RESEARCH CENTRE

Grenoble - Rhône-Alpes

2020 ACTIVITY REPORT

Project-Team THOTH

Learning visual models from large-scale data

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

DOMAIN

Perception, Cognition and Interaction

THEME

Vision, perception and multimedia interpretation

Contents

Project-Team THOTH

Creation of the Team: 2016 January 01, updated into Project-Team: 2016 March 01

Keywords

Computer sciences and digital sciences

A3.4. – Machine learning and statistics

A5.3. – Image processing and analysis

A5.4. – Computer vision

A5.9. – Signal processing

A6.2.6. – Optimization

A8.2. – Optimization

A9.2. – Machine learning

A9.3. – Signal analysis

A9.7. – AI algorithmics

Other research topics and application domains

B9.5.6. – Data science

1 Team members, visitors, external collaborators

Research Scientists

- Julien Mairal [Team leader, Inria, Researcher, HDR]
- Karteek Alahari [Inria, Researcher, HDR]
- Pierre Gaillard [Inria, Researcher, from Sep 2020]
- Cordelia Schmid [Inria, Senior Researcher, until Mar 2020, HDR]

Faculty Member

- Jocelyn Chanussot [Institut polytechnique de Grenoble, Professor, HDR]

Post-Doctoral Fellows

- Huu Dien Khue Le [Inria, from Jun 2020]
- Margot Selosse [Univ de Grenoble, from Nov 2020]

PhD Students

- Minttu Alakuijala [Google, CIFRE]
- Florent Bartoccioni [Valeo AI, CIFRE]
- Mathilde Caron [Facebook, CIFRE]
- Dexiong Chen [Univ Grenoble Alpes]
- Maha Elbayad [Univ Grenoble Alpes, until Aug 2020]
- Valentin Gabeur [Google, CIFRE]
- Pierre Louis Guhur [Univ Paris-Saclay]
- Ekaterina Iakovleva [Univ Grenoble Alpes]
- Roman Klokov [Inria]
- Andrei Kulunchakov [Inria]
- Bruno Lecouat [Inria]
- Hubert Leterme [Univ Grenoble Alpes]
- Lina Mezghani [Facebook, CIFRE]
- Gregoire Mialon [Inria]
- Alexander Pashevich [Inria]
- Alexandre Sablayrolles [Facebook, CIFRE]
- Mert Bulent Sariyildiz [Naver Labs, CIFRE, from Sep 2020]
- Robin Strudel [École Normale Supérieure de Paris]
- Vladyslav Sydorov [Inria]
- Houssam Zenati [Criteo, CIFRE, from Jul 2020]
- Alexandre Zouaoui [Inria, from Oct 2020]

Technical Staff

- Gaspard Beugnot [Inria, Engineer, from Sep 2020]
- Theo Bodrito [Inria, Engineer, from Nov 2020]
- Mikita Dvornik [Inria, Engineer, until Oct 2020]
- Ricardo Jose Garcia Pinel [Inria, Engineer]
- Xavier Martin [Inria, Engineer, until Aug 2020]
- Gedeon Muhawenayo [Inria, Engineer, from Dec 2020]
- Houssam Zenati [Criteo, Engineer, until Jun 2020]
- Alexandre Zouaoui [Inria, Engineer, until Sep 2020]

Administrative Assistant

- Nathalie Gillot [Inria]

Visiting Scientists

- Ning Huyan [Xidian University, Xi'an, China, until Nov 2020]
- Huan Ni [School of Remote Sensing Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China]
- Dou Quan [Xidian University, Xi'an, China, until Nov 2020]

