

# Deep Learning for Computer Vision

Edited by

Daniel Cremers<sup>1</sup>, Laura Leal-Taixé<sup>2</sup>, and René Vidal<sup>3</sup>

<sup>1</sup> TU München, DE, [cremers@tum.de](mailto:cremers@tum.de)

<sup>2</sup> TU München, DE, [leal.taixe@tum.de](mailto:leal.taixe@tum.de)

<sup>3</sup> Johns Hopkins University – Baltimore, US, [rvidal@cis.jhu.edu](mailto:rvidal@cis.jhu.edu)

---

## Abstract

The field of computer vision engages in the goal to enable and enhance a machine's ability to infer knowledge and information from spatial and visual input data. Recent advances in data-driven learning approaches, accelerated by increasing parallel computing power and a ubiquitous availability of large amounts of data, pushed the boundaries of almost every vision related subdomain. The most prominent example of these machine learning approaches is a so called deep neural network (DNN), which works as a general function approximator and can be trained to learn a mapping between given input and target output data. Research on and with these DNN is generally referred to as Deep Learning. Despite its high dimensional and complex input space, research in the field of computer vision was and still is one of the main driving forces for new development in machine and deep learning, and vice versa.

This seminar aims to discuss recent works on theoretical and practical advances in the field of deep learning itself as well as state-of-the-art results achieved by applying learning based approaches to various vision problems. Our diverse spectrum of topics includes theoretical and mathematical insights focusing on a better understanding of the fundamental concepts behind deep learning and a multitude of specific research topics facilitating an exchange of knowledge between peers of the research community.

**Seminar** September 24–29, 2017 – <http://www.dagstuhl.de/17391>

**1998 ACM Subject Classification** I.2.10 Vision and Scene Understanding, I.2.6 Learning

**Keywords and phrases** computer vision, convolutional networks, deep learning, machine learning

**Digital Object Identifier** 10.4230/DagRep.7.9.109

**Edited in cooperation with** Tim Meinhardt

## 1 Executive Summary

*René Vidal*

*Daniel Cremers*

*Laura Leal-Taixé*

**License**  Creative Commons BY 3.0 Unported license  
© René Vidal, Daniel Cremers, and Laura Leal-Taixé

The paradigm that a machine can learn from examples much like humans learn from experience has fascinated researchers since the advent of computers. It has triggered numerous research developments and gave rise to the concept of artificial neural networks as a computational paradigm designed to mimic aspects of signal and information processing in the human brain.

There have been several key advances in this area including the concept of back-propagation learning (essentially gradient descent and chain rule differentiation on the network weight vectors) by Werbos in 1974, later popularized in the celebrated 1984 paper of Rumelhart,



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Deep Learning for Computer Vision, *Dagstuhl Reports*, Vol. 7, Issue 09, pp. 109–125

Editors: Daniel Cremers, Laura Leal-Taixé, and René Vidal



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Hinton and Williams. Despite a certain success in pattern recognition challenges like handwritten digit classification, artificial neural networks dropped in popularity in the 1990s with alternative techniques such as support vector machines gaining attention.

With increasing computational power (and in particular highly parallel GPU architectures) and more sophisticated training strategies such as layer-by-layer pretraining, supervised backpropagation and dropout learning, neural networks regained popularity in the 2000s and the 2010s. With deeper network architectures and more training data, their performance has drastically improved. Over the last couple of years, they have outperformed numerous existing algorithms on a variety of computer vision challenges such as object recognition, semantic segmentation and even stereo and optical flow estimation.

The aim of this Dagstuhl Seminar is to bring together leading experts from the area of machine learning and computer vision and discuss the state-of-the-art in deep learning for computer vision. During our seminar, we will address a variety of both experimental and theoretical questions such as:

1. In which types of challenges do deep learning techniques work well?
2. In which types of challenges do they fail? Are there variations of the network architectures that may enable us to tackle these challenges as well?
3. Which type of network architectures exist (convolutional networks, recurrent networks, deep belief networks, long short-term memory networks, deep Turing machines)? What advantages and drawbacks does each network architecture bring about?
4. Which aspects are crucial for the practical performance of deep network approaches?
5. Which theoretical guarantees can be derived for neural network learning?
6. What properties assure the impressive practical performance despite respective cost functions being generally non-convex?

## 2 Table of Contents

### Executive Summary

*René Vidal, Daniel Cremers, and Laura Leal-Taixé* . . . . . 109

### Overview of Talks

Statistics, Computation and Learning with Graph Neural Networks <i>Joan Bruna Estrach</i> . . . . .	113
A picture of the energy landscape of deep neural networks <i>Pratik Chaudhari</i> . . . . .	113
Recent Advances in Deep Learning at the Computer Vision Group TUM <i>Daniel Cremers</i> . . . . .	113
Deep Learning for Sensorimotor Control <i>Alexey Dosovitskiy</i> . . . . .	114
Do semantic parts emerge in Convolutional Neural Networks? <i>Vittorio Ferrari</i> . . . . .	114
Proximal Backpropagation <i>Thomas Frerix</i> . . . . .	114
Recurrent Neural Networks and Open Sets <i>Jürgen Gall</i> . . . . .	114
Towards deep multi view stereo <i>Silvano Galliani</i> . . . . .	115
Optimization with Deep Learning <i>Raja Giryes</i> . . . . .	115
Semantic Jitter and Look-Around Policies <i>Kristen Grauman</i> . . . . .	115
Global Optimization of Positively Homogeneous Deep Networks <i>Benjamin Haeffele</i> . . . . .	115
Deep Depth From Focus <i>Caner Hazirbas</i> . . . . .	116
Learning by Association and Associative Domain Adaptation <i>Philip Häusser</i> . . . . .	116
Towards universal networks: from Ubertnet to Densereg <i>Iasonas Kokkinos</i> . . . . .	116
Hybrid RNN-HMM models for action recognition <i>Hildegard Kühne</i> . . . . .	117
Continual and lifelong learning <i>Christoph H. Lampert</i> . . . . .	117
Out with the old? CNNs for image-based localization <i>Laura Leal-Taixé</i> . . . . .	117
Structured Multiscale Deep Networks <i>Stephane Mallat</i> . . . . .	117

Learning Proximal Operators	
<i>Michael Möller</i> . . . . .	118
Geometric Deep Learning	
<i>Emanuele Rodolà</i> . . . . .	118
Self-Supervised Deep Learning from Video	
<i>Rahul Sukthankar</i> . . . . .	118
Challenges for Deep Learning in Robotic Vision	
<i>Niko Sünderhauf</i> . . . . .	119
VQA, and why it's asking the wrong questions	
<i>Anton van den Hengel</i> . . . . .	119
Learning CNN filter resolution using multi-scale structured receptive fields	
<i>Jan Van Gemert</i> . . . . .	119
Self-learning visual objects	
<i>Andrea Vedaldi</i> . . . . .	120
Mathematics of Deep Learning	
<i>René Vidal</i> . . . . .	120
A Primal Dual Network for Low-Level Vision Problems	
<i>Christoph Vogel</i> . . . . .	120
<b>Working groups</b>	
Open Sets and Robotics	
<i>Jürgen Gall</i> . . . . .	121
Object Recognition	
<i>Kristen Grauman, Pratik Chaudhari, Vittorio Ferrari, Raja Giryes, Iasonas Kokkinos, and Andrea Vedaldi</i> . . . . .	122
Deep Learning for 3D and Graphs	
<i>Emanuele Rodolà</i> . . . . .	122
Deep Learning for Time Series	
<i>Jan Van Gemert</i> . . . . .	123
<b>Participants</b> . . . . .	125

### 3 Overview of Talks

#### 3.1 Statistics, Computation and Learning with Graph Neural Networks

*Joan Bruna Estrach (New York University, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Joan Bruna Estrach

Many problems across applied sciences, from particle physics to recommender systems, are formulated in terms of signals defined over non-Euclidean geometries, and also come with strong geometric stability priors. In this talk, I presented techniques that exploit geometric stability in general geometries with appropriate graph neural network architectures. I showed that these techniques can all be framed in terms of local graph generators such as the graph Laplacian. I presented some stability certificates, as well as applications to computer graphics, particle physics and graph estimation problems. In particular, I described how graph neural networks can be used to reach statistical detection thresholds in community detection, and attack hard combinatorial optimization problems, such as the Quadratic Assignment Problem.

#### 3.2 A picture of the energy landscape of deep neural networks

*Pratik Chaudhari (UCLA, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Pratik Chaudhari

I discussed some peculiarities of the residual surface discovered in the statistical physics literature, and a recently developed algorithm named Entropy-SGD that exploits them. Entropy-SGD can be shown to compute the solution of a viscous Hamilton-Jacobi PDE, which leads to a non-greedy, stochastic optimal control counterpart of SGD that is provably faster. This analysis establishes a previously unknown link between tools of statistical physics, non-convex optimization and the theory of PDEs. In the limit as the viscosity goes to zero, the non-viscous Hamilton-Jacobi PDE leads to the well-known proximal point iteration via the Hopf-Lax formula, thereby providing another link to classical techniques in convex optimization. Moreover, Entropy-SGD includes as special cases some of the most popular algorithms in the deep learning literature, e.g., distributed algorithms like Elastic-SGD as well as algorithms in Bayesian machine learning. It enjoys exceptional empirical performance and obtains state-of-the-art generalization errors with optimal convergence rates, all without any extra hyper-parameters to tune.

#### 3.3 Recent Advances in Deep Learning at the Computer Vision Group TUM

*Daniel Cremers (TU München, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Daniel Cremers

A short overview of Deep Learning Activities in our group.

### 3.4 Deep Learning for Sensorimotor Control

*Alexey Dosovitskiy (Intel Deutschland GmbH – Feldkirchen, DE)*

License  Creative Commons BY 3.0 Unported license  
© Alexey Dosovitskiy

Deep learning, including deep reinforcement learning, is recently being actively used for learning sensorimotor control – producing useful actions in environments based on sensory inputs. I talked about deep RL, imitation learning, and issues with end-to-end learning for sensorimotor control.

### 3.5 Do semantic parts emerge in Convolutional Neural Networks?


*Vittorio Ferrari (University of Edinburgh, GB)*

License  Creative Commons BY 3.0 Unported license  
© Vittorio Ferrari

Study of whether CNNs trained for object classification spontaneously learn semantic parts in their internal representation.

### 3.6 Proximal Backpropagation

*Thomas Frerix (TU München, DE)*

License  Creative Commons BY 3.0 Unported license  
© Thomas Frerix

We offer a generalized point of view on the backpropagation algorithm, currently the most common technique to train neural networks via stochastic gradient descent and variants thereof. Specifically, we show that backpropagation of a prediction error is equivalent to sequential gradient descent steps on a quadratic penalty energy. This energy comprises the network activations as variables of the optimization and couples them to the network parameters. Based on this viewpoint, we illustrate the limitations on step sizes when optimizing a nested function with gradient descent. Rather than taking explicit gradient steps, where step size restrictions are an impediment for optimization, we propose proximal backpropagation (ProxProp) as a novel algorithm that takes implicit gradient steps to update the network parameters.

### 3.7 Recurrent Neural Networks and Open Sets

*Jürgen Gall (Universität Bonn, DE)*

License  Creative Commons BY 3.0 Unported license  
© Jürgen Gall

The first part of the talk dealt with recurrent neural networks (RNNs). RNNs are used in Computer Vision in two ways. Either for modeling temporal relations or for converting an iterative algorithm into an end-to-end learning system. For the latter case, I gave an example that converts a BoW model based on KMeans into a neural network. The second

part of the talk addresses the limitations of closed sets benchmarks, which do not reflect the characteristics of real-world problems. I briefly discussed an domain adaptation approach with open sets.

### 3.8 Towards deep multi view stereo

*Silvano Galliani (ETH Zürich, CH)*

**License** © Creative Commons BY 3.0 Unported license  
© Silvano Galliani

Is it possible to learn Multi View Stereo? I presented recent advances using supervised and unsupervised learning to solve Multi View Stereo.

### 3.9 Optimization with Deep Learning

*Raja Giryes (Tel Aviv University, IL)*

**License** © Creative Commons BY 3.0 Unported license  
© Raja Giryes

Brief description on recent results related to how it is possible to solve optimization problems efficiently with deep learning.

### 3.10 Semantic Jitter and Look-Around Policies

*Kristen Grauman (University of Texas – Austin, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Kristen Grauman

I presented our recent work using deep learning to predict missing visual data, for two very different goals. The first considers the problem where the missing data are entire training examples; we propose “semantic jitter” to generate realistic synthetic images for training fine-grained relative attributes. The second considers the case where the missing data are yet-unseen portions of a scene or 3D object; we proposed a reinforcement learning approach to obtain exploratory policies that actively select new observations to reconstruct the full scene or 3D object, thereby learning how to look around intelligently.

### 3.11 Global Optimization of Positively Homogeneous Deep Networks


*Benjamin Haeffele (Johns Hopkins University – Baltimore, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Benjamin Haeffele

I discussed sufficient conditions of the neural network training problem that guarantee when local minima are globally optimal and show that a global minimum can be reached via local descent from any initialization. The key required conditions are that both the network output and the regularization on the network parameters needs to be positively homogeneous of the same degree, with the regularization specifically constructed to control the network size.

### 3.12 Deep Depth From Focus

*Caner Hazirbas (TU München, DE)*

License  Creative Commons BY 3.0 Unported license  
© Caner Hazirbas

Depth from Focus (DFF) is one of the classical ill-posed inverse problems in computer vision. Most approaches recover the depth at each pixel based on the focal setting which exhibits maximal sharpness. Yet, it is not obvious how to reliably estimate the sharpness level, particularly in low-textured areas. In this paper, we propose *Deep Depth From Focus* (DDFF) as the first end-to-end learning approach to this problem. Towards this goal, we create a novel real-scene indoor benchmark composed of 4D light-field images obtained from a plenoptic camera and ground truth depth obtained from a registered RGB-D sensor. Compared to existing benchmarks our dataset is 30 times larger, enabling the use of machine learning for this inverse problem. We compare our results with state-of-the-art DFF methods and we also analyze the effect of several key deep architectural components. These experiments show that DDFFNet achieves state-of-the-art performance in all scenes, reducing depth error by more than 70% w.r.t. classic DFF methods.

### 3.13 Learning by Association and Associative Domain Adaptation


*Philip Häusser (TU München, DE)*

License  Creative Commons BY 3.0 Unported license  
© Philip Häusser

In many real-world scenarios, labeled data for a specific machine learning task is costly to obtain. Semi-supervised training methods make use of abundantly available unlabeled data and a smaller number of labeled examples. We propose a new framework for semi-supervised training of deep neural networks inspired by learning in humans. “Associations” are made from embeddings of labeled samples to those of unlabeled ones and back. The optimization schedule encourages correct association cycles that end up at the same class from which the association was started and penalizes wrong associations ending at a different class. The implementation is easy to use and can be added to any existing end-to-end training setup. We demonstrate the capabilities of learning by association on several data sets and show that it can improve performance on classification tasks tremendously by making use of additionally available unlabeled data. Finally, we show that the proposed association loss produces embeddings that are more effective for domain adaptation compared to methods employing maximum mean discrepancy as a similarity measure in embedding space.

### 3.14 Towards universal networks: from Ubertnet to Densereg

*Iasonas Kokkinos (Facebook AI Research, FR and University College London, GB)*

License  Creative Commons BY 3.0 Unported license  
© Iasonas Kokkinos

I presented an overview of a work on training one network to solve many computer vision tasks – and learning one task (image-to-surface correspondence) to solve many computer vision problems. We also covered works on links between MRFs and CNNs.



### 3.15 Hybrid RNN-HMM models for action recognition

*Hildegard Kühne (Universität Bonn, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Hildegard Kühne

With the increasing demand for large scale data, the need of hand annotated data for the training of such system becomes more and more impractical. One way to avoid frame-based human annotation is the use of action order information to learn the respective action classes. In this context, we propose a hierarchical approach to address the problem by combining a framewise RNN model with a coarse probabilistic inference.

### 3.16 Continual and lifelong learning

*Christoph H. Lampert (IST Austria – Klosterneuburg, AT)*

**License** © Creative Commons BY 3.0 Unported license  
© Christoph H. Lampert

I discussed some recent work from our group towards continual and lifelong learning, and showed examples of open problems.

### 3.17 Out with the old? CNNs for image-based localization

*Laura Leal-Taixé (TU München, DE)*

**License** © Creative Commons BY 3.0 Unported license  
© Laura Leal-Taixé

Deep learning solutions for camera pose regression for indoor and outdoor scenes have become popular in recent years. CNNs allow us to learn suitable feature representations for localization that are robust against motion blur and illumination changes. LSTM units can then be used on the CNN output for structured dimensionality reduction, leading to drastic improvements in localization performance. Nonetheless, quantitative experiments show that SIFT-based methods are still superior to CNN-based methods, except for indoor localization, where textureless regions and repetitive structures are common. I will also present our recent advances in relative pose estimation and start a discussion on how to teach epipolar geometry to Deep Learning architectures.

### 3.18 Structured Multiscale Deep Networks

*Stephane Mallat (Ecole Polytechnique – Palaiseau, FR)*

**License** © Creative Commons BY 3.0 Unported license  
© Stephane Mallat

Deep neural networks compute invariants to translations, deformations and much more complex transformations. I explained how one can structure these networks with progressively more prior information for computer vision. The goal is to better understand their mathematical properties and reduce the number of training samples.

### 3.19 Learning Proximal Operators

*Michael Möller (Universität Siegen, DE)*

License  Creative Commons BY 3.0 Unported license  
© Michael Möller

There are two typical approaches to solving inverse problems in imaging. Variational methods define an energy based on possible solutions in such a way that a low value of the energy corresponds to a solution with desirable properties, and subsequently minimize this energy. Learning based approaches define a parameterized function from the data to the space of solutions and learn the optimal parameters of this mapping from a set of exemplary training images. In this talk I discussed and analyzed the approach to replace the proximal operator of a regularization within a minimization algorithm by a denoising neural network to combine some advantages of both types of methods, variational and learning based techniques.

### 3.20 Geometric Deep Learning

*Emanuele Rodolà (Uni. of Lugano, CH & Sapienza Univ – Rome, IT)*

License  Creative Commons BY 3.0 Unported license  
© Emanuele Rodolà

The past decade in computer vision research has witnessed the re-emergence of “deep learning”, and in particular convolutional neural network (CNN) techniques, allowing to learn powerful image feature representations from large collections of examples. CNNs achieve a breakthrough in performance in a wide range of applications such as image classification, segmentation, detection and annotation. Nevertheless, when attempting to apply the CNN paradigm to 3D shapes (feature-based description, similarity, correspondence, retrieval, etc.) one has to face fundamental differences between images and geometric objects. Shape analysis and geometry processing pose new challenges that are non-existent in image analysis, and deep learning methods have only recently started penetrating into these communities. The purpose of this talk was to overview the foundations and the current state of the art on learning techniques for geometric data analysis. Special focus was put on deep learning techniques (CNN) applied to Euclidean and non-Euclidean manifolds for tasks of shape classification, retrieval and correspondence. The talk presented in a new light the problems of shape analysis and geometry processing, emphasizing the analogies and differences with the classical 2D setting, and showing how to adapt popular learning schemes in order to deal with 3D shapes.

### 3.21 Self-Supervised Deep Learning from Video


*Rahul Sukthankar (Google Research – Mountain View, US)*

License  Creative Commons BY 3.0 Unported license  
© Rahul Sukthankar

I covered a couple of recent efforts from my group on extracting depth, ego-motion, etc. from unlabeled video with the hope of stimulating discussion/interest rather than presenting finished results.

### 3.22 Challenges for Deep Learning in Robotic Vision


*Niko Sünderhauf (Queensland University of Technology – Brisbane, AU)*

License  Creative Commons BY 3.0 Unported license  
© Niko Sünderhauf

I talked about current challenges for deep learning in robotic vision. Where does deep learning fail when deployed on a robot? What unique challenges arise from operation in unconstrained, everyday scenarios? During the workshop, I discussed ideas for a new benchmark challenge to complement COCO and similar existing challenges, with a focus on robotic vision related problems such as open-set operations, incremental learning, active learning, and active vision.

### 3.23 VQA, and why it's asking the wrong questions

*Anton van den Hengel (University of Adelaide, AU)*

License  Creative Commons BY 3.0 Unported license  
© Anton van den Hengel

Visual Question Answering is one of the applications of Deep Learning that is pushing towards real Artificial Intelligence. It turns the training process around by only defining the question after the training has taken place, and in the process, changes the task fundamentally. This talk covered some of the high-level questions about the types of challenges Deep Learning can be applied to, and how we might separate the things its good at from those that it's not.

### 3.24 Learning CNN filter resolution using multi-scale structured receptive fields

*Jan Van Gemert (TU Delft, NL)*

License  Creative Commons BY 3.0 Unported license  
© Jan Van Gemert

The design of subsampling layers in a CNNs is a matter of trial and error. In addition, the filter-size in a single layer is typically hard-coded to be the same. Here I questioned this design. Instead of hard-coding we aim to learn the resolution. We do this by coupling resolution to the standard deviation ( $\sigma$ ) of a Gaussian blur kernel, and then learn CNN filters by learning coefficients of a local differential Gaussian basis. Preliminary results show that global resolution can be learned by optimizing  $\sigma$ , and –in contrast to pixel filters CNNs– is robust to removing the subsampling layers.

### 3.25 Self-learning visual objects

*Andrea Vedaldi (University of Oxford, GB)*

License  Creative Commons BY 3.0 Unported license  
© Andrea Vedaldi

I showed an approach to learn the intrinsic structure of visual object categories without the use of any manual supervision. The method learns a mapping that associates to image pixels a 2D embedding which identifies points of the 3D surface of the underlying deformable object class. The key property of this mapping is to be equivariant, i.e. consistent with image transformations. While these transformations naturally arise from a viewpoint change or an object deformation, we showed that synthetic warps are sufficient to learn such an embedding well. This is a simple yet powerful technique that can learn reliable object landmarks without the need to label them manually in examples. For human faces, I also showed empirically that the learned landmarks are highly-consistent with manually-supervised ones.

### 3.26 Mathematics of Deep Learning

*René Vidal (Johns Hopkins University – Baltimore, US)*

License  Creative Commons BY 3.0 Unported license  
© René Vidal

A mini-tutorial summarizing recent advances in understanding the mathematical properties of deep networks, including stability, generalization, and optimization.

### 3.27 A Primal Dual Network for Low-Level Vision Problems

*Christoph Vogel (TU Graz, AT)*

License  Creative Commons BY 3.0 Unported license  
© Christoph Vogel

In the past, classic energy optimization techniques were the driving force in many innovations and are a building block for almost any problem in computer vision. Efficient algorithms are mandatory to achieve real-time processing, needed in many applications like autonomous driving. However, energy models – even if designed by human experts – might never be able to fully capture the complexity of natural scenes and images. Similar to optimization techniques, Deep Learning has changed the landscape of computer vision in recent years and has helped to push the performance of many models to never experienced heights. Our idea of a primal-dual network is to combine the structure of regular energy optimization techniques, in particular of first order methods, with the flexibility of Deep Learning to adapt to the statistics of the input data.

## 4 Working groups

### 4.1 Open Sets and Robotics

*Jürgen Gall (Universität Bonn, DE)*


License © Creative Commons BY 3.0 Unported license  
© Jürgen Gall

Standard benchmarks in computer vision use a closed set evaluation protocol, i.e. training and test data are sampled from the same data source and it is ensured that all categories in the test data are also present in the training data. For many applications like robotics, however, the closed set assumption of standard benchmarks is unrealistic and just increasing the number of classes and size of the datasets will not advance the field towards realistic open set scenarios, where the test data contains many classes that are not part of the training data. Indeed, for any application that requires to interact within an unconstrained environment, it will be infeasible to cover all cases. In particular rare cases are unlikely to be present in the training data. For an open set scenario, there are several levels of difficulty a system has to tackle. In the simplest setting, the system has to recognize that it does know the class, when it sees an instance from a class that has not been part of the training data. The system, however, has also to recognize rare instances of the classes that do not occur in the training data. On the next level, the system has to be able to learn new classes or adapt to changes in the categories, for instances, when a category needs to be split into several finer categories. This can include active learning, where a system actively requests labels from a human. On the long term, systems or robots have to solve tasks that they have not previously executed or trained on. Although there are several publications on open sets and a few workshops on this topic have been recently organized, open sets have so far received only little attention in the research community. One reason is the lack of benchmarks for open sets, which require a more elaborate evaluation process than closed set benchmarks.

The working group also discussed the ideal representation of the environment for an autonomous system. Categories, as they are commonly used for analyzing images or videos, have been developed for communication and human categories do not necessarily make sense from a perceptual or robotics perspective. It is also clear that categories are insufficient to represent the world, but the ideal representation of a scene remains an open research question. It might be a combination of categories, attributes, affordances but also some fluent representations that cannot be captured by any taxonomy. It is also an open question if a generic perception model should be developed first, which can be used for all kinds of tasks, or if the perception model should be learned from a large set of tasks, which need to be solved by a robot.

## 4.2 Object Recognition

*Kristen Grauman (University of Texas – Austin, US), Pratik Chaudhari (UCLA, US), Vittorio Ferrari (University of Edinburgh, GB), Raja Giryes (Tel Aviv University, IL), Iasonas Kokkinos (Facebook AI Research, FR and University College London, GB), and Andrea Vedaldi (University of Oxford, GB)*

**License**  Creative Commons BY 3.0 Unported license

© Kristen Grauman, Pratik Chaudhari, Vittorio Ferrari, Raja Giryes, Iasonas Kokkinos, and Andrea Vedaldi

We discussed four topics. The first was whether the output of object recognition and/or detection systems can be used to solve other tasks better. For example, if we can recognize objects and their location in an image, we can use this information to infer their albedo, or perhaps their 3D shape. It was noted that monocular depth estimation may not require explicit recognition of objects, but at the same time it appears to be less transferable across domains and settings. If the semantic labels generated by a model are available, they can be used to run local and class-specific models for another task. Alternatively, semantic labels can be used to constrain the output of a neural networks to be more consistent with the “meaning” of the image. For example, you can use that in order to reconstruct an image of a scene compatible with a given semantic layout.


The second topic that was discussed was whether there exists a “holy-grail” supervisory signal, i.e. a learning task that would be sufficient to learn all interesting properties of natural images, replacing sub tasks such as boundary detection, object recognition, normal estimation, depth estimation etc. This appears to be related to finding a “vision complete” problem and is probably not an image-based signal, but rather a RL signal. It was also noted that certain problems such as 3D reconstruction are very generically applicable and do not tap human-induced semantics. Active vision can also be used to more efficiently transfer representations across domain changes.

The third topic that was discussed was how to construct networks that can reason about object instances. There are two approaches: fully-convolutions and region based networks, the latter explicitly iterating on each object instance. There was no consensus on which one will win out, but it was noted that the region-based approaches have clear advantages in terms of representation invariance and simplification of the network architecture. The problem of exploring different instances is connected to the one of neuron binding in the brain, which is also not well understood.

The final topic was the one of self-supervision. We raised a new distinction between natural self-supervision (as predicting motion frames in a video from other frames) and artificial (such as super-resolution, jigsaw puzzles, and so on). It seems that there are self-supervision tasks that can learn an ample array of vision “primitives” from unsupervised data.

## 4.3 Deep Learning for 3D and Graphs

*Emanuele Rodolà (Univ. of Lugano, CH & Sapienza Univ – Rome, IT)*

**License**  Creative Commons BY 3.0 Unported license

© Emanuele Rodolà

The main focus of discussion for the “Deep Learning for 3D and Graphs” group was on the adoption of deep learning techniques to address problems involving geometric data, and vice

versa, on the use of geometric reasoning within deep learning pipelines. To the surprise of the participants, we found that most deep learning approaches dealing with 3D or geometric data place the main focus of attention on synthesis problems – as opposed to traditional “axiomatic” approaches where the bulk of the work is on the analysis and processing of 3D data. While on the one hand this shift of focus is somewhat surprising, on the other hand it can be justified by noting that deep learning is primarily useful precisely in this kind of problems, i.e., whenever training data can be easily generated and the inverse function is hard (if not impossible) to model axiomatically. The second topic of discussion (what can geometry do for deep learning) has received limited attention from the community so far, and the working group faced difficulties when attempting to identify approaches that explore this direction, as well as potential avenues for future research. The discussion converged toward “constrained deep learning”, namely on the increasing necessity to impose hard constraints in deep learning pipelines, which traditionally operate in an unconstrained regime. Constrained problems often arise when dealing with geometry, for example, when the objects of interest are camera motions (which live on the Lie group of rotations), normal fields (which live on the sphere), or Laplace-like operators (whose discretizations live on the manifold of fixed-rank positive semi-definite matrices). While imposing such constraints as hard requirements is commonly achieved “implicitly” by resorting to ad-hoc representations (e.g., quaternions for rotations), the group agreed that there is a global need for ways to satisfy hard constraints when such workarounds are not possible.

## 4.4 Deep Learning for Time Series

*Jan Van Gemert (TU Delft, NL)*

**License**  Creative Commons BY 3.0 Unported license  
© Jan Van Gemert

The working group focused on deep learning for time series, with video and music, as examples of this domain.

The consensus was that motion in video is difficult for deep networks to learn. Current approaches such as FlowNet and FlowNet 2.0 learn motion in the form of optical flow. These methods do well, but do not (yet) enhance over traditional ‘hand-crafted’ optical flow methods. One issue could be the size of the data. Ground truth motion patterns are difficult to obtain, limiting current datasets either to synthetic ground truth, or using hand-crafted optical flow as the labels.

The analogy of object recognition in images is action recognition in video. The field has not converged on a specific solution and results are typically fused with ‘hand-crafted’ methods such as optical flow. We identified three main approaches in action recognition:

1. LSTM on frames: Extract pre-trained deep network features per frame, and learn a recurrent network (LSTM) on these frames. LSTMs have proved quite successful for modeling words in a text. Yet, for video the LSTMs have proved applicable, e.g. for video captioning, but for action recognition it does not yet lead to great breakthroughs.
2. 3D convolutions. The natural extension of convolutions in 2D images to 3D videos is using 3D convolutions. There has been quite some work on this, but it also does not yet lead to significant improvements. The reasons could be that all possible motion and appearance has to be learned from scratch. Which would require orders or magnitude more data than is available.

3. Two-stream fusion of optical flow and RGB. The (arguably) best results are obtained by the 'two-stream' approach, which processes a video in two streams: an 'appearance' stream using only RGB and a 'motion' stream by using optical flow. The results of the two streams are then merged together to obtain a final prediction.

Various combinations of these 3 building blocks have also been observed; e.g.: 3D convolution on optical flow.

For music recognition, it is remarkable that the best results have been obtained by first converting the music to a 2D spectrogram. A spectrogram bins a 1D signal in various frequencies, which is then processed over the whole signal with an overlapping window. This gives a 2D output, and the current best practice is using a 2D CNN to do image recognition on this 'image'.

In conclusion, seems to be a consensus in the working group that in contrast to image classification, object detection, semantic segmentation where deep learning yielded a substantial increase in performance motion is different. The lack of clear improvements points to some fundamental difference between motion and spatial information. Compared to a 3D image (e.g. MRI) a video is quite different. The motion is potentially causal, and perhaps should be treated differently, and the question is how to proceed.



## Participants

- Joan Bruna Estrach  
New York University, US
- Pratik Chaudhari  
UCLA, US
- Daniel Cremers  
TU München, DE
- Alexey Dosovitskiy  
Intel Deutschland GmbH –  
Feldkirchen, DE
- Vittorio Ferrari  
University of Edinburgh, GB
- Thomas Frerix  
TU München, DE
- Jürgen Gall  
Universität Bonn, DE
- Silvano Galliani  
ETH Zürich, CH
- Ravi Garg  
University of Adelaide, AU
- Raja Giryes  
Tel Aviv University, IL
- Kristen Grauman  
University of Texas – Austin, US
- Benjamin Haeffele  
Johns Hopkins University –  
Baltimore, US
- Philip Häusser  
TU München, DE
- Caner Hazirbas  
TU München, DE
- Iasonas Kokkinos  
Facebook AI Research, FR and  
University College London, GB
- Hildegard Kühne  
Universität Bonn, DE
- Christoph H. Lampert  
IST Austria –  
Klosterneuburg, AT
- Laura Leal-Taixé  
TU München, DE
- Stephane Mallat  
Ecole Polytechnique –  
Palaiseau, FR
- Michael Möller  
Universität Siegen, DE
- Emanuele Rodolà  
Uni. of Lugano, CH & Sapienza  
Univ – Rome, IT
- Niko Sünderhauf  
Queensland University of  
Technology – Brisbane, AU
- Rahul Sukthankar  
Google Research –  
Mountain View, US
- Anton van den Hengel  
University of Adelaide, AU
- Jan Van Gemert  
TU Delft, NL
- Andrea Vedaldi  
University of Oxford, GB
- René Vidal  
Johns Hopkins University –  
Baltimore, US
- Christoph Vogel  
TU Graz, AT

