CNN-PS: CNN-Based Photometric
Stereo for General Non-convex Surfaces
Satoshi Ikehata(B)
National Institute of Informatics, Tokyo, Japan
sikehata@nii.ac.jp
Abstract. Most conventional photometric stereo algorithms inversely
solve a BRDF-based image formation model. However, the actual imag-
ing process is often far more complex due to the global light transport on
the non-convex surfaces. This paper presents a photometric stereo net-work that
directly learns relationships between the photometric stereo
input and surface normals of a scene. For handling unordered, arbitrary
number of input images, we merge all the input data to the intermediaterepresent
ation called observation map that has a ■xed shape, is able to
be fed into a CNN. To improve both training and prediction, we take
into account the rotational pseudo-invariance of the observation mapthat is deri
ved from the isotropic constraint. For training the network,
we create a synthetic photometric stereo dataset that is generated by a
physics-based renderer, therefore the global light transport is considered.Our e
xperimental results on both synthetic and real datasets show that
our method outperforms conventional BRDF-based photometric stereo
algorithms especially when scenes are highly non-convex.
Keywords: Photometric stereo
·Convolutional neural networks

**********
Dynamic Conditional Networks
for Few-Shot Learning
Fang Zhao1, Jian Zhao1,2(B), Shuicheng Yan1,3, and Jiashi Feng1
1National University of Singapore, Singapore, Singapore
{elezhf,eleyans,elefjia }@nus.edu.sg
2National University of Defense Technology, Hunan, China
3Qihoo 360 AI Institute, Beijing, China
zhaojian90@u.nus.edu
https://zhaoj9014.github.io/
Abstract. This paper proposes a novel Dynamic Conditional Convolu-
tional Network (DCCN) to handle conditional few-shot learning, i.e, only
a few training samples are available for each condition. DCCN consists of
dual subnets: DyConvNet contains a dynamic convolutional layer witha bank of bas
is ■lters; CondiNet predicts a set of adaptive weights from
conditional inputs to linearly combine the basis ■lters. In this manner, a
speci■c convolutional kernel can be dynamically obtained for each condi-tional i
nput. The ■lter bank is shared between all conditions thus only a
low-dimension weight vector needs to be learned. This signi■cantly facil-
itates the parameter learning across di■erent conditions when trainingdata are l
imited. We evaluate DCCN on four tasks which can be formu-
lated as conditional model learning, including speci■c object counting,
multi-modal image classi■cation, phrase grounding and identity basedface generat
ion. Extensive experiments demonstrate the superiority of
the proposed model in the conditional few-shot learning setting.
Keywords: Conditional model
·Few-shot learning ·Deep learning
Dynamic convolution ·Filter bank

**********
Deep Factorised Inverse-Sketching
Kaiyue Pang1(B),D aL i1, Jifei Song1, Yi-Zhe Song1, Tao Xiang1,
and Timothy M. Hospedales1,2
1SketchX, Queen Mary University of London, London, UK
{kaiyue.pang,da.li,yizhe.song,j.song,t.xiang }@qmul.ac.uk
2The University of Edinburgh, Edinburgh, UK

t.hospedales@ed.ac.uk
Abstract. Modelling human free-hand sketches has become topical
recently, driven by practical applications such as ■ne-grained sketch
based image retrieval (FG-SBIR). Sketches are clearly related to photoedge-maps,
 but a human free-hand sketch of a photo is not simply a
clean rendering of that photo’s edge map. Instead there is a fundamen-
tal process of abstraction and iconic rendering, where overall geometry iswarped
 and salient details are selectively included. In this paper we study
this sketching process and attempt to invert it. We model this inversion
by translating iconic free-hand sketches to contours that resemble more
geometrically realistic projections of object boundaries, and separately
factorise out the salient added details. This factorised re-representationmakes
it easier to match a free-hand sketch to a photo instance of an
object. Speci■cally, we propose a novel unsupervised image style transfer
model based on enforcing a cyclic embedding consistency constraint. Adeep FG-SBI
R model is then formulated to accommodate complemen-
tary discriminative detail from each factorised sketch for better matching
with the corresponding photo. Our method is evaluated both qualita-tively and qu
antitatively to demonstrate its superiority over a number
of state-of-the-art alternatives for style transfer and FG-SBIR.
1
**********

Look Deeper into Depth: Monocular
Depth Estimation with Semantic Booster
and Attention-Driven Loss
Jianbo Jiao1,2(B), Ying Cao1, Yibing Song3, and Rynson Lau1
1City University of Hong Kong, Kowloon, Hong Kong SAR
jianbjiao2-c@my.cityu.edu.hk, caoying59@gmail.com,
dynamicstevenson@gmail.com
2University of Illinois at Urbana-Champaign, Urbana, USA
3Tencent AI Lab, Shenzhen, China
rynson.lau@cityu.edu.hk
Abstract. Monocular depth estimation bene■ts greatly from learning
based techniques. By studying the training data, we observe that the
per-pixel depth values in existing datasets typically exhibit a long-taileddistr
ibution. However, most previous approaches treat all the regions in
the training data equally regardless of the imbalanced depth distribu-
tion, which restricts the model performance particularly on distant depthregions
. In this paper, we investigate the long tail property and delve
deeper into the distant depth regions ( i.e.the tail part) to propose an
attention-driven loss for the network supervision. In addition, to betterleverag
e the semantic information for monocular depth estimation, we pro-
pose a synergy network to automatically learn the information sharing
strategies between the two tasks. With the proposed attention-driven loss
and synergy network, the depth estimation and semantic labeling tasks
can be mutually improved. Experiments on the challenging indoor datasetshow that
 the proposed approach achieves state-of-the-art performance on
both monocular depth estimation and semantic labeling tasks.
Keywords: Monocular depth
·Semantic labeling ·Attention loss
1
**********

Summarizing First-Person Videos from
Third Persons’ Points of Views
Hsuan-I Ho1, Wei-Chen Chiu2, and Yu-Chiang Frank Wang1(B)
1Department of Electrical Engineering, National Taiwan University, Taiwan,
Republic of China
{b01901029,ycwang }@ntu.edu.tw
2Department of Computer Science, National Chiao Tung University, Taiwan,
Republic of China

walon@cs.nctu.edu.tw
Abstract. Video highlight or summarization is among interesting topics
in computer vision, which bene■ts a variety of applications like viewing,
searching, or storage. However, most existing studies rely on training
data of third-person videos, which cannot easily generalize to highlightthe ■rst
-person ones. With the goal of deriving an e■ective model to
summarize ■rst-person videos, we propose a novel deep neural network
architecture for describing and discriminating vital spatiotemporal infor-mation
across videos with di■erent points of view. Our proposed model
is realized in a semi-supervised setting, in which fully annotated third-
person videos, unlabeled ■rst-person videos, and a small number of anno-tated ■r
st-person ones are presented during training. In our experiments,
qualitative and quantitative evaluations on both benchmarks and our
collected ■rst-person video datasets are presented.
Keywords: Video summarization
·First-person vision
Transfer learning ·Metric learning

**********

Learning Single-View 3D Reconstruction
with Limited Pose Supervision

Guandao Yang1(B), Yin Cui1,2, Serge Belongie1,2, and Bharath Hariharan1
1Department of Computer Science, Cornell University, Ithaca, USA
gy46@cornell.edu
2Cornell Tech, New York, USA
Abstract. It is expensive to label images with 3D structure or precise
camera pose. Yet, this is precisely the kind of annotation required to train
single-view 3D reconstruction models. In contrast, unlabeled images or
images with just category labels are easy to acquire, but few currentmodels can
use this weak supervision. We present a uni■ed framework
that can combine both types of supervision: a small amount of cam-
era pose annotations are used to enforce pose-invariance and view-point
consistency, and unlabeled images combined with an adversarial loss are
used to enforce the realism of rendered, generated models. We use thisuni■ed fra
mework to measure the impact of each form of supervision in
three paradigms: semi-supervised, multi-task, and transfer learning. We
show that with a combination of these ideas, we can train single-viewreconstruct
ion models that improve up to 7 points in performance (AP)
when using only 1% pose annotated training data.
Keywords: Single-image 3D-reconstruction
·Few-shot learning
GANs

**********

Weakly- and Semi-supervised Panoptic
Segmentation

Qizhu Li, Anurag Arnab(B), and Philip H.S. Torr
University of Oxford, Oxford, UK
{liqizhu,aarnab,phst }@robots.ox.ac.uk
Abstract. We present a weakly supervised model that jointly per-
forms both semantic- and instance-segmentation – a particularly rele-vant proble
m given the substantial cost of obtaining pixel-perfect anno-
tation for these tasks. In contrast to many popular instance segmen-
tation approaches based on object detectors, our method does notpredict any over
lapping instances. Moreover, we are able to segment
both “thing” and “stu■” classes, and thus explain all the pixels in
the image. “Thing” classes are weakly-supervised with bounding boxes,and “stu■”
with image-level tags. We obtain state-of-the-art results on
Pascal VOC, for both full and weak supervision (which achieves about
95% of fully-supervised performance). Furthermore, we present the ■rstweakly-sup

ervised results on Cityscapes for both semantic- and instance-segmentation. Finally, we use our weakly supervised framework to analyse the relationship between annotation quality and predictive perfor-mance, which is of interest to dataset creators.