2 Overall objectives

In 2021, it is expected that nearly 82% of the Internet traffic will be due to videos, and that it would take an individual over 5 million years to watch the amount of video that will cross global IP networks each month by then. Thus, there is a pressing and in fact increasing demand to annotate and index this visual content for home and professional users alike. The available text and speech-transcript metadata is typically not sufficient by itself for answering most queries, and visual data must come into play. On the other hand, it is not imaginable to learn the models of visual content required to answer these queries by manually and precisely annotating every relevant concept, object, scene, or action category in a representative sample of everyday conditions—if only because it may be difficult, or even impossible to decide a priori what are the relevant categories and the proper granularity level. This suggests reverting back to the original metadata as source of annotation, despite the fact that the information it provides is typically sparse (e.g., the location and overall topic of newscasts in a video archive) and noisy (e.g., a movie script may tell us that two persons kiss in some scene, but not when, and the kiss may occur off screen or not have survived the final cut). On the other hand, this weak form of "embedded annotation" is rich and diverse, and mining the corresponding visual data from the web, TV or film archives guarantees that it is representative of the many different scene settings depicted in situations typical of on-line content. Thus, leveraging this largely untapped source of information, rather than attempting to hand label all possibly relevant visual data, is a key to the future use of on-line imagery.

Today's object recognition and scene understanding technology operates in a very different setting; it mostly relies on fully supervised classification engines, and visual models are essentially (piecewise) rigid templates learned from hand labeled images. The sheer scale of on-line data and the nature of the embedded annotation call for a departure from this fully supervised scenario. The main idea of the Thoth project-team is to develop a new framework for learning the structure and parameters of visual models by actively exploring large digital image and video sources (off-line archives as well as growing on-line content, with millions of images and thousands of hours of video), and exploiting the weak supervisory signal provided by the accompanying metadata. This huge volume of visual training

data will allow us to learn complex non-linear models with a large number of parameters, such as deep convolutional networks and higher-order graphical models. This is an ambitious goal, given the sheer volume and intrinsic variability of the visual data available on-line, and the lack of a universally accepted formalism for modeling it. Yet, the potential payoff is a breakthrough in visual object recognition and scene understanding capabilities. Further, recent advances at a smaller scale suggest that this is realistic. For example, it is already possible to determine the identity of multiple people from news images and their captions, or to learn human action models from video scripts. There has also been recent progress in adapting supervised machine learning technology to large-scale settings, where the training data is very large and potentially infinite, and some of it may not be labeled. Methods that adapt the structure of visual models to the data are also emerging, and the growing computational power and storage capacity of modern computers are enabling factors that should of course not be neglected.

One of the main objectives of Thoth is to transform massive visual data into trustworthy knowledge libraries. For that, it addresses several challenges.

- Designing and learning structured models capable of representing complex visual information.
- Learning visual models from minimal supervision or unstructured meta-data.
- Large-scale learning and optimization.

3 Research program

3.1 Designing and learning structured models

The task of understanding image and video content has been interpreted in several ways over the past few decades, namely image classification, detecting objects in a scene, recognizing objects and their spatial extents in an image, estimating human poses, recovering scene geometry, recognizing activities performed by humans. However, addressing all these problems individually provides us with a partial understanding of the scene at best, leaving much of the visual data unexplained.

One of the main goals of this research axis is to go beyond the initial attempts that consider only a subset of tasks jointly, by developing novel models for a more complete understanding of scenes to address all the component tasks. We propose to incorporate the structure in image and video data explicitly into the models. In other words, our models aim to satisfy the complex sets of constraints that exist in natural images and videos. Examples of such constraints include: (i) relations between objects, like signs for shops indicate the presence of buildings, people on a road are usually walking or standing, (ii) higher-level semantic relations involving the type of scene, geographic location, and the plausible actions as a global constraint, e.g., an image taken at a swimming pool is unlikely to contain cars, (iii) relating objects occluded in some of the video frames to content in other frames, where they are more clearly visible as the camera or the object itself move, with the use of long-term trajectories and video object proposals.

This research axis will focus on three topics. The first is developing deep features for video. This involves designing rich features available in the form of long-range temporal interactions among pixels in a video sequence to learn a representation that is truly spatio-temporal in nature. The focus of the second topic is the challenging problem of modeling human activities in video, starting from human activity descriptors to building intermediate spatio-temporal representations of videos, and then learning the interactions among humans, objects and scenes temporally. The last topic is aimed at learning models that capture the relationships among several objects and regions in a single image scene, and additionally, among scenes in the case of an image collection or a video. The main scientific challenges in this topic stem from learning the structure of the probabilistic graphical model as well as the parameters of the cost functions quantifying the relationships among its entities. In the following we will present work related to all these three topics and then elaborate on our research directions.

- **Deep features for vision.** Deep learning models provide a rich representation of complex objects but in return have a large number of parameters. Thus, to work well on difficult tasks, a large amount of data is required. In this context, video presents several advantages: objects are observed from a large range of viewpoints, motion information allows the extraction of moving objects and

parts, and objects can be differentiated by their motion patterns. We initially plan to develop deep features for videos that incorporate temporal information at multiple scales. We then plan to further exploit the rich content in video by incorporating additional cues, such as the detection of people and their body-joint locations in video, minimal prior knowledge of the object of interest, with the goal of learning a representation that is more appropriate for video understanding. In other words, a representation that is learned from video data and targeted at specific applications. For the application of recognizing human activities, this involves learning deep features for humans and their body parts with all their spatiotemporal variations, either directly from raw video data or "pre-processed" videos containing human detections. For the application of object tracking, this task amounts to learning object-specific deep representations, further exploiting the limited annotation provided to identify the object.

- **Modeling human activities in videos.** Humans and their activities are not only one of the most frequent and interesting subjects in videos but also one of the hardest to analyze owing to the complexity of the human form, clothing and movements. As part of this task, the Thoth project-team plans to build on state-of-the-art approaches for spatio-temporal representation of videos. This will involve using the dominant motion in the scene as well as the local motion of individual parts undergoing a rigid motion. Such motion information also helps in reasoning occlusion relationships between people and objects, and the state of the object. This novel spatio-temporal representation ultimately provides the equivalent of object proposals for videos, and is an important component for learning algorithms using minimal supervision. To take this representation even further, we aim to integrate the proposals and the occlusion relationships with methods for estimating human pose in videos, thus leveraging the interplay among body-joint locations, objects in the scene, and the activity being performed. For example, the locations of shoulder, elbow and wrist of a person drinking coffee are constrained to move in a certain way, which is completely different from the movement observed when a person is typing. In essence, this step will model human activities by dynamics in terms of both low-level movements of body-joint locations and global high-level motion in the scene.

- **Structured models.** The interactions among various elements in a scene, such as the objects and regions in it, the motion of object parts or entire objects themselves, form a key element for understanding image or video content. These rich cues define the structure of visual data and how it evolves spatio-temporally. We plan to develop a novel graphical model to exploit this structure. The main components in this graphical model are spatio-temporal regions (in the case of video or simply image regions), which can represent object parts or entire objects themselves, and the interactions among several entities. The dependencies among the scene entities are defined with a higher order or a global cost function. A higher order constraint is a generalization of the pairwise interaction term, and is a cost function involving more than two components in the scene, e.g., several regions, whereas a global constraint imposes a cost term over the entire image or vide such as a prior knowledge on the number of people expected in the scene. The constraints we plan to include generalize several existing methods, which are limited to pairwise interactions or a small restrictive set of higher-order costs. In addition to learning the parameters of these novel functions, we will focus on learning the structure of the graph itself—a challenging problem that is seldom addressed in current approaches. This provides an elegant way to go beyond state-of-the-art deep learning methods, which are limited to learning the high-level interaction among parts of an object, by learning the relationships among objects.

3.2 Learning of visual models from minimal supervision

Today's approaches to visual recognition learn models for a limited and fixed set of visual categories with fully supervised classification techniques. This paradigm has been adopted in the early 2000's, and within it enormous progress has been made over the last decade.

The scale and diversity in today's large and growing image and video collections (such as, e.g., broadcast archives, and personal image/video collections) call for a departure from the current paradigm. This is the case because to answer queries about such data, it is unfeasible to learn the models of visual content by manually and precisely annotating every relevant concept, object, scene, or action category in

a representative sample of everyday conditions. For one, it will be difficult, or even impossible to decide a-priori what are the relevant categories and the proper granularity level. Moreover, the cost of such annotations would be prohibitive in most application scenarios. One of the main goals of the Thoth project-team is to develop a new framework for learning visual recognition models by actively exploring large digital image and video sources (off-line archives as well as growing on-line content), and exploiting the weak supervisory signal provided by the accompanying metadata (such as captions, keywords, tags, subtitles, or scripts) and audio signal (from which we can for example extract speech transcripts, or exploit speaker recognition models).

Textual metadata has traditionally been used to index and search for visual content. The information in metadata is, however, typically sparse (e.g., the location and overall topic of newscasts in a video archive [1]) and noisy (e.g., a movie script may tell us that two persons kiss in some scene, but not when, and the kiss may occur off the screen or not have survived the final cut). For this reason, metadata search should be complemented by visual content based search, where visual recognition models are used to localize content of interest that is not mentioned in the metadata, to increase the usability and value of image/video archives. *The key insight that we build on in this research axis is that while the metadata for a single image or video is too sparse and noisy to rely on for search, the metadata associated with large video and image databases collectively provide an extremely versatile source of information to learn visual recognition models.* This form of "embedded annotation" is rich, diverse and abundantly available. Mining these correspondences from the web, TV and film archives, and online consumer generated content sites such as Flickr, Facebook, or YouTube, guarantees that the learned models are representative for many different situations, unlike models learned from manually collected fully supervised training data sets which are often biased.

The approach we propose to address the limitations of the fully supervised learning paradigm aligns with "Big Data" approaches developed in other areas: we rely on the orders-of-magnitude-larger training sets that have recently become available with metadata to compensate for less explicit forms of supervision. This will form a sustainable approach to learn visual recognition models for a much larger set of categories with little or no manual intervention. Reducing and ultimately removing the dependency on manual annotations will dramatically reduce the cost of learning visual recognition models. This in turn will allow such models to be used in many more applications, and enable new applications based on visual recognition beyond a fixed set of categories, such as natural language based querying for visual content. This is an ambitious goal, given the sheer volume and intrinsic variability of the every day visual content available on-line, and the lack of a universally accepted formalism for modeling it. Yet, the potential payoff is a breakthrough in visual object recognition and scene understanding capabilities.

This research axis is organized into the following three sub-tasks:

- **Weakly supervised learning.** For object localization we will go beyond current methods that learn one category model at a time and develop methods that learn models for different categories concurrently. This allows "explaining away" effects to be leveraged, i.e., if a certain region in an image has been identified as an instance of one category, it cannot be an instance of another category at the same time. For weakly supervised detection in video we will consider detection proposal methods. While these are effective for still images, recent approaches for the spatio-temporal domain need further improvements to be similarly effective. Furthermore, we will exploit appearance and motion information jointly over a set of videos. In the video domain we will also continue to work on learning recognition models from subtitle and script information. The basis of leveraging the script data which does not have a temporal alignment with the video is to use matches in the narrative in the script and the subtitles (which do have a temporal alignment with the video). We will go beyond simple correspondences between names and verbs relating to self-motion, and match more complex sentences related to interaction with objects and other people. To deal with the limited number of occurrences of such actions in a single movie, we will consider approaches that learn action models across a collection of movies.

- **Online learning of visual models.** As a larger number of visual category models is being learned, online learning methods become important, since new training data and categories will arrive over

[1] For example at the Dutch national broadcast archive Netherlands Institute of Sound and Vision, with whom we collaborated in the EU FP7 project AXES, typically one or two sentences are used in the metadata to describe a one hour long TV program.

time. We will develop online learning methods that can incorporate new examples for existing category models, and learn new category models from few examples by leveraging similarity to related categories using multi-task learning methods. Here we will develop new distance-based classifiers and attribute and label embedding techniques, and explore the use of NLP techniques such as skipgram models to automatically determine between which classes transfer should occur. Moreover, NLP will be useful in the context of learning models for many categories to identify synonyms, and to determine cases of polysemy (e.g. jaguar car brand v.s. jaguar animal), and merge or refine categories accordingly. Ultimately this will result in methods that are able to learn an"encyclopedia" of visual models.

- **Visual search from unstructured textual queries.** We will build on recent approaches that learn recognition models on-the-fly (as the query is issued) from generic image search engines such as Google Images. While it is feasible to learn models in this manner in a matter of seconds, it is challenging to use the model to retrieve relevant content in real-time from large video archives of more than a few thousand hours. To achieve this requires feature compression techniques to store visual representations in memory, and cascaded search techniques to avoid exhaustive search. This approach, however, leaves untouched the core problem of how to associate visual material with the textual query in the first place. The second approach we will explore is based on image annotation models. In particular we will go beyond image-text retrieval methods by using recurrent neural networks such as Elman networks or long short-term memory (LSTM) networks to generate natural language sentences to describe images.

3.3 Large-scale learning and optimization

We have entered an era of massive data acquisition, leading to the revival of an old scientific utopia: it should be possible to better understand the world by automatically converting data into knowledge. It is also leading to a new economic paradigm, where data is a valuable asset and a source of activity. Therefore, developing scalable technology to make sense of massive data has become a strategic issue. Computer vision has already started to adapt to these changes.

In particular, very high-dimensional models such as deep networks are becoming highly popular and successful for visual recognition. This change is closely related to the advent of big data. On the one hand, these models involve a huge number of parameters and are rich enough to represent well complex objects such as natural images or text corpora. On the other hand, they are prone to overfitting (fitting too closely to training data without being able to generalize to new unseen data) despite regularization; to work well on difficult tasks, they require a large amount of labeled data that has been available only recently. Other cues may explain their success: the deep learning community has made significant engineering efforts, making it possible to learn in a day on a GPU large models that would have required weeks of computations on a traditional CPU, and it has accumulated enough empirical experience to find good hyper-parameters for its networks.

To learn the huge number of parameters of deep hierarchical models requires scalable optimization techniques and large amounts of data to prevent overfitting. This immediately raises two major challenges: how to learn without large amounts of labeled data, or with weakly supervised annotations? How to efficiently learn such huge-dimensional models? To answer the above challenges, we will concentrate on the design and theoretical justifications of deep architectures including our recently proposed deep kernel machines, with a focus on weakly supervised and unsupervised learning, and develop continuous and discrete optimization techniques that push the state of the art in terms of speed and scalability.

This research axis will be developed into three sub-tasks:

- **Deep kernel machines for structured data.** Deep kernel machines combine advantages of kernel methods and deep learning. Both approaches rely on high-dimensional models. Kernels implicitly operate in a space of possibly infinite dimension, whereas deep networks explicitly construct high-dimensional nonlinear data representations. Yet, these approaches are complementary: Kernels can be built with deep learning principles such as hierarchies and convolutions, and approximated by multilayer neural networks. Furthermore, kernels work with structured data and have well understood theoretical principles. Thus, a goal of the Thoth project-team is to design and optimize the training of such deep kernel machines.

- **Large-scale parallel optimization.** Deep kernel machines produce nonlinear representations of input data points. After encoding these data points, a learning task is often formulated as a *large-scale convex optimization problem*; for example, this is the case for linear support vector machines, logistic regression classifiers, or more generally many empirical risk minimization formulations. We intend to pursue recent efforts for making convex optimization techniques that are dedicated to machine learning more scalable. Most existing approaches address scalability issues either in model size (meaning that the function to minimize is defined on a domain of very high dimension), or in the amount of training data (typically, the objective is a large sum of elementary functions). There is thus a large room for improvements for techniques that jointly take these two criteria into account.

- **Large-scale graphical models.** To represent structured data, we will also investigate graphical models and their optimization. The challenge here is two-fold: designing an adequate cost function and minimizing it. While several cost functions are possible, their utility will be largely determined by the efficiency and the effectiveness of the optimization algorithms for solving them. It is a combinatorial optimization problem involving billions of variables and is NP-hard in general, requiring us to go beyond the classical approximate inference techniques. The main challenges in minimizing cost functions stem from the large number of variables to be inferred, the inherent structure of the graph induced by the interaction terms (e.g., pairwise terms), and the high-arity terms which constrain multiple entities in a graph.

3.4 Datasets and evaluation

Standard benchmarks with associated evaluation measures are becoming increasingly important in computer vision, as they enable an objective comparison of state-of-the-art approaches. Such datasets need to be relevant for real-world application scenarios; challenging for state-of-the-art algorithms; and large enough to produce statistically significant results.

A decade ago, small datasets were used to evaluate relatively simple tasks, such as for example interest point matching and detection. Since then, the size of the datasets and the complexity of the tasks gradually evolved. An example is the Pascal Visual Object Challenge with 20 classes and approximately 10,000 images, which evaluates object classification and detection. Another example is the ImageNet challenge, including thousands of classes and millions of images. In the context of video classification, the TrecVid Multimedia Event Detection challenges, organized by NIST, evaluate activity classification on a dataset of over 200,000 video clips, representing more than 8,000 hours of video, which amounts to 11 months of continuous video.

Almost all of the existing image and video datasets are annotated by hand; it is the case for all of the above cited examples. In some cases, they present limited and unrealistic viewing conditions. For example, many images of the ImageNet dataset depict upright objects with virtually no background clutter, and they may not capture particularly relevant visual concepts: most people would not know the majority of subcategories of snakes cataloged in ImageNet. This holds true for video datasets as well, where in addition a taxonomy of action and event categories is missing.

Our effort on data collection and evaluation will focus on two directions. First, we will design and assemble video datasets, in particular for action and activity recognition. This includes defining relevant taxonomies of actions and activities. Second, we will provide data and define evaluation protocols for weakly supervised learning methods. This does not mean of course that we will forsake human supervision altogether: some amount of ground-truth labeling is necessary for experimental validation and comparison to the state of the art. Particular attention will be paid to the design of efficient annotation tools.

Not only do we plan to collect datasets, but also to provide them to the community, together with accompanying evaluation protocols and software, to enable a comparison of competing approaches for action recognition and large-scale weakly supervised learning. Furthermore, we plan to set up evaluation servers together with leader-boards, to establish an unbiased state of the art on held out test data for which the ground-truth annotations are not distributed. This is crucial to avoid tuning the parameters for a specific dataset and to guarantee a fair evaluation.

- **Action recognition.** We will develop datasets for recognizing human actions and human-object interactions (including multiple persons) with a significant number of actions. Almost all of today's action recognition datasets evaluate classification of short video clips into a number of predefined categories, in many cases a number of different sports, which are relatively easy to identify by their characteristic motion and context. However, in many real-world applications the goal is to identify and localize actions in entire videos, such as movies or surveillance videos of several hours. The actions targeted here are "real-world" and will be defined by compositions of atomic actions into higher-level activities. One essential component is the definition of relevant taxonomies of actions and activities. We think that such a definition needs to rely on a decomposition of actions into poses, objects and scenes, as determining all possible actions without such a decomposition is not feasible. We plan to provide annotations for spatio-temporal localization of humans as well as relevant objects and scene parts for a large number of actions and videos.

- **Weakly supervised learning.** We will collect weakly labeled images and videos for training. The collection process will be semi-automatic. We will use image or video search engines such as Google Image Search, Flickr or YouTube to find visual data corresponding to the labels. Initial datasets will be obtained by manually correcting whole-image/video labels, i.e., the approach will evaluate how well the object model can be learned if the entire image or video is labeled, but the object model has to be extracted automatically. Subsequent datasets will feature noisy and incorrect labels. Testing will be performed on PASCAL VOC'07 and ImageNet, but also on more realistic datasets similar to those used for training, which we develop and manually annotate for evaluation. Our dataset will include both images and videos, the categories represented will include objects, scenes as well as human activities, and the data will be presented in realistic conditions.

- **Joint learning from visual information and text.** Initially, we will use a selection from the large number of movies and TV series for which scripts are available on-line, see for example `http://www.dailyscript.com` and `http://www.weeklyscript.com`. These scripts can easily be aligned with the videos by establishing correspondences between script words and (timestamped) spoken ones obtained from the subtitles or audio track. The goal is to jointly learn from visual content and text. To measure the quality of such a joint learning, we will manually annotate some of the videos. Annotations will include the space-time locations of the actions as well as correct parsing of the sentence. While DVDs will, initially, receive most attention, we will also investigate the use of data obtained from web pages, for example images with captions, or images and videos surrounded by text. This data is by nature more noisy than scripts.

4 Application domains

4.1 Visual applications

Any solution to automatically understanding images and videos on a semantic level will have an immediate impact on a wide range of applications. For example:

- Semantic-level image and video access is highly relevant for visual search on the Web, in professional archives and personal collections.

- Visual data organization is applicable to organizing family photo and video albums as well as to large-scale information retrieval.

- Visual object recognition has potential applications ranging from surveillance, service robotics for assistance in day-to-day activities as well as the medical domain.

- Real-time scene understanding is relevant for human interaction through devices such as HoloLens, Oculus Rift.

4.2 Pluri-disciplinary research

Machine learning is intrinsically pluri-disciplinary. By developing large-scale machine learning models and algorithms for processing data, the Thoth team became naturally involved in pluri-disciplinary collaborations that go beyond visual modelling. During the last few years, Thoth has conducted several collaborations in other fields such as neuroimaging, bioinformatics, natural language processing, and remote sensing.

5 Highlights of the year

5.1 Awards

- Alberto Bietti received the best PhD award of Université Grenoble Alpes.
- Jocelyn Chanussot received the label 2020 Highly Cited Researcher: top 0.1% most influential researchers, according to Thomson Reuters / Clarivate Analytics
- Valentin Gabeur, Chen Sun (Google), Karteek Alahari and Cordelia Schmid won the first place of the CVPR 2020 Video Pentathlon Challenge with their Multi-modal Transformer for Video Retrieval.

6 New software and platforms

6.1 New software

6.1.1 Cyanure

Name: Cyanure: An Open-Source Toolbox for Empirical Risk Minimization

Keyword: Machine learning

Functional Description: Cyanure is an open-source C++ software package with a Python interface. The goal of Arsenic is to provide state-of-the-art solvers for learning linear models, based on stochastic variance-reduced stochastic optimization with acceleration mechanisms and Quasi-Newton principles. Arsenic can handle a large variety of loss functions (logistic, square, squared hinge, multinomial logistic) and regularization functions (l2, l1, elastic-net, fused Lasso, multi-task group Lasso). It provides a simple Python API, which is very close to that of scikit-learn, which should be extended to other languages such as R or Matlab in a near future.

Release Contributions: version initiale

URL: http://thoth.inrialpes.fr/people/mairal/arsenic/welcome.html

Author: Julien Mairal

Contact: Julien Mairal

Participant: Julien Mairal

6.1.2 graph-ckn

Name: Convolