors the same in small bounding boxes versus large bounding boxes. A small error in a large box is generally benign but a small error in a small box has a much greater effect on IOU. Our main source of error is incorrect localizations.

3. Comparison to Other Detection Systems

Object detection is a core problem in computer vision. Detection pipelines generally start by extracting a set of robust features from input images (Haar [25], SIFT [23], HOG [4], convolutional features [6]). Then, classifiers [35, 21, 13, 10] or localizers [1, 31] are used to identify objects in the feature space. These classifiers or localizers are run either in sliding window fashion over the whole image or on some subset of regions in the image [34, 15, 38]. We compare the YOLO detection system to several top detection frameworks, highlighting key similarities and differences.

Deformable parts models. Deformable parts models (DPM) use a sliding window approach to object detection [10]. DPM uses a disjoint pipeline to extract static features, classify regions, predict bounding boxes for high scoring regions, etc. Our system replaces all of these disparate parts with a single convolutional neural network. The network performs feature extraction, bounding box prediction, non-maximal suppression, and contextual reasoning all concurrently. Instead of static features, the network trains the features in-line and optimizes them for the detection task. Our unified architecture leads to a faster, more accurate model than DPM.

R-CNN. R-CNN and its variants use region proposals instead of sliding windows to find objects in images. Selective

Search [34] generates potential bounding boxes, a convolutional network extracts features, an SVM scores the boxes, a linear model adjusts the bounding boxes, and non-max suppression eliminates duplicate detections. Each stage of this complex pipeline must be precisely tuned independently and the resulting system is very slow, taking more than 40 seconds per image at test time [14].

YOLO shares some similarities with R-CNN. Each grid cell proposes potential bounding boxes and scores those boxes using convolutional features. However, our system puts spatial constraints on the grid cell proposals which helps mitigate multiple detections of the same object. Our system also proposes far fewer bounding boxes, only 98 per image compared to about 2000 from Selective Search. Finally, our system combines these individual components into a single, jointly optimized model.

Other Fast Detectors Fast and Faster R-CNN focus on speeding up the R-CNN framework by sharing computation and using neural networks to propose regions instead of Selective Search [14] [27]. While they offer speed and accuracy improvements over R-CNN, both still fall short of real-time performance.

Many research efforts focus on speeding up the DPM pipeline [30] [37] [5]. They speed up HOG computation, use cascades, and push computation to GPUs. However, only 30Hz DPM [30] actually runs in real-time.

Instead of trying to optimize individual components of a large detection pipeline, YOLO throws out the pipeline entirely and is fast by design.

Detectors for single classes like faces or people can be highly optimized since they have to deal with much less variation [36]. YOLO is a general purpose detector that learns to detect a variety of objects simultaneously.

Deep MultiBox. Unlike R-CNN, Szegedy et al. train a convolutional neural network to predict regions of interest [8] instead of using Selective Search. MultiBox can also perform single object detection by replacing the confidence prediction with a single class prediction. However, MultiBox cannot perform general object detection and is still just a piece in a larger detection pipeline, requiring further image patch classification. Both YOLO and MultiBox use a convolutional network to predict bounding boxes in an image but YOLO is a complete detection system.

OverFeat. Sermanet et al. train a convolutional neural network to perform localization and adapt that localizer to perform detection [31]. OverFeat efficiently performs sliding window detection but it is still a disjoint system. OverFeat optimizes for localization, not detection performance. Like DPM, the localizer only sees local information when making a prediction. OverFeat cannot reason about global context and thus requires significant post-processing to produce coherent detections.

MultiGrasp. Our work is similar in design to work on

grasp detection by Redmon et al [26]. Our grid approach to bounding box prediction is based on the MultiGrasp system for regression to grasps. However, grasp detection is a much simpler task than object detection. MultiGrasp only needs to predict a single graspable region for an image containing one object. It doesn't have to estimate the size, location, or boundaries of the object or predict it's class, only find a region suitable for grasping. YOLO predicts both bounding boxes and class probabilities for multiple objects of multiple classes in an image.

4. Experiments

First we compare YOLO with other real-time detection systems on PASCAL VOC 2007. To understand the differences between YOLO and R-CNN variants we explore the errors on VOC 2007 made by YOLO and Fast R-CNN, one of the highest performing versions of R-CNN [14]. Based on the different error profiles we show that YOLO can be used to rescore Fast R-CNN detections and reduce the errors from background false positives, giving a significant performance boost. We also present VOC 2012 results and compare mAP to current state-of-the-art methods. Finally, we show that YOLO generalizes to new domains better than other detectors on two artwork datasets.