Keywords: Weak supervision
·Instance segmentation
Semantic segmentation ·Scene understanding

1

**********

## Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation

Markus Oberweger1(B), Mahdi Rad1, and Vincent Lepetit2,1
1Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria
{oberweger,rad,lepetit }@icg.tugraz.at
2Laboratoire Bordelais de Recherche en Informatique, Universit´ ed eB o r d e a u x ,
Bordeaux, France

Abstract. We introduce a novel method for robust and accurate 3D object pose estimation from a single color image under large occlusions. Following recent approaches, we ■rst predict the 2D projections of 3Dpoints related to the target object and then compute the 3D pose from these correspondences using a geometric method. Unfortunately, as the results of our experiments show, predicting these 2D projections usinga regular CNN or a Convolutional Pose Machine is highly sensitive to partial occlusions, even when these methods are trained with partially occluded examples. Our solution is to predict heatmaps from multiplesmall patches independently and to accumulate the results to obtain accurate and robust predictions. Training subsequently becomes challenging because patches with similar appearances but di■erent positionson the object correspond to di■erent heatmaps. However, we provide a simple yet e■ective solution to deal with such ambiguities. We show that our approach outperforms existing methods on two challengingdatasets: The Occluded LineMOD dataset and the YCB-Video dataset, both exhibiting cluttered scenes with highly occluded objects.

Keywords: 3D object pose estimation
·Heatmaps ·Occlusions

1

**********

## Deep Co-Training for Semi-Supervised Image Recognition

Siyuan Qiao1(B), Wei Shen1,2, Zhishuai Zhang1,B oW a n g3, and Alan Yuille1
1Johns Hopkins University, Baltimore, USA
siyuan.qiao@jhu.edu
2Shanghai University, Shanghai, China
3Hikvision Research Institute, Hangzhou, China

Abstract. In this paper, we study the problem of semi-supervised image recognition, which is to learn classi■ers using both labeled and unlabeled images. We present Deep Co-Training, a deep learning based methodinspired by the Co-Training framework. The original Co-Training learns two classi■ers on two views which are data from di■erent sources that describe the same instances. To extend this concept to deep learning, Deep Co-Training trains multiple deep neural networks to be the di■erent views and exploits adversarial examples to encourage view di■erence, in order to prevent the networks from collapsing into each other. As a result,the co-trained networks provide di■erent and complementary information about the data, which is necessary for the Co-Training framework to achieve good results. We test our method on SVHN, CIFAR-10/100and ImageNet datasets, and our method outperforms the previous state-of-the-art methods by a large margin.

Keywords: Co-Training
·Deep networks ·Semi-supervised learning
1
**********
Visual Coreference Resolution in Visual
Dialog Using Neural Module Networks
Satwik Kottur1,2(B),J o s ´eM .F .M o u r a2, Devi Parikh1,3, Dhruv Batra1,3,
and Marcus Rohrbach1
1Facebook AI Research, Menlo Park, USA
2Carnegie Mellon University, Pittsburgh, USA
skottur@andrew.cmu.edu
3Georgia Institute of Technology, Atlanta, USA
Abstract. Visual dialog entails answering a series of questions grounded
in an image, using dialog history as context. In addition to the challenges
found in visual question answering (VQA), which can be seen as one-
round dialog, visual dialog encompasses several more. We focus on onesuch proble
m called visual coreference resolution that involves determin-
ing which words, typically noun phrases and pronouns, co-refer to the
same entity/object instance in an image. This is crucial, especially forpronouns
 (e.g., ‘ it’), as the dialog agent must ■rst link it to a previous
coreference (e.g., ‘boat’), and only then can rely on the visual grounding
of the coreference ‘boat’ to reason about the pronoun ‘ it’. Prior work
(in visual dialog) models visual coreference resolution either (a) implic-
itly via a memory network over history, or (b) at a coarse level for the
entire question; and not explicitly at a phrase level of granularity. Inthis wor
k, we propose a neural module network architecture for visual
dialog by introducing two novel modules— Refer andExclude —that per-
form explicit, grounded, coreference resolution at a ■ner word level. Wedemonstr
ate the e■ectiveness of our model on MNIST Dialog, a visually
simple yet coreference-wise complex dataset, by achieving near perfect
accuracy, and on VisDial, a large and challenging visual dialog dataseton real i
mages, where our model outperforms other approaches, and is
more interpretable, grounded, and consistent qualitatively.
1
**********
Learning Blind Video Temporal
Consistency
Wei-Sheng Lai1(B), Jia-Bin Huang2, Oliver Wang3, Eli Shechtman3,
Ersin Yumer4, and Ming-Hsuan Yang1,5
1UC Merced, Merced, USA
wlai24@ucmerced.edu
2Virginia Tech, Blacksburg, USA
3Adobe Research, Seattle, USA
4Argo AI, Mountain View, USA
5Google Cloud AI, Sunnyvale, USA
Abstract. Applying image processing algorithms independently to
each frame of a video often leads to undesired inconsistent results overtime. De
veloping temporally consistent video-based extensions, however,
requires domain knowledge for individual tasks and is unable to general-
ize to other applications. In this paper, we present an e■cient approachbased on
 a deep recurrent network for enforcing temporal consistency in
a video. Our method takes the original and per-frame processed videos as
inputs to produce a temporally consistent video. Consequently, our app-roach is
agnostic to speci■c image processing algorithms applied to the
original video. We train the proposed network by minimizing both short-
term and long-term temporal losses as well as a perceptual loss to strikea balan
ce between temporal coherence and perceptual similarity with the
processed frames. At test time, our model does not require computing
optical ■ow and thus achieves real-time speed even for high-resolutionvideos. We
 show that our single model can handle multiple and unseen

tasks, including but not limited to artistic style transfer, enhancement, colorization, image-to-image translation and intrinsic image decomposi-tion. Ext ensive objective evaluation and subject study demonstrate that the proposed approach performs favorably against the state-of-the-art methods on various types of videos.

1

**********

Salient Objects in Clutter:
Bringing Salient Object Detection
to the Foreground

Deng-Ping Fan1, Ming-Ming Cheng1(B), Jiang-Jiang Liu1,
Shang-Hua Gao1, Qibin Hou1, and Ali Borji2
1College of Computer Science, Nankai University, Tianjin, China
cmm@nankai.edu.cn
2CRCV, University of Central Florida, Orlando, Florida, US
http://mmcheng.net/SOCBenchmark/

Abstract. We provide a comprehensive evaluation of salient object detection (SOD) models. Our analysis identi■es a serious design bias of existing SOD datasets which assumes that each image contains atleast one clea rly outstanding salient object in low clutter. The design bias has led to a saturated high performance for state-of-the-art SOD models when evaluated on existing datasets. The models, however, stillperform fa r from being satisfactory when applied to real-world daily scenes. Based on our analyses, we ■rst identify 7 crucial aspects that a comprehensive and balanced dataset should ful■ll. Then, we propose anew high qua lity dataset and update the previous saliency benchmark. Speci■cally, our SOC (Salient Objects in Clutter) dataset, includes images with salient and non-salient objects from daily object categories.Beyond object category annotations, each salient image is accompanied by attributes that re■ect common challenges in real-world scenes. Finally, we report attribute-based performance assessment on our dataset.

Keywords: Salient object detection
·Saliency benchmark
Dataset ·Attribute

1

**********

Gray-Box Adversarial Training

B. S. Vivek(B), Konda Reddy Mopuri , and R. Venkatesh Babu
Indian Institute of Science, Bangalore, India
svivek@iisc.ac.in ,kondamopuri@iisc.ac.in ,venky@iisc.ac.in

Abstract. Adversarial samples are perturbed inputs crafted to mislead the machine learning systems. A training mechanism, called adversarial training, which presents adversarial samples along with clean sampleshas been in troduced to learn robust models. In order to scale adversarial training for large datasets, these perturbations can only be crafted using fast and simple methods (e.g., gradient ascent). However, it is shownthat advers arial training converges to a degenerate minimum, where the model appears to be robust by generating weaker adversaries. As a result, the models are vulnerable to simple black-box attacks.
In this paper we, (i) demonstrate the shortcomings of existing eval-uation policy, (ii) introduce novel variants of white-box and black-boxattacks, dubbed “gray-box adversarial attacks” based on which we pro-pose novel evaluation method to assess the robustness of the learned models, and (iii) propose a novel variant of adversarial training, named“Gray-bo x Adversarial Training” that uses intermediate versions of the models to seed the adversaries. Experimental evaluation demonstrates that the models trained using our method exhibit better robustness com-pared to both undefended and adversarially trained models.

Keywords: Adversarial perturbations
Attacks on machine learning models

·Adversarial training
Robust machine learning models
1
**********
Visual Question Answering as a Meta
Learning Task
Damien Teney(B)and Anton van den Hengel
Australian Institute for Machine Learning, University of Adelaide,
Adelaide, Australia
{damien.teney,anton.vandenhengel }@adelaide.edu.au
Abstract. The predominant approach to Visual Question Answering
(VQA) demands that the model represents within its weights all of the
information required to answer any question about any image. Learningthis inform
ation from any real training set seems unlikely, and represent-
ing it in a reasonable number of weights doubly so. We propose instead
to approach VQA as a meta learning task, thus separating the ques-tion answering
 method from the information required. At test time, the
method is provided with a support set of example questions/answers,
over which it reasons to resolve the given question. The support set isnot ■xed
and can be extended without retraining, thereby expanding the
capabilities of the model. To exploit this dynamically provided informa-
tion, we adapt a state-of-the-art VQA model with two techniques fromthe recent m
eta learning literature, namely prototypical networks and
meta networks. Experiments demonstrate the capability of the system to
learn to produce completely novel answers ( i.e. never seen during train-
ing) from examples provided at test time. In comparison to the existing
state of the art, the proposed method produces qualitatively distinct
results with higher recall of rare answers, and a better sample e■ciencythat all
ows training with little initial data. More importantly, it repre-
sents an important step towards vision-and-language methods that can
learn and reason on-the-■y.
The task of Visual Question Answering (VQA) demands that an agent correctly
answer a previously unseen question about a previously unseen image. The fact
that neither the question nor the image is speci■ed until test time means that t
heagent must embody most of the achievements of Computer Vision and Natural
Language Processing, and many of those of Arti■cial Intelligence.
VQA is typically framed in a purely supervised learning setting. A large
training set of example questions, images, and their correct answers is used to
train a method to map a question and image to scores over a predetermined,
■xed vocabulary of possible answers using the maximum likelihood [ 39]. This
approach has inherent scalability issues, as it attempts to represent all world
Electronic supplementary material The online version of this chapter ( https://
doi.org/10.1007/978-3-030-01267-0 14) contains supplementary material, which is
available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11219, pp. 229–245, 2018.https://doi.o
rg/10.1007/978-3-030-01267-0
_1
**********
On O■ine Evaluation of Vision-Based
Driving Models
Felipe Codevilla1, Antonio M. L´ opez1, Vladlen Koltun2,
and Alexey Dosovitskiy3(B)
1Computer Vision Center, Universitat Aut` onoma de Barcelona, Barcelona, Spain
fcodevilla@cvc.uab.es
2Intel Labs, Santa Clara, USA
3Intel Labs, Munich, Germany
adosovitskiy@gmail.com
Abstract. Autonomous driving models should ideally be evaluated by
deploying them on a ■eet of physical vehicles in the real world. Unfortu-nately,

this approach is not practical for the vast majority of researchers.
An attractive alternative is to evaluate models o■ine, on a pre-collected
validation dataset with ground truth annotation. In this paper, we inves-
tigate the relation between various online and o■ine metrics for evalua-
tion of autonomous driving models. We ■nd that o■ine prediction erroris not nece
ssarily correlated with driving quality, and two models with
identical prediction error can di■er dramatically in their driving perfor-
mance. We show that the correlation of o■ine evaluation with drivingquality can
be signi■cantly improved by selecting an appropriate valida-
tion dataset and suitable o■ine metrics.
Keywords: Autonomous driving
·Deep learning
1
**********

Visual Psychophysics for Making Face
Recognition Algorithms More Explainable
Brandon RichardWebster1(B), So Yon Kwon2, Christopher Clarizio1,
Samuel E. Anthony2,3, and Walter J. Scheirer1
1University of Notre Dame, Notre Dame, IN 46556, USA
brichar1@nd.edu
2Perceptive Automata, Inc., Somerville, USA
3Harvard University, Cambridge, MA 02138, USA
Abstract. Scienti■c ■elds that are interested in faces have developed
their own sets of concepts and procedures for understanding how a tar-get model
system (be it a person or algorithm) perceives a face under
varying conditions. In computer vision, this has largely been in the form
of dataset evaluation for recognition tasks where summary statistics areused to
measure progress. While aggregate performance has continued
to improve, understanding individual causes of failure has been di■cult,
as it is not always clear why a particular face fails to be recognized, orwhy an
impostor is recognized by an algorithm. Importantly, other ■elds
studying vision have addressed this via the use of visual psychophysics:
the controlled manipulation of stimuli and careful study of the responsesthey ev
oke in a model system. In this paper, we suggest that visual
psychophysics is a viable methodology for making face recognition algo-
rithms more explainable. A comprehensive set of procedures is developedfor asses
sing face recognition algorithm behavior, which is then deployed
over state-of-the-art convolutional neural networks and more basic, yet
still widely used, shallow and handcrafted feature-based approaches.
Keywords: Face recognition
·Biometrics ·Explainable AI
Visual psychophysics ·Biometric menagerie
1
**********

Conditional Prior Networks for Optical
Flow
Yanchao Yang(B)and Stefano Soatto
UCLA Vision Lab, University of California, Los Angeles, CA 90095, USA
{yanchao.yang,soatto }@cs.ucla.edu
Abstract. Classical computation of optical ■ow involves generic pri-
ors (regularizers) that capture rudimentary statistics of images, but notlong-ra
nge correlations or semantics. On the other hand, fully supervised
methods learn the regularity in the annotated data, without explicit reg-
ularization and with the risk of over■tting. We seek to learn richer priorson th
e set of possible ■ows that are statistically compatible with an
image. Once the prior is learned in a supervised fashion, one can eas-
ily learn the full map to infer optical ■ow directly from two or more
images, without any need for (additional) supervision. We introduce a
novel architecture, called Conditional Prior Network (CPN), and show
how to train it to yield a conditional prior. When used in conjunctionwith a sim

ple optical ■ow architecture, the CPN beats all variational methods and all unsupervised learning-based ones using the same data term. It performs comparably to fully supervised ones, that however are ■ne-tuned to a particular dataset. Our method, on the other hand, performs well even when transferred between datasets. Code is available at:https://github.com/YanchaoYang/Conditional-Prior-Networks .



**********

Robust Optical Flow in Rainy Scenes

Ruoteng Li1(B), Robby T. Tan1,2, and Loong-Fah Cheong1
1National University of Singapore, Singapore, Singapore
liruoteng@gmail.com
2Yale-NUS College, Singapore, Singapore

Abstract. Optical ■ow estimation in rainy scenes is challenging due to degradation caused by rain streaks and rain accumulation, where thelatter refers to the poor visibility of remote scenes due to intense rainfall. To resolve the problem, we introduce a residue channel, a single channel (gray) image that is free from rain, and its colored version, acolored-residue image. We propose to utilize these two rain-free images in computing optical ■ow. To deal with the loss of contrast and the attendant sensitivity to noise, we decompose each of the input imagesinto a piecewise-smooth structure layer and a high-frequency ■ne-detail texture layer. We combine the colored-residue images and structure layers in a uni■ed objective function, so that the estimation of optical ■ow can be more robust. Results on both synthetic and real images show that our algorithm outperforms existing methods on di■erent types ofrain sequences. To our knowledge, this is the ■rst optical ■ow method speci■cally dealing with rain. We also provide an optical ■ow dataset consisting of both synthetic and real rain images.

Keywords: Optical ■ow
·Rain·Decomposition ·Residue channel



**********

Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-o■s in Video Classi■cation

Saining Xie1,2, Chen Sun1(B), Jonathan Huang1, Zhuowen Tu1,2, and Kevin Murphy1
1Google Research, Mountain View, USA
chensun@google.com
2University of California San Diego, San Diego, USA

Abstract. Despite the steady progress in video analysis led by the adoption of convolutional neural networks (CNNs), the relative improvement has been less drastic as that in 2D static image classi■cation. Three mainchallenges exist including spatial (image) feature representation, temporal information representation, and model/computation complexity. It was recently shown by Carreira and Zisserman that 3D CNNs, in■atedfrom 2D networks and pretrained on ImageNet, could be a promising way for spatial and temporal representation learning. However, as for model/computation complexity, 3D CNNs are much more expensive than2D CNNs and prone to over■t. We seek a balance between speed and accuracy by building an e■ective and e■cient video classi■cation system through systematic exploration of critical network design choices. Inparticular, we show that it is possible to replace many of the 3D convolutions by low-cost 2D convolutions. Rather surprisingly, best result (in both speed and accuracy) is achieved when replacing the 3D convolutionsat the bottom of the network, suggesting that temporal representation learning on high-level “semantic” features is more useful. Our conclusion generalizes to datasets with very di■erent properties. When com-bined with several other cost-e■ective designs including separable spa-

tial/temporal convolution and feature gating, our system results in an e■ective video classi■cation system that that produces very competitive results on several action classi■cation benchmarks (Kinetics, Something-something, UCF101 and HMDB), as well as two action detection (local-ization) ben chmarks (JHMDB and UCF101-24).
1
**********
Variational Wasserstein Clustering
Liang Mi1(B), Wen Zhang1, Xianfeng Gu2, and Yalin Wang1
1Arizona State University, Tempe, USA
{liangmi,wzhan139 }@asu.edu
ylwang@asu.edu
2Stony Brook University, Stony Brook, USA
gu@cs.stonybrook.edu
Abstract. We propose a new clustering method based on optimal trans-portation. We discuss the connection between optimal transportation and k-means clustering, solve optimal transportation with the variational principle, and investigate the use of power diagrams as transportationplans for aggregating arbitrary domains into a ■xed number of clusters. We drive cluster centroids through the target domain while maintaining the minimum clustering energy by adjusting the power diagram. Thus, we simultaneously pursue clustering and the Wasserstein distance between the centroids and the target domain, resulting in a measure-preservingmapping. W e demonstrate the use of our method in domain adaptation, remeshing, and learning representations on synthetic and real data.
Keywords: Clustering
·Discrete distribution ·K-means
Measure preserving ·Optimal transportation ·Wasserstein distance
1
**********
Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data
Xihui Liu1, Hongsheng Li1(B), Jing Shao2, Dapeng Chen1, and Xiaogang Wang1
1The Chinese University of Hong Kong, Hong Kong, China
{hsli,xgwang }@ee.cuhk.edu.hk, xihui-liu@link.cuhk.edu.hk, dpchen@cuhk.edu.hk
2SenseTime Research, Hong Kong, China
shaojing@sensetime.com
Abstract. The aim of image captioning is to generate captions by machine to describe image contents. Despite many e■orts, generating discriminative captions for images remains non-trivial. Most traditional approaches imitate the language structure patterns, thus tend to fallinto a ster eotype of replicating frequent phrases or sentences and neglect unique aspects of each image. In this work, we propose an image caption-ing framework with a self-retrieval module as training guidance, whichencourages generating discriminative captions. It brings unique advan-tages: (1) the self-retrieval guidance can act as a metric and an evaluator of caption discriminativeness to assure the quality of generated captions.(2) Th e correspondence between generated captions and images are natu-rally incorporated in the generation process without human annotations, and hence our approach could utilize a large amount of unlabeled imagesto boost captioning performance with no additional annotations. We demonstrate the e■ectiveness of the proposed retrieval-guided method on COCO and Flickr30k captioning datasets, and show its superior cap-tioning per formance with more discriminative captions.
Keywords: Image captioning
·Language and vision
Text-image retrieval

1
**********
Contour Knowledge Transfer for Salient
Object Detection
Xin Li1,F a nY a n g1(B), Hong Cheng1, Wei Liu1, and Dinggang Shen2
1University of Electronic Science and Technology of China, Chengdu 611731, China
xinli uestc@hotmail.com ,fanyang uestc@hotmail.com ,hcheng@uestc.edu.cn
2Department of Radiology and BRIC, University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599, USA
dgshen@med.unc.edu
Abstract. In recent years, deep Convolutional Neural Networks (CNNs)
have broken all records in salient object detection. However, training sucha dee
p model requires a large amount of manual annotations. Our goal is
to overcome this limitation by automatically converting an existing deep
contour detection model into a salient object detection model withoutusing any m
anual salient object masks. For this purpose, we have created
a deep network architecture, namely Contour-to-Saliency Network (C2S-
Net), by grafting a new branch onto a well-trained contour detection
network. Therefore, our C2S-Net has two branches for performing two
di■erent tasks: (1) predicting contours with the original contour branch,and (2)
 estimating per-pixel saliency score of each image with the newly-
added saliency branch. To bridge the gap between these two tasks, we
further propose a contour-to-saliency transferring method to automati-cally gene
rate salient object masks which can be used to train the saliency
branch from outputs of the contour branch. Finally, we introduce a novel
alternating training pipeline to gradually update the network parame-ters. In th
is scheme, the contour branch generates saliency masks for
training the saliency branch, while the saliency branch, in turn, feeds
back saliency knowledge in the form of saliency-aware contour labels,for ■ne-tun
ing the contour branch. The proposed method achieves state-
of-the-art performance on ■ve well-known benchmarks, outperforming
existing fully supervised methods while also maintaining high e■ciency.
Keywords: Saliency detection
·Deep learning ·Transfer learning
1
**********
Learning Category-Speci■c Mesh
Reconstruction from Image Collections
Angjoo Kanazawa(B), Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik
University of California, Berkeley, USA
{kanazawa,shubhtuls,efros,malik }@eecs.berkeley.edu
Abstract. We present a learning framework for recovering the 3D
shape, camera, and texture of an object from a single image. The shape isreprese
nted as a deformable 3D mesh model of an object category where
a shape is parameterized by a learned mean shape and per-instance pre-
dicted deformation. Our approach allows leveraging an annotated image
collection for training, where the deformable model and the 3D predic-
tion mechanism are learned without relying on ground-truth 3D or multi-view supe
rvision. Our representation enables us to go beyond existing 3D
prediction approaches by incorporating texture inference as prediction of
an image in a canonical appearance space. Additionally, we show thatsemantic key
points can be easily associated with the predicted shapes.
We present qualitative and quantitative results of our approach on CUB
and PASCAL3D datasets and show that we can learn to predict diverseshapes and te
xtures across objects using only annotated image collec-
tions. The project website can be found at https://akanazawa.github.io/
cmr/.
1
**********
Learning to Forecast and Re■ne Residual

Motion for Image-to-Video Generation
Long Zhao1(B), Xi Peng2, Yu Tian1, Mubbasir Kapadia1,
and Dimitris Metaxas1
1Rutgers University, Piscataway, USA
{lz311,yt219,mk1353,dnm }@cs.rutgers.edu
2Binghamton University, Binghamton, USA
xpeng@binghamton.edu
Abstract. We consider the problem of image-to-video translation,
where an input image is translated into an output video containing
motions of a single object. Recent methods for such problems typically
train transformation networks to generate future frames conditioned onthe struct
ure sequence. Parallel work has shown that short high-quality
motions can be generated by spatiotemporal generative networks that
leverage temporal knowledge from the training data. We combine thebene■ts of bot
h approaches and propose a two-stage generation frame-
work where videos are generated from structures and then re■ned by
temporal signals. To model motions more e■ciently, we train networksto learn res
idual motion between the current and future frames, which
avoids learning motion-irrelevant details. We conduct extensive experi-
ments on two image-to-video translation tasks: facial expression retarget-ing an
d human pose forecasting. Superior results over the state-of-the-art
methods on both tasks demonstrate the e■ectiveness of our approach.
Keywords: Video generation
·Motion forecasting ·Residual learning

**********
Teaching Machines to Understand
Baseball Games: Large-Scale Baseball
Video Database for Multiple Video
Understanding Tasks
Minho Shim , Young Hwi Kim , Kyungmin Kim , and Seon Joo Kim(B)
Yonsei University, Seoul, South Korea
{minhoshim,younghwikim,kyungminkim,seonjookim }@yonsei.ac.kr
Abstract. A major obstacle in teaching machines to understand videos
is the lack of training data, as creating temporal annotations for longvideos re
quires a huge amount of human e■ort. To this end, we introduce
a new large-scale baseball video dataset called the BBDB, which is pro-
duced semi-automatically by using play-by-play texts available online.The BBDB c
ontains 4200+hr of baseball game videos with 400k+ tem-
porally annotated activity segments. The new dataset has several major
challenging factors compared to other datasets: (1) the dataset containsa large
number of visually similar segments with di■erent labels. (2) It
can be used for many video understanding tasks including video recogni-
tion, localization, text-video alignment, video highlight generation, anddata im
balance problem. To observe the potential of the BBDB, we con-
ducted extensive experiments by running many di■erent types of video
understanding algorithms on our new dataset. The database is availableathttps://
sites.google.com/site/eccv2018bbdb/ .
Keywords: Video understanding
·Large-scale video dataset
Action recognition ·Temporal localization

**********
SketchyScene: Richly-Annotated Scene
Sketches
Changqing Zou1, Qian Yu2,R u o f e iD u1, Haoran Mo3, Yi-Zhe Song2,
Tao Xiang2, Chengying Gao3, Baoquan Chen4(B), and Hao Zhang5
1University of Maryland, College Park, US
2Queen Mary University of London, London, UK
3Sun Yat-sen University, Guangzhou, China

4Shandong University, Jinan, China
baoquan@sdu.edu.cn
5Simon Fraser University, Burnaby, Canada
Abstract. We contribute the ■rst large-scale dataset of scene sketches , SketchyScene , with the goal of advancing research on sketch understanding at both the object and scene level. The dataset is created through a novel and carefully designed crowdsourcing pipeline, enabling users to e■ciently generate large quantities of realistic and diverse scene sketches. SketchyScene contains more than 29,000 scene-level sketches, 7,000+ pairs of scene templates and photos, and 11,000+ object sketches.All objects in the scene sketches have ground-truth semantic and instance masks. The dataset is also highly scalable and extensible, easily allowing augmenting and/or changing scene composition. We demonstratethe potential impact of SketchyScene by training new computational models for semantic segmentation of scene sketches and showing how the new dataset enables several applications including image retrieval,sketch colorization, editing, and captioning, etc. The dataset and code can be found at https://github.com/SketchyScene/SketchyScene .
Keywords: Sketch dataset ·Scene sketch ·Sketch segmentation

TU-Berlin [ 9] is the ■rst such large-scale crowd-sourced sketch dataset which was primarily designed for sketch recognition. It consist of 20,000 sketches spanning over 250 categories. The more recent QuickDraw[ 10] dataset is much larger, with 50 million sketches across 345 categories. Albeit being large enough to facilitate stroke-level analysis[ 6], sketches sourced in these datasets were produced by sketching towards a semantic concept (e.g., “cat”, “house”), without a reference photo or mental recollection of natural scene/objects. This greatly limits the level of visual detail and variations depicted, therefore making them un■ttingfor ■ne-grained matching and scene-level parsing. For example, faces are almost all in their frontal view, and depicted as a smiley in QuickDraw.

The concurrent work of [ 36] and [ 19] progressed the ■eld further by collecting object instance sketches for FG-SBIR. QMUL database [ 36] consists of 716 sketch-photo pairs across two object categories (shoe and chair), with reference photos crawled from on-line shopping websites. Sketchy [ 19] contains 75,471 sketches and 12,500 corresponding photos across a much wider selection of categories (125 in total). Object instance sketches are produced by asking crowd-sources to depict their mental recollection of a reference photo. In comparison with concept sketches [ 9,10], they by and large exhibit more object details and have matching poses with the reference photos. However, a common drawbackfor both, for the purpose this project, lies with their limited pose selection and object con■gurations. QMUL sketches exhibit only one object pose (side view) under a single object con■guration. Scene sketches albeit exhibits more objectposes and con■gurations, are still restricted since their reference photos mainly consists of single objects centered on relatively plain backgrounds (thus depicts no object interactions). This drawback essentially renders them both unsuitablefor our task of scene sketch parsing, where complex mutual object interactions dictate high degree of object pose and con■guration variations, as well as subtle details. For example, within a picnic scene depicted in Figure 1, people appear in di■erent poses and con■gurations with subtle eye contacts among each other. Fig.2(c) shows a composition result using sketches from Sketchy and TU-Berlin.

SketchyScene is the ■rst large-scale dataset speci■cally designed for scene-level sketch understanding. It di■ers from all aforementioned datasets in that it goes beyond single object sketch understanding to tackle scene sketch, andpurposefully includes an assorted selection of object sketches with diverse poses, con■gurations and object details to accommodate the complex scene-level object interactions. Although the existing dataset Abstract Scenes [ 38] serves a similar
motivation for understanding high-level semantic information in visual data, they focus on abstract scenes composed using clip arts, which include much more visual cues such as color and texture. In addition, their scenes are restricted indescribing interactions between two characters and a handful of objects, while the scene contents and mutual object interactions in SketchyScene are a lot more diverse.

2.2 Sketch Understanding

Sketch recognition is perhaps the most studied problem in sketch understanding. Since the release of TU-Berlin dataset [ 9], many works have been proposed an

**********

Learn-to-Score: E■cient 3D Scene
Exploration by Predicting View Utility

Benjamin Hepp1,2(B), Debadeepta Dey2, Sudipta N. Sinha2, Ashish Kapoor2,
Neel Joshi2, and Otmar Hilliges1
1ETH Zurich, Zurich, Switzerland
benjamin.hepp@inf.ethz.ch
2Microsoft Research, Redmond, USA

Abstract. Camera equipped drones are nowadays being used to explore large scenes and reconstruct detailed 3D maps. When free space in the scene is approximately known, an o■ine planner can generate optimalplans to e■ciently explore the scene. However, for exploring unknown
scenes, the planner must predict and maximize usefulness of where to go on the ■y. Traditionally, this has been achieved using handcrafted utility functions. We propose to learn a better utility function that predicts the usefulness of future viewpoints. Our learned utility function is based ona 3D convolutional neural network. This network takes as input a novel volumetric scene representation that implicitly captures previously visited viewpoints and generalizes to new scenes. We evaluate our methodon several large 3D models of urban scenes using simulated depth cameras. We show that our method outperforms existing utility measures in terms of reconstruction performance and is robust to sensor noise.

Keywords: 3D reconstruction
·Exploration ·Active vision ·3D CNN



**********

Revisiting RCNN: On Awakening
the Classi■cation Power of Faster RCNN

Bowen Cheng1, Yunchao Wei1(B), Honghui Shi2, Rogerio Feris2,
Jinjun Xiong2, and Thomas Huang1
1University of Illinois at Urbana-Champaign, Champaign, IL, USA
{bcheng9,yunchao,t-huang1 }@illinois.edu
2IBM T.J. Watson Research Center, Ossining, NY, USA
Honghui.Shi@ibm.com, {rsferis,jinjun }@us.ibm.com

Abstract. Recent region-based object detectors are usually built with separate classi■cation and localization branches on top of shared feature extraction networks. In this paper, we analyze failure cases of state-of-the-art detectors and observe that most hard false positives result from classi■cation instead of localization. We conjecture that: (1) Shared feature representation is not optimal due to the mismatched goals of feature learning for classi■cation and localization; (2) multi-task learning helps, yet optimization of the multi-task loss may result in sub-optimal for individual tasks; (3) large receptive ■eld for di■erent scales leadsto redundant context information for small objects. We demonstrate

the potential of detector classi■cation power by a simple, e■ective, and widely-applicable Decoupled Classi■cation Re■nement (DCR) network. DCR samples hard false positives from the base classi■er in Faster RCNN and trains a RCNN-styled strong classi■er. Experiments show new state-of-the-art results on PASCAL VOC and COCO without any bells andwhistles.
Keywords: Object detection
1
**********

Semi-supervised Generative Adversarial Hashing for Image Retrieval

Guan'an Wang1,3, Qinghao Hu2,3, Jian Cheng2,3,4, and Zengguang Hou1,3,4(B)
1The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{wangguanan2015,zengguang.hou }@ia.ac.cn
2National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{qinghao.hu,jcheng }@nlpr.ia.ac.cn
3University of Chinese Academy of Sciences, Beijing, China
4Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

Abstract. With explosive growth of image and video data on the Internet, hashing technique has been extensively studied for large-scale visual search. Bene■ting from the advance of deep learning, deep hashing methods have achieved promising performance. However, those deep hashingmodels are usually trained with supervised information, which is rare and expensive in practice, especially class labels. In this paper, inspired by the idea of generative models and the minimax two-player game, wepropose a novel semi-supervised generative adversarial hashing (SSGAH) approach. Firstly, we unify a generative model, a discriminative model and a deep hashing model in a framework for making use of triplet-wise information and unlabeled data. Secondly, we design novel structure of the generative model and the discriminative model to learn the distri-bution of triplet-wise information in a semi-supervised way. In addition, we propose a semi-supervised ranking loss and an adversary ranking loss to learn binary codes which preserve semantic similarity for bothlabeled data and unlabeled data. Finally, by optimizing the whole model in an adversary training way, the learned binary codes can capture better semantic information of all data. Extensive empirical evaluations ontwo widely-used benchmark datasets show that our proposed approach signi■cantly outperforms state-of-the-art hashing methods.
Keywords: Information retrieval
·Hashing ·Deep learning ·GANs
1
**********

Person Re-identi■cation with Deep Similarity-Guided Graph Neural Network

Yantao Shen1, Hongsheng Li1(B),S h u a iY i2, Dapeng Chen1, and Xiaogang Wang1
1CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong, Hong Kong, People's Republic of China
{ytshen,hsli,dpchen }@ee.cuhk.edu.hk
2SenseTime Research, Hong Kong, People's Republic of China
yishuai@sensetime.com

Abstract. The person re-identi■cation task requires to robustly estimate visual similarities between person images. However, existing person re-identi■cation models mostly estimate the similarities of di■erentimage pairs of probe and gallery images independently while ignores the relationship information between di■erent probe-gallery pairs. As a result, the similarity estimation of some hard samples might not be accu-rate. I

n this paper, we propose a novel deep learning framework, named Similarity-Guided Graph Neural Network (SGGNN) to overcome such limitations. Given a probe image and several gallery images, SGGNNcreates a graph to represent the pairwise relationships between probe-gallery pairs (nodes) and utilizes such relationships to update the probe-gallery relation features in an end-to-end manner. Accurate similarityestimation can be achieved by using such updated probe-gallery relation features for prediction. The input features for nodes on the graph are the relation features of di■erent probe-gallery image pairs. The probe-galleryrelation feature updating is then performed by the messages passing in SGGNN, which takes other nodes' information into account for similarity estimation. Di■erent from conventional GNN approaches, SGGNN learnsthe edge weights with rich labels of gallery instance pairs directly, which provides relation fusion more precise information. The e■ectiveness of our proposed method is validated on three public person re-identi■cationdatasets.

Keywords: Deep learning
·Person re-identi■cation
Graph Neural Networks

1

**********

## Learning and Matching Multi-View Descriptors for Registration of Point Clouds

Lei Zhou1,S i y uZ h u1, Zixin Luo1, Tianwei Shen1, Runze Zhang1(B), Mingmin Zhen1, Tian Fang2, and Long Quan1
1Hong Kong University of Science and Technology, Hong Kong, China
{lzhouai,szhu,zluoag,tshenaa,rzhangaj,mzhen,quan }@cse.ust.hk
2Shenzhen Zhuke Innovation Technology (Altizure), Shenzhen, China
fangtian@altizure.com

Abstract. Critical to the registration of point clouds is the establishment of a set of accurate correspondences between points in 3D space. The correspondence problem is generally addressed by the design of dis-criminative 3D local descriptors on the one hand, and the development of robust matching strategies on the other hand. In this work, we ■rst propose a multi-view local descriptor, which is learned from the imagesof multiple views, for the description of 3D keypoints. Then, we develop a robust matching approach, aiming at rejecting outlier matches based on the e■cient inference via belief propagation on the de■ned graphicalmodel. We have demonstrated the boost of our approaches to registration on the public scanning and multi-view stereo datasets. The superior performance has been veri■ed by the intensive comparisons against a variety of descriptors and matching methods.

Keywords: Point cloud registration
·3D descriptor ·Robust matching

1

**********

## Revisiting Autofocus for Smartphone Cameras

Abdullah Abuolaim(B), Abhijith Punnappurath, and Michael S. Brown
Department of Electrical Engineering and Computer Science,
Lassonde School of Engineering, York University, Toronto, Canada
{abuolaim,pabhijith,mbrown }@eecs.yorku.ca

Abstract. Autofocus (AF) on smartphones is the process of determining how to move a camera's lens such that certain scene content is in focus. The underlying algorithms used by AF systems, such as contrast detection and phase di■erencing, are well established. However, deter-mining a high-level objective regarding how to best focus a particular scene is less clear. This is evident in part by the fact that di■erent smartphone cameras employ di■erent AF criteria; for example, some attemptto keep item

s in the center in focus, others give priority to faces while
others maximize the sharpness of the entire scene. The fact that di■er-
ent objectives exist raises the research question of whether there is apreferred
objective. This becomes more interesting when AF is applied
to videos of dynamic scenes. The work in this paper aims to revisit AF
for smartphones within the context of temporal image data. As part ofthis e■ort,
we describe the capture of a new 4D dataset that provides
access to a full focal stack at each time point in a temporal sequence.
Based on this dataset, we have developed a platform and associatedapplication pr
ogramming interface (API) that mimic real AF systems,
restricting lens motion within the constraints of a dynamic environment
and frame capture. Using our platform we evaluated several high-levelfocusing ob
jectives and found interesting insight into what users pre-
fer. We believe our new temporal focal stack dataset, AF platform, and
initial user-study ■ndings will be useful in advancing AF research.
Keywords: Autofocus
·Focal stack ·AF platform
Low-level computer vision
1
**********

Deep Burst Denoising
Cl´ement Godard1(B), Kevin Matzen2, and Matt Uyttendaele2
1University College London, London, UK
c.godard@cs.ucl.ac.uk
2Facebook, Seattle, USA
Abstract. Noise is an inherent issue of low-light image capture, which
is worsened on mobile devices due to their narrow apertures and smallsensors. On
e strategy for mitigating noise in low-light situations is to
increase the shutter time, allowing each photosite to integrate more light
and decrease noise variance. However, there are two downsides of longexposures:
(a) bright regions can exceed the sensor range, and (b) camera
and scene motion will cause blur. Another way of gathering more light is
to capture multiple short (thus noisy) frames in a burst and intelligentlyintegr
ate the content, thus avoiding the above downsides. In this paper,
we use the burst-capture strategy and implement the intelligent integra-
tion via a recurrent fully convolutional deep neural net (CNN). We build
our novel, multi-frame architecture to be a simple addition to any single
frame denoising model. The resulting architecture denoises all frames ina sequen
ce of arbitrary length. We show that it achieves state of the
art denoising results on our burst dataset, improving on the best pub-
lished multi-frame techniques, such as VBM4D and FlexISP. Finally, weexplore oth
er applications of multi-frame image enhancement and show
that our CNN architecture generalizes well to image super-resolution.
1
**********

ISNN: Impact Sound Neural Network
for Audio-Visual Object Classi■cation
Auston Sterling1, Justin Wilson1(B), Sam Lowe1, and Ming C. Lin1,2
1Department of Computer Science, University of North Carolina at Chapel Hill,
Chapel Hill, USA
{austonst,wilson,samlowe,lin }@cs.unc.edu
2Department of Computer Science, University of Maryland, College Park, USA
lin@cs.umd.edu
Abstract. 3D object geometry reconstruction remains a challenge when
working with transparent, occluded, or highly re■ective surfaces. Whilerecent me
thods classify shape features using raw audio, we present a mul-
timodal neural network optimized for estimating an object’s geometry
and material. Our networks use spectrograms of recorded and synthe-sized object
impact sounds and voxelized shape estimates to extend the
capabilities of vision-based reconstruction. We evaluate our method on

multiple datasets of both recorded and synthesized sounds. We furtherpresent an interactive application for real-time scene reconstruction in which a user can strike objects, producing sound that can instantly classify and segment the struck object, even if the object is transparent orvisually occluded.

1

**********

StereoNet: Guided Hierarchical Re■nement for Real-Time Edge-Aware Depth Prediction

Sameh Khamis(B), Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi
Google Inc., Mountain View, USA
sameh@google.com

Abstract. This paper presents StereoNet, the ■rst end-to-end deep architecture for real-time stereo matching that runs at 60fps on an NVidia Titan X, producing high-quality, edge-preserved, quantization-free disparity maps. A key insight of this paper is that the network achieves a sub-pixel matching precision than is a magnitude higher than those of traditional stereo matching approaches. This allows us to achievereal-time performance by using a very low resolution cost volume that encodes all the information needed to achieve high disparity precision. Spatial precision is achieved by employing a learned edge-aware upsam-pling function. Our model uses a Siamese network to extract features from the left and right image. A ■rst estimate of the disparity is computed in a very low resolution cost volume, then hierarchically the modelre-introduces high-frequency details through a learned upsampling function that uses compact pixel-to-pixel re■nement networks. Leveraging color input as a guide, this function is capable of producing high-qualityedge-aware output. We achieve compelling results on multiple benchmarks, showing how the proposed method o■ers extreme ■exibility at an acceptable computational budget.

Keywords: Stereo matching ·Depth estimation Edge-aware re■nement ·Cost volume ■ltering ·Deep learning

1

**********

Attention-Aware Deep Adversarial Hashing for Cross-Modal Retrieval

Xi Zhang1,2, Hanjiang Lai1,2(B), and Jiashi Feng3
1School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China
zhangx368@mail2.sysu.edu.cn,
laihanj3@mail.sysu.edu.cn
2Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China
3Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
elefjia@nus.edu.sg

Abstract. Due to the rapid growth of multi-modal data, hashing methods for cross-modal retrieval have received considerable attention. However, ■nding content similarities between di■erent modalities of data is still challenging due to an existing heterogeneity gap. To further addressthis problem, we propose an adversarial hashing network with an attention mechanism to enhance the measurement of content similarities by selectively focusing on the informative parts of multi-modal data. The proposed new deep adversarial network consists of three building blocks: (1) the feature learning module to obtain the feature representations; (2) theattention module to generate an attention mask, which is used to divide the feature representations into the attended and unattended feature representations; and (3) the hashing module to learn hash functions thatpreserve th

e similarities between di■erent modalities. In our framework,
the attention and hashing modules are trained in an adversarial way: the
attention module attempts to make the hashing module unable to pre-serve the sim
ilarities of multi-modal data w.r.t. the unattended feature
representations, while the hashing module aims to preserve the similari-
ties of multi-modal data w.r.t. the attended and unattended feature repre-sentat
ions. Extensive evaluations on several benchmark datasets demon-
strate that the proposed method brings substantial improvements over
other state-of-the-art cross-modal hashing methods.
Keywords: Hashing
·Adversarial learning ·Attention mechanism
Cross modal retrieval
1
**********

3DFeat-Net: Weakly Supervised Local 3D
Features for Point Cloud Registration
Zi Jian Yew(B)and Gim Hee Lee
Department of Computer Science, National University of Singapore,
Singapore, Singapore
{zijian.yew,gimhee.lee }@comp.nus.edu.sg
Abstract. In this paper, we propose the 3DFeat-Net which learns both
3D feature detector and descriptor for point cloud matching using weak
supervision. Unlike many existing works, we do not require manual anno-
tation of matching point clusters. Instead, we leverage on alignment and
attention mechanisms to learn feature correspondences from GPS/INS
tagged 3D point clouds without explicitly specifying them. We createtraining and
 benchmark outdoor Lidar datasets, and experiments show
that 3DFeat-Net obtains state-of-the-art performance on these gravity-
aligned datasets.
Keywords: Point cloud
·Registration ·Deep learning
Weak supervision
1
**********

Deep Domain Generalization via
Conditional Invariant Adversarial
Networks
Ya Li1, Xinmei Tian1(B), Mingming Gong2,3, Yajing Liu1, Tongliang Liu4,
Kun Zhang2, and Dacheng Tao4
1CAS Key Laboratory of Technology in Geo-Spatial Information
Processing and Application Systems,
University of Science and Technology of China, Hefei, China
{muziyiye,lyj123 }@mail.ustc.edu.cn, xinmei@ustc.edu.cn
2Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA
gongmingnju@gmail.com, kunz1@cmu.edu
3Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, USA
4UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney, Sydney, Australia
tliang.liu@gmail.com, dacheng.tao@sydney.edu.au
Abstract. Domain generalization aims to learn a classi■cation model
from multiple source domains and generalize it to unseen target domains.
A critical problem in domain generalization involves learning domain-
invariant representations. Let Xand Ydenote the features and the
labels, respectively. Under the assumption that the conditional distri-
bution P(Y|X) remains unchanged across domains, earlier approaches
to domain generalization learned the invariant representation T(X)b y
minimizing the discrepancy of the marginal distribution P(T(X)). How-
ever, such an assumption of stable P(Y|X) does not necessarily hold in
practice. In addition, the representation learning function T(X) is usu-
ally constrained to a simple linear transformation or shallow networks. To
address the above two drawbacks, we propose an end-to-end conditional

invariant deep domain generalization approach by leveraging deep neu-
ral networks for domain-invariant representation learning. The domain-
invariance property is guaranteed through a conditional invariant adver-sarial n
etwork that can learn domain-invariant representations w.r.t. the
joint distribution P(T(X),Y) if the target domain data are not severely
class unbalanced. We perform various experiments to demonstrate thee■ectiveness
of the proposed method.
Keywords: Domain generalization
·Adversarial networks
Domain invariant representation

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11219, pp. 647–663, 2018.https://doi.o
rg/10.1007/978-3-030-01267-0
_3
**********

Using LIP to Gloss Over Faces
in Single-Stage Face Detection Networks

Siqi Yang(B), Arnold Wiliem, Shaokang Chen, and Brian C. Lovell
The University of Queensland, Brisbane, Australia
siqi.yang@uq.net.au ,a.wiliem@uq.edu.au ,shaokangchenuq@gmail.com ,
lovell@itee.uq.edu.au

Abstract. This work shows that it is possible to fool/attack recent
state-of-the-art face detectors which are based on the single-stage net-works. S
uccessfully attacking face detectors could be a serious malware
vulnerability when deploying a smart surveillance system utilizing face
detectors. In addition, for the privacy concern, it helps prevent facesbeing har
vested and stored in the server. We show that existing adver-
sarial perturbation methods are not e■ective to perform such an attack,
especially when there are multiple faces in the inut image. This is becausethe a
dversarial perturbation speci■cally generated for one face may dis-
rupt the adversarial perturbation for another face. In this paper, we call
this problem the Instance Perturbation Interference (IPI) problem. ThisIPI probl
em is addressed by studying the relationship between the deep
neural network receptive ■eld and the adversarial perturbation. Besides
the single-stage face detector, we ■nd that the IPI problem also exists onthe ■r
st stage of the Faster-RCNN, the commonly used two-stage object
detector. As such, we propose the Localized Instance Perturbation (LIP)
that con■nes the adversarial perturbation inside the E■ective ReceptiveField (ER
F) of a target to perform the attack. Experimental results show
the LIP method massively outperforms existing adversarial perturbation
generation methods – often by a factor of 2 to 10.
Keywords: Adversarial
·Interference ·E■ective Receptive Field
Single-stage network ·Detection
1
**********

HiDDeN: Hiding Data With Deep
Networks

Jiren Zhu(B), Russell Kaplan, Justin Johnson, and Li Fei-Fei
Computer Science Department, Stanford University, Stanford, USA
{jirenz,rjkaplan,jcjohns,feifeili }@cs.stanford.edu

Abstract. Recent work has shown that deep neural networks are highly
sensitive to tiny perturbations of input images, giving rise to adversar-
ial examples . Though this property is usually considered a weakness of
learned models, we explore whether it can be bene■cial. We ■nd that
neural networks can learn to use invisible perturbations to encode a richamount
of useful information. In fact, one can exploit this capability for
the task of data hiding. We jointly train encoder and decoder networks,
where given an input message and cover image, the encoder producesa visually ind
istinguishable encoded image, from which the decoder can

recover the original message. We show that these encodings are competitive with existing data hiding algorithms, and further that they can bemade robust to noise: our models learn to reconstruct hidden information in an encoded image despite the presence of Gaussian blurring, pixelwise dropout, cropping, and JPEG compression. Even though JPEG isnon-di■erentiable, we show that a robust model can be trained using di■erentiable approximations. Finally, we demonstrate that adversarial training improves the visual quality of encoded images.

Keywords: Adversarial networks ·Steganography Robust blind watermarking ·Deep learning ·Convolutional networks

1

**********

Multimodal Dual Attention Memory for Video Story Question Answering

Kyung-Min Kim1, Seong-Ho Choi2, Jin-Hwa Kim3, and Byoung-Tak Zhang2,4(B)
1Clova AI Research, NAVER Corp, Seongnam 13561, South Korea
kmkim@bi.snu.ac.kr
2Computer Science and Engineering, Seoul National University, Seoul 08826, South Korea
{shchoi,btzhang }@bi.snu.ac.kr
3SK T-Brain, Seoul 04539, South Korea
jhkim@bi.snu.ac.kr
4Surromind Robotics, Seoul 08826, South Korea

Abstract. We propose a video story question-answering (QA) architecture, Multimodal Dual Attention Memory (MDAM). The key ideais to use a dual attention mechanism with late fusion. MDAM uses self-attention to learn the latent concepts in scene frames and captions. Given a question, MDAM uses the second attention over these latent concepts.Multimodal fusion is performed after the dual attention processes (late fusion). Using this processing pipeline, MDAM learns to infer a high-level vision-language joint representation from an abstraction of the full video content. We evaluate MDAM on PororoQA and MovieQA datasets which have large-scale QA annotations on cartoon videos and movies, respec-tively. For both datasets, MDAM achieves new state-of-the-art results with signi■cant margins compared to the runner-up models. We con■rm the best performance of the dual attention mechanism combined withlate fusion by ablation studies. We also perform qualitative analysis by visualizing the inference mechanisms of MDAM.

Keywords: Video story QA ·Visual QA ·Attention mechanism Multimodal learning ·Deep learning

1

**********

Deep Variational Metric Learning

Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu(B), and Jie Zhou
Tsinghua University, Beijing 100084, China
stein6@163.com, {duanyq14,dqy15 }@mails.tsinghua.edu.cn,
{lujiwen,jzhou }@tsinghua.edu.cn

Abstract. Deep metric learning has been extensively explored recently, which trains a deep neural network to produce discriminative embeddingfeatures. Most existing methods usually enforce the model to be indiscriminating to intra-class variance, which makes the model over-■tting to the training set to minimize loss functions on these speci■c changesand leads to low generalization power on unseen classes. However, these methods ignore a fact that in the central latent space, the distribution of variance within classes is actually independent on classes. In this paper,we propose a deep variational metric learning (DVML) framework to explicitly model the intra-class variance and disentangle the intra-class

invariance, namely, the class centers. With the learned distribution ofintra-cla
ss variance, we can simultaneously generate discriminative sam-
ples to improve robustness. Our method is applicable to most of existing
metric learning algorithms, and extensive experiments on three bench-mark datase
ts including CUB-200-2011, Cars196 and Stanford Online
Products show that our DVML signi■cantly boosts the performance of
currently popular deep metric learning methods.
Keywords: Metric learning
·Variational auto-encoder
Discriminative samples generating
1
**********

HGMR: Hierarchical Gaussian Mixtures
for Adaptive 3D Registration
Benjamin Eckart(B), Kihwan Kim, and Jan Kautz
NVIDIA Research, Santa Clara, CA, USA
beckart@nvidia.com
Abstract. Point cloud registration sits at the core of many important
and challenging 3D perception problems including autonomous naviga-
tion, SLAM, object/scene recognition, and augmented reality. In thispaper, we pr
esent a new registration algorithm that is able to achieve
state-of-the-art speed and accuracy through its use of a Hierarchical
Gaussian Mixture representation. Our method, Hierarchical Gaussian
Mixture Registration (HGMR), constructs a top-down multi-scale rep-
resentation of point cloud data by recursively running many small-scaledata like
lihood segmentations in parallel on a GPU. We leverage the
resulting representation using a novel optimization criterion that adap-
tively ■nds the best scale to perform data association between spatialsubsets of
 point cloud data. Compared to previous Iterative Closest Point
and GMM-based techniques, our tree-based point association algorithm
performs data association in logarithmic-time while dynamically adjust-ing the l
evel of detail to best match the complexity and spatial distri-
bution characteristics of local scene geometry. In addition, unlike other
GMM methods that restrict covariances to be isotropic, our new PCA-based optimiz
ation criterion well-approximates the true MLE solution
even when fully anisotropic Gaussian covariances are used. E■cient data
association, multi-scale adaptability, and a robust MLE approximationproduce an
algorithm that is up to an order of magnitude both faster
and more accurate than current state-of-the-art on a wide variety of 3D
datasets captured from LiDAR to structured light.
1
**********

Bi-Real Net: Enhancing the Performance
of 1-Bit CNNs with Improved
Representational Capability
and Advanced Training Algorithm
Zechun Liu1, Baoyuan Wu2, Wenhan Luo2,X i nY a n g3(B), Wei Liu2,
and Kwang-Ting Cheng1
1Hong Kong University of Science and Technology, Hong Kong, China
zliubq@connect.ust.hk ,timcheng@ust.hk
2Tencent AI Lab, Beijing, China
wubaoyuan1987@gmail.com ,whluo.china@gmail.com ,wliu@ee.columbia.edu
3Huazhong University of Science and Technology, Wuhan, China
xinyang2014@hust.edu.cn
Abstract. In this work, we study the 1-bit convolutional neural net-
works (CNNs), of which both the weights and activations are binary.While being e
■cient, the classi■cation accuracy of the current 1-bit
CNNs is much worse compared to their counterpart real-valued CNN
models on the large-scale dataset, like ImageNet. To minimize the per-formance g
ap between the 1-bit and real-valued CNN models, we propose

a novel model, dubbed Bi-Real net, which connects the real activations
(after the 1-bit convolution and/or BatchNorm layer, before the signfunction) to
activations of the consecutive block, through an identity
shortcut. Consequently, compared to the standard 1-bit CNN, the rep-
resentational capability of the Bi-Real net is signi■cantly enhanced andthe addi
tional cost on computation is negligible. Moreover, we develop
a speci■c training algorithm including three technical novelties for 1-
bit CNNs. Firstly, we derive a tight approximation to the derivative ofthe non-d
i■erentiable sign function with respect to activation. Secondly,
we propose a magnitude-aware gradient with respect to the weight for
updating the weight parameters. Thirdly, we pre-train the real-valuedCNN model w
ith a clip function, rather than the ReLU function, to bet-
ter initialize the Bi-Real net. Experiments on ImageNet show that the
Bi-Real net with the proposed training algorithm achieves 56.4% and62.2% top-1 a
ccuracy with 18 layers and 34 layers, respectively. Com-
pared to the state-of-the-arts ( e.g., XNOR Net), Bi-Real net achieves
up to 10% higher top-1 accuracy with more memory saving and lowercomputational c
ost.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11219, pp. 747–763, 2018.https://doi.o
rg/10.1007/978-3-030-01267-0
_4
**********

Orthogonal Deep Features Decomposition
for Age-Invariant Face Recognition

Yitong Wang , Dihong Gong , Zheng Zhou ,X i n gJ i ,H a oW a n g ,
Zhifeng Li(B), Wei Liu(B), and Tong Zhang
Tencent AI Lab, Beijing, China
{yitongwang,encorezhou,denisji,hawelwang,michaelzfli }@tencent.com,
gongdihong@gmail.com, wl2223@columbia.edu, tongzhang@tongzhang-ml.org

Abstract. As facial appearance is subject to signi■cant intra-class vari-
ations caused by the aging process over time, age-invariant face recogni-
tion (AIFR) remains a major challenge in face recognition community.To reduce th
e intra-class discrepancy caused by the aging, in this paper
we propose a novel approach (namely, Orthogonal Embedding CNNs,
or OE-CNNs) to learn the age-invariant deep face features. Speci■cally,we decomp
ose deep face features into two orthogonal components to
represent age-related and identity-related features. As a result, identity-
related features that are robust to aging are then used for AIFR. Besides,for co
mplementing the existing cross-age datasets and advancing the
research in this ■eld, we construct a brand-new large-scale Cross-Age
Face dataset (CAF). Extensive experiments conducted on the three pub-
lic domain face aging datasets (MORPH Album 2, CACD-VS and FG-
NET) have shown the e■ectiveness of the proposed approach and thevalue of the co
nstructed CAF dataset on AIFR. Benchmarking our algo-
rithm on one of the most popular general face recognition (GFR) dataset
LFW additionally demonstrates the comparable generalization perfor-mance on GFR.

Keywords: Age-invariant face recognition
Convolutional neural networks
·Cross-age face dataset
1
**********

Broadcasting Convolutional Network for
Visual Relational Reasoning

Simyung Chang1,2, John Yang1, SeongUk Park1, and Nojun Kwak1(B)
1Seoul National University, Seoul, South Korea
{yjohn,swpark0703,nojunk }@snu.ac.kr
2Samsung Electronics, Suwon, South Korea
timelighter@snu.ac.kr

Abstract. In this paper, we propose the Broadcasting Convolutional

Network (BCN) that extracts key object features from the global ■eld of an entire input image and recognizes their relationship with localfeatures. BCN is a simple network module that collects e■ective spatial features, embeds location information and broadcasts them to the entire feature maps. We further introduce the Multi-Relational Network (multiRN) that improves the existing Relation Network (RN) by utilizing the BCN module. In pixel-based relation reasoning problems, with the help of BCN, multiRN extends the concept of ‘pairwise relations’ inconventional RNs to ‘multiwise relations’ by relating each object with multiple objects at once. This yields in O(n) complexity for nobjects, which is a vast computational gain from RNs that take O(n 2). Through experiments, multiRN has achieved a state-of-the-art performance on CLEVR dataset, which proves the usability of BCN on relation reasoning problems.

Keywords: Visual relational reasoning ·BCN·Broadcast ·CLEVR Multi-RN ·Visuo-spatial features 1

**********

## Improving Spatiotemporal Self-supervision by Deep Reinforcement Learning

Uta B¨ uchler(B), Biagio Brattoli(B), and Bj¨ orn Ommer
Heidelberg University, HCI/IWR, Heidelberg, Germany
{uta.buechler,biagio.brattoli,bjoern.ommer }@iwr.uni-heidelberg.de

Abstract. Self-supervised learning of convolutional neural networks can harness large amounts of cheap unlabeled data to train powerful feature representations. As surrogate task, we jointly address ordering of visual data in the spatial and temporal domain. The permutations of trainingsamples, which are at the core of self-supervision by ordering, have so far been sampled randomly from a ■xed preselected set. Based on deep reinforcement learning we propose a sampling policy that adapts to thestate of the network, which is being trained. Therefore, new permutations are sampled according to their expected utility for updating the convolutional feature representation. Experimental evaluation on unsu-pervised and transfer learning tasks demonstrates competitive performance on standard benchmarks for image and video classi■cation and nearest neighbor retrieval.

Keywords: Deep reinforcement learning ·Self-supervision ·Shu■ing Action recognition ·Image understanding 1

**********

## Learning to Look around Objects for Top-View Representations of Outdoor Scenes

Samuel Schulter1(B), Menghua Zhai2, Nathan Jacobs2, and Manmohan Chandraker1,3
1NEC-Laboratories, Cupertino, CA 95014, USA
samuel@nec-labs.com
2University of Kentucky, Lexington, KY 40506, USA
3University of California San Diego, La Jolla, CA 92093, USA

Abstract. Given a single RGB image of a complex outdoor road scene in the perspective view, we address the novel problem of estimating an occlusion-reasoned semantic scene layout in the top-view. This challeng-ing problem not only requires an accurate understanding of both the 3D geometry and the semantics of the visible scene, but also of occluded areas. We propose a convolutional neural network that learns to pre-dict occluded portions of the scene layout by looking around foreground objects like cars or pedestrians. But instead of hallucinating RGB val-

ues, we show that directly predicting the semantics and depths in theoccluded ar
eas enables a better transformation into the top-view. We fur-
ther show that this initial top-view representation can be signi■cantly
enhanced by learning priors and rules about typical road layouts fromsimulated o
r, if available, map data. Crucially, training our model does
not require costly or subjective human annotations for occluded areas or
the top-view, but rather uses readily available annotations for standardsemantic
 segmentation in the perspective view. We extensively evaluate
and analyze our approach on the KITTI and Cityscapes data sets.
Keywords: 3D scene understanding
·Occlusion reasoning
Semantic top-view representations
1
**********

Hierarchical Metric Learning
and Matching for 2D and 3D Geometric
Correspondences
Mohammed E. Fathy1, Quoc-Huy Tran2(B), M. Zeeshan Zia3, Paul Vernaza2,
and Manmohan Chandraker2,4
1Google Cloud AI, New York, USA
2NEC Laboratories America, Inc., Princeton, USA
qhtran@nec.labs.com
3Microsoft Hololens, London, UK
4University of California, San Diego, US
Abstract. Interest point descriptors have fueled progress on almost
every problem in computer vision. Recent advances in deep neural net-works have
enabled task-speci■c learned descriptors that outperform
hand-crafted descriptors on many problems. We demonstrate that com-
monly used metric learning approaches do not optimally leverage thefeature hiera
rchies learned in a Convolutional Neural Network (CNN),
especially when applied to the task of geometric feature matching. While
a metric loss applied to the deepest layer of a CNN, is often expectedto yield i
deal features irrespective of the task, in fact the growing recep-
tive ■eld as well as striding e■ects cause shallower features to be better
at high precision matching tasks. We leverage this insight together with
explicit supervision at multiple levels of the feature hierarchy for better
regularization, to learn more e■ective descriptors in the context of geo-metric
matching tasks. Further, we propose to use activation maps at
di■erent layers of a CNN, as an e■ective and principled replacement for
the multi-resolution image pyramids often used for matching tasks. Wepropose con
crete CNN architectures employing these ideas, and evaluate
them on multiple datasets for 2D and 3D geometric matching as well
as optical ■ow, demonstrating state-of-the-art results and generalizationacross
datasets.
Keywords: Hierarchical metric learning
·Hierarchical matching
Geometric correspondences ·Dense correspondences
Part of this work was done during M. E. Fathy’s internship at NEC Labs America.C
ode and models will be made available at http://www.nec-labs.com/
~mas/HiLM/ .
Electronic supplementary material The online version of this chapter ( https://
doi.org/10.1007/978-3-030-01267-0 49) contains supplementary material, which is
available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11219, pp. 832–850, 2018.https://doi.o
rg/10.1007/978-3-030-01267-0
_4
**********

Deep Component Analysis
via Alternating Direction Neural

Networks
Calvin Murdock(B), Ming-Fang Chang, and Simon Lucey
Carnegie Mellon University, Pittsburgh, USA
{cmurdock,mingfanc,slucey }@cs.cmu.edu
Abstract. Despite a lack of theoretical understanding, deep neural net-
works have achieved unparalleled performance in a wide range of appli-
cations. On the other hand, shallow representation learning with com-ponent anal
ysis is associated with rich intuition and theory, but smaller
capacity often limits its usefulness. To bridge this gap, we introduce Deep
Component Analysis (DeepCA), an expressive multilayer model formu-lation that en
forces hierarchical structure through constraints on latent
variables in each layer. For inference, we propose a di■erentiable opti-
mization algorithm implemented using recurrent Alternating DirectionNeural Netwo
rks (ADNNs) that enable parameter learning using stan-
dard backpropagation. By interpreting feed-forward networks as single-
iteration approximations of inference in our model, we provide botha novel persp
ective for understanding them and a practical technique
for constraining predictions with prior knowledge. Experimentally, we
demonstrate performance improvements on a variety of tasks, includingsingle-imag
e depth prediction with sparse output constraints.
Keywords: Component analysis
·Deep learning ·Constraints
1
**********
ADVISE: Symbolism and External
Knowledge for Decoding Advertisements
Keren Ye(B)and Adriana Kovashka
University of Pittsburgh, Pittsburgh, PA 15260, USA
{yekeren,kovashka }@cs.pitt.edu
Abstract. In order to convey the most content in their limited space,
advertisements embed references to outside knowledge via symbolism.For example,
a motorcycle stands for adventure (a positive property the
ad wants associated with the product being sold), and a gun stands for
danger (a negative property to dissuade viewers from undesirable behav-iors). We
 show how to use symbolic references to better understand the
meaning of an ad. We further show how anchoring ad understanding
in general-purpose object recognition and image captioning improvesresults. We f
ormulate the ad understanding task as matching the ad
image to human-generated statements that describe the action that the
ad prompts, and the rationale it provides for taking this action. Our pro-posed
method outperforms the state of the art on this task, and on an
alternative formulation of question-answering on ads. We show additional
applications of our learned representations for matching ads to slogans,and clus
tering ads according to their topic, without extra training.
Keywords: Advertisements
·Symbolism ·Question answering
External knowledge ·Vision and language ·Representation learning
1
**********