4.1. Comparison to Other Real-Time Systems

Many research efforts in object detection focus on making standard detection pipelines fast. [5] [37] [30] [14] [17] [27] However, only Sadeghi et al. actually produce a detection system that runs in real-time (30 frames per second or better) [30]. We compare YOLO to their GPU implementation of DPM which runs either at 30Hz or 100Hz. While the other efforts don't reach the real-time milestone we also compare their relative mAP and speed to examine the accuracy-performance tradeoffs available in object detection systems.

Fast YOLO is the fastest object detection method on PASCAL; as far as we know, it is the fastest extant object detector. With 52.7% mAP, it is more than twice as accurate as prior work on real-time detection. YOLO pushes mAP to 63.4% while still maintaining real-time performance.

We also train YOLO using VGG-16. This model is more accurate but also significantly slower than YOLO. It is useful for comparison to other detection systems that rely on VGG-16 but since it is slower than real-time the rest of the paper focuses on our faster models.

Fastest DPM effectively speeds up DPM without sacrificing much mAP but it still misses real-time performance by a factor of 2 [37]. It also is limited by DPM's relatively low accuracy on detection compared to neural network approaches.

R-CNN minus R replaces Selective Search with static bounding box proposals [20]. While it is much faster than

Real-Time Detectors	Train	mAP	FPS
100Hz DPM [30]	2007	16.0	100
30Hz DPM [30]	2007	26.1	30
Fast YOLO	2007+2012	52.7	155
YOLO	2007+2012	63.4	45
Less Than Real-Time			
Fastest DPM [37]	2007	30.4	15
R-CNN Minus R [20]	2007	53.5	6
Fast R-CNN [14]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[27]	2007+2012	73.2	7
Faster R-CNN ZF [27]	2007+2012	62.1	18
YOLO VGG-16	2007+2012	66.4	21

Table 1: Real-Time Systems on PASCAL VOC 2007. Comparing the performance and speed of fast detectors. Fast YOLO is the fastest detector on record for PASCAL VOC detection and is still twice as accurate as any other real-time detector. YOLO is 10 mAP more accurate than the fast version while still well above real-time in speed.

R-CNN, it still falls short of real-time and takes a significant accuracy hit from not having good proposals.

Fast R-CNN speeds up the classification stage of R-CNN but it still relies on selective search which can take around 2 seconds per image to generate bounding box proposals. Thus it has high mAP but at 0.5 fps it is still far from real-time.

The recent Faster R-CNN replaces selective search with a neural network to propose bounding boxes, similar to Szegedy et al. [8] In our tests, their most accurate model achieves 7 fps while a smaller, less accurate one runs at 18 fps. The VGG-16 version of Faster R-CNN is 10 mAP higher but is also 6 times slower than YOLO. The Zeiler-Fergus Faster R-CNN is only 2.5 times slower than YOLO but is also less accurate.

4.2. VOC 2007 Error Analysis

To further examine the differences between YOLO and state-of-the-art detectors, we look at a detailed breakdown of results on VOC 2007. We compare YOLO to Fast R-CNN since Fast R-CNN is one of the highest performing detectors on PASCAL and it's detections are publicly available.

We use the methodology and tools of Hoiem et al. [19] For each category at test time we look at the top N predictions for that category. Each prediction is either correct or it is classified based on the type of error:

- Correct: correct class and $\text{IOU} > .5$
- Localization: correct class, $.1 < \text{IOU} < .5$
- Similar: class is similar, $\text{IOU} > .1$

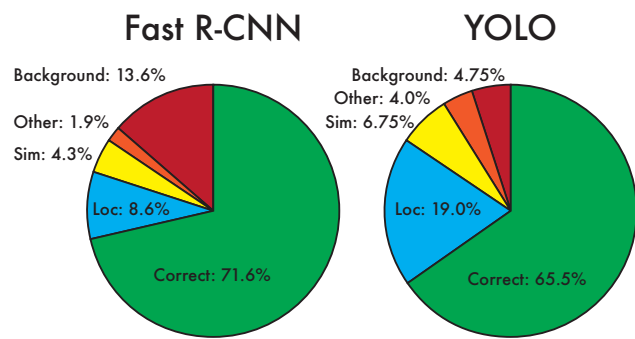


Figure 4: Error Analysis: Fast R-CNN vs. YOLO These charts show the percentage of localization and background errors in the top N detections for various categories ($N = \#$ objects in that category).

- Other: class is wrong, $\text{IOU} > .1$
- Background: $\text{IOU} < .1$ for any object

Figure 4 shows the breakdown of each error type averaged across all 20 classes.

YOLO struggles to localize objects correctly. Localization errors account for more of YOLO's errors than all other sources combined. Fast R-CNN makes much fewer localization errors but far more background errors. 13.6% of it's top detections are false positives that don't contain any objects. Fast R-CNN is almost 3x more likely to predict background detections than YOLO.

4.3. Combining Fast R-CNN and YOLO

YOLO makes far fewer background mistakes than Fast R-CNN. By using YOLO to eliminate background detections from Fast R-CNN we get a significant boost in performance. For every bounding box that R-CNN predicts we check to see if YOLO predicts a similar box. If it does, we give that prediction a boost based on the probability predicted by YOLO and the overlap between the two boxes.

The best Fast R-CNN model achieves a mAP of 71.8% on the VOC 2007 test set. When combined with YOLO, its

	mAP	Combined	Gain
Fast R-CNN	71.8	-	-
Fast R-CNN (2007 data)	66.9	72.4	.6
Fast R-CNN (VGG-M)	59.2	72.4	.6
Fast R-CNN (CaffeNet)	57.1	72.1	.3
YOLO	63.4	75.0	3.2

Table 2: Model combination experiments on VOC 2007. We examine the effect of combining various models with the best version of Fast R-CNN. Other versions of Fast R-CNN provide only a small benefit while YOLO provides a significant performance boost.

VOC 2012 test	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
MR_CNN_MORE_DATA [11]	73.9	85.5	82.9	76.6	57.8	62.7	79.4	77.2	86.6	55.0	79.1	62.2	87.0	83.4	84.7	78.9	45.3	73.4	65.8	80.3	74.0
HyperNet_VGG	71.4	84.2	78.5	73.6	55.6	53.7	78.7	79.8	87.7	49.6	74.9	52.1	86.0	81.7	83.3	81.8	48.6	73.5	59.4	79.9	65.7
HyperNet_SP	71.3	84.1	78.3	73.3	55.5	53.6	78.6	79.6	87.5	49.5	74.9	52.1	85.6	81.6	83.2	81.6	48.4	73.2	59.3	79.7	65.6
Fast R-CNN + YOLO	70.7	83.4	78.5	73.5	55.8	43.4	79.1	73.1	89.4	49.4	75.5	57.0	87.5	80.9	81.0	74.7	41.8	71.5	68.5	82.1	67.2
MR_CNN_S_CNN [11]	70.7	85.0	79.6	71.5	55.3	57.7	76.0	73.9	84.6	50.5	74.3	61.7	85.5	79.9	81.7	76.4	41.0	69.0	61.2	77.7	72.1
Faster R-CNN [27]	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
DEEP_ENS_COCO	70.1	84.0	79.4	71.6	51.9	51.1	74.1	72.1	88.6	48.3	73.4	57.8	86.1	80.0	80.7	70.4	46.6	69.6	68.8	75.9	71.4
NoC [28]	68.8	82.8	79.0	71.6	52.3	53.7	74.1	69.0	84.9	46.9	74.3	53.1	85.0	81.3	79.5	72.2	38.9	72.4	59.5	76.7	68.1
Fast R-CNN [14]	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0	55.0	87.5	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
UMICH_FGS_STRUCT	66.4	82.9	76.1	64.1	44.6	49.4	70.3	71.2	84.6	42.7	68.6	55.8	82.7	77.1	79.9	68.7	41.4	69.0	60.0	72.0	66.2
NUS_NIN_C2000 [7]	63.8	80.2	73.8	61.9	43.7	43.0	70.3	67.6	80.7	41.9	69.7	51.7	78.2	75.2	76.9	65.1	38.6	68.3	58.0	68.7	63.3
BabyLearning [7]	63.2	78.0	74.2	61.3	45.7	42.7	68.2	66.8	80.2	40.6	70.0	49.8	79.0	74.5	77.9	64.0	35.3	67.9	55.7	68.7	62.6
NUS_NIN	62.4	77.9	73.1	62.6	39.5	43.3	69.1	66.4	78.9	39.1	68.1	50.0	77.2	71.3	76.1	64.7	38.4	66.9	56.2	66.9	62.7
R-CNN VGG BB [13]	62.4	79.6	72.7	61.9	41.2	41.9	65.9	66.4	84.6	38.5	67.2	46.7	82.0	74.8	76.0	65.2	35.6	65.4	54.2	67.4	60.3
R-CNN VGG [13]	59.2	76.8	70.9	56.6	37.5	36.9	62.9	63.6	81.1	35.7	64.3	43.9	80.4	71.6	74.0	60.0	30.8	63.4	52.0	63.5	58.7
YOLO	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
Feature Edit [32]	56.3	74.6	69.1	54.4	39.1	33.1	65.2	62.7	69.7	30.8	56.0	44.6	70.0	64.4	71.1	60.2	33.3	61.3	46.4	61.7	57.8
R-CNN BB [13]	53.3	71.8	65.8	52.0	34.1	32.6	59.6	60.0	69.8	27.6	52.0	41.7	69.6	61.3	68.3	57.8	29.6	57.8	40.9	59.3	54.1
SDS [16]	50.7	69.7	58.4	48.5	28.3	28.8	61.3	57.5	70.8	24.1	50.7	35.9	64.9	59.1	65.8	57.1	26.0	58.8	38.6	58.9	50.7
R-CNN [13]	49.6	68.1	63.8	46.1	29.4	27.9	56.6	57.0	65.9	26.5	48.7	39.5	66.2	57.3	65.4	53.2	26.2	54.5	38.1	50.6	51.6

Table 3: PASCAL VOC 2012 Leaderboard. YOLO compared with the full comp4 (outside data allowed) public leaderboard as of November 6th, 2015. Mean average precision and per-class average precision are shown for a variety of detection methods. YOLO is the only real-time detector. Fast R-CNN + YOLO is the forth highest scoring method, with a 2.3% boost over Fast R-CNN.

mAP increases by 3.2% to 75.0%. We also tried combining the top Fast R-CNN model with several other versions of Fast R-CNN. Those ensembles produced small increases in mAP between .3 and .6%, see Table 2 for details.

The boost from YOLO is not simply a byproduct of model ensembling since there is little benefit from combining different versions of Fast R-CNN. Rather, it is precisely because YOLO makes different kinds of mistakes at test time that it is so effective at boosting Fast R-CNN’s performance.

Unfortunately, this combination doesn’t benefit from the speed of YOLO since we run each model separately and then combine the results. However, since YOLO is so fast it doesn’t add any significant computational time compared to Fast R-CNN.

4.4. VOC 2012 Results

On the VOC 2012 test set, YOLO scores 57.9% mAP. This is lower than the current state of the art, closer to the original R-CNN using VGG-16, see Table 3. Our system struggles with small objects compared to its closest competitors. On categories like bottle, sheep, and tv/monitor YOLO scores 8-10% lower than R-CNN or Feature Edit. However, on other categories like cat and train YOLO achieves higher performance.

Our combined Fast R-CNN + YOLO model is one of the highest performing detection methods. Fast R-CNN gets a 2.3% improvement from the combination with YOLO, boosting it 5 spots up on the public leaderboard.

4.5. Generalizability: Person Detection in Artwork

Academic datasets for object detection draw the training and testing data from the same distribution. In real-world applications it is hard to predict all possible use cases and

the test data can diverge from what the system has seen before [3]. We compare YOLO to other detection systems on the Picasso Dataset [12] and the People-Art Dataset [3], two datasets for testing person detection on artwork.

Figure 5 shows comparative performance between YOLO and other detection methods. For reference, we give VOC 2007 detection AP on person where all models are trained only on VOC 2007 data. On Picasso models are trained on VOC 2012 while on People-Art they are trained on VOC 2010.

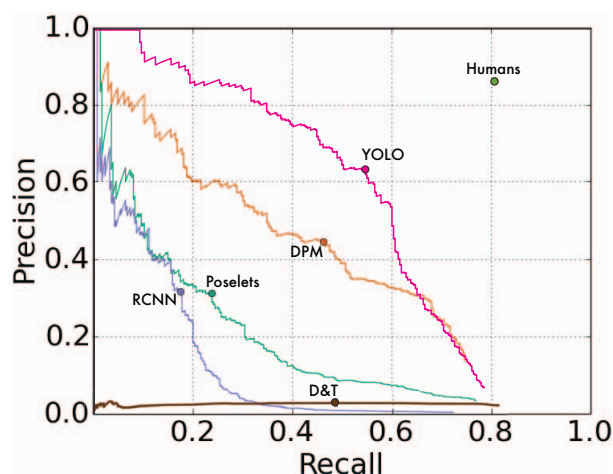
R-CNN has high AP on VOC 2007. However, R-CNN drops off considerably when applied to artwork. R-CNN uses Selective Search for bounding box proposals which is tuned for natural images. The classifier step in R-CNN only sees small regions and needs good proposals.

DPM maintains its AP well when applied to artwork. Prior work theorizes that DPM performs well because it has strong spatial models of the shape and layout of objects. Though DPM doesn’t degrade as much as R-CNN, it starts from a lower AP.

YOLO has good performance on VOC 2007 and its AP degrades less than other methods when applied to artwork. Like DPM, YOLO models the size and shape of objects, as well as relationships between objects and where objects commonly appear. Artwork and natural images are very different on a pixel level but they are similar in terms of the size and shape of objects, thus YOLO can still predict good bounding boxes and detections.

5. Real-Time Detection In The Wild

YOLO is a fast, accurate object detector, making it ideal for computer vision applications. We connect YOLO to a webcam and verify that it maintains real-time performance,



(a) Picasso Dataset precision-recall curves.

	VOC 2007 AP	Picasso AP Best F_1	People-Art AP
YOLO	59.2	53.3 0.590	45
R-CNN	54.2	10.4 0.226	26
DPM	43.2	37.8 0.458	32
Poselets [2]	36.5	17.8 0.271	
D&T [4]	-	1.9 0.051	

(b) Quantitative results on the VOC 2007, Picasso, and People-Art Datasets. The Picasso Dataset evaluates on both AP and best F_1 score.

Figure 5: Generalization results on Picasso and People-Art datasets.

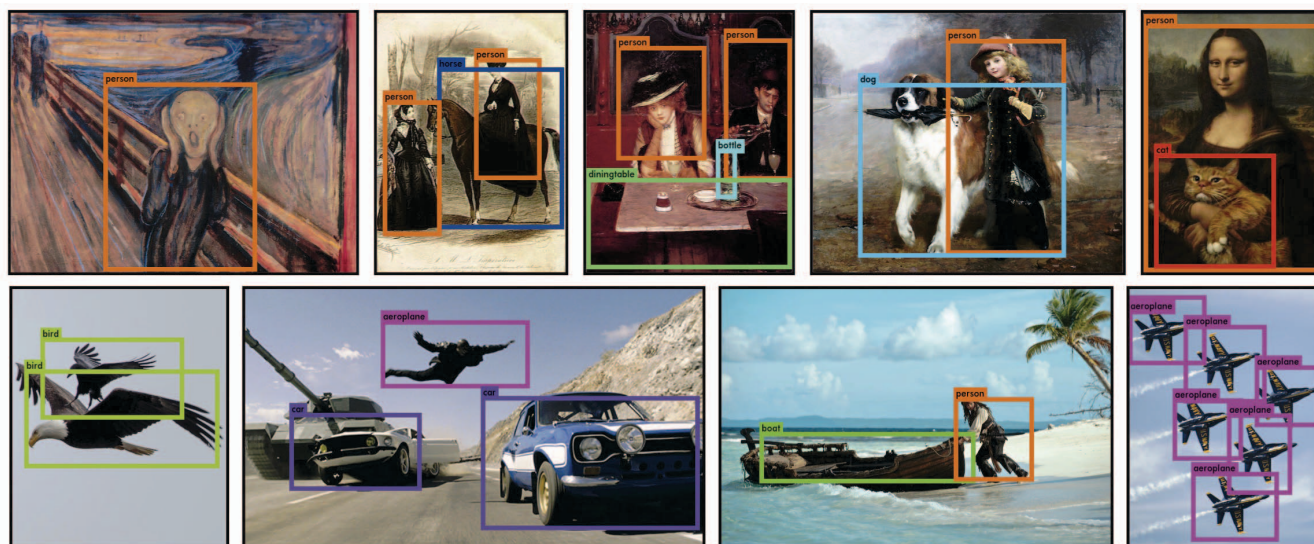


Figure 6: Qualitative Results. YOLO running on sample artwork and natural images from the internet. It is mostly accurate although it does think one person is an airplane.

including the time to fetch images from the camera and display the detections.

The resulting system is interactive and engaging. While YOLO processes images individually, when attached to a webcam it functions like a tracking system, detecting objects as they move around and change in appearance. A demo of the system and the source code can be found on our project website: <http://pjreddie.com/yolo/>.

6. Conclusion

We introduce YOLO, a unified model for object detection. Our model is simple to construct and can be trained

directly on full images. Unlike classifier-based approaches, YOLO is trained on a loss function that directly corresponds to detection performance and the entire model is trained jointly.

Fast YOLO is the fastest general-purpose object detector in the literature and YOLO pushes the state-of-the-art in real-time object detection. YOLO also generalizes well to new domains making it ideal for applications that rely on fast, robust object detection.

Acknowledgements: This work is partially supported by ONR N00014-13-1-0720, NSF IIS-1338054, and The Allen Distinguished Investigator Award.

References

- [1] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *Computer Vision–ECCV 2008*, pages 2–15. Springer, 2008. 4
- [2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*, 2009. 8
- [3] H. Cai, Q. Wu, T. Corradi, and P. Hall. The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs. *arXiv preprint arXiv:1505.00110*, 2015. 7
- [4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 4, 8
- [5] T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, J. Yagnik, et al. Fast, accurate detection of 100,000 object classes on a single machine. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1814–1821. IEEE, 2013. 5
- [6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 4
- [7] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and semantic segmentation. In *Computer Vision–ECCV 2014*, pages 299–314. Springer, 2014. 7
- [8] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2155–2162. IEEE, 2014. 5, 6
- [9] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 2
- [10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1, 4
- [11] S. Gidaris and N. Komodakis. Object detection via a multi-region & semantic segmentation-aware CNN model. *CoRR*, abs/1505.01749, 2015. 7
- [12] S. Ginosar, D. Haas, T. Brown, and J. Malik. Detecting people in cubist art. In *Computer Vision–ECCV 2014 Workshops*, pages 101–116. Springer, 2014. 7
- [13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014. 1, 4, 7
- [14] R. B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. 2, 5, 6, 7
- [15] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *Advances in neural information processing systems*, pages 655–663, 2009. 4
- [16] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *Computer Vision–ECCV 2014*, pages 297–312. Springer, 2014. 7
- [17] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *arXiv preprint arXiv:1406.4729*, 2014. 5
- [18] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 4
- [19] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *Computer Vision–ECCV 2012*, pages 340–353. Springer, 2012. 6
- [20] K. Lenc and A. Vedaldi. R-cnn minus r. *arXiv preprint arXiv:1506.06981*, 2015. 5, 6
- [21] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–900. IEEE, 2002. 4
- [22] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 2
- [23] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 4
- [24] D. Mishkin. Models accuracy on imagenet 2012 val. <https://github.com/BVLC/caffe/wiki/Models-accuracy-on-ImageNet-2012-val>. Accessed: 2015-10-2. 3
- [25] C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Computer vision, 1998. sixth international conference on*, pages 555–562. IEEE, 1998. 4
- [26] J. Redmon and A. Angelova. Real-time grasp detection using convolutional neural networks. *CoRR*, abs/1412.3128, 2014. 5
- [27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 5, 6, 7
- [28] S. Ren, K. He, R. B. Girshick, X. Zhang, and J. Sun. Object detection networks on convolutional feature maps. *CoRR*, abs/1504.06066, 2015. 3, 7
- [29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015. 3
- [30] M. A. Sadeghi and D. Forsyth. 30hz object detection with dpm v5. In *Computer Vision–ECCV 2014*, pages 65–79. Springer, 2014. 5, 6
- [31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. 4, 5
- [32] Z. Shen and X. Xue. Do more dropouts in pool5 feature maps for better object detection. *arXiv preprint arXiv:1409.6911*, 2014. 7

- [33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 2
- [34] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 4, 5
- [35] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 4:34–47, 2001. 4
- [36] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 5
- [37] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The fastest deformable part model for object detection. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2497–2504. IEEE, 2014. 5, 6
- [38] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014*, pages 391–405. Springer, 2014. 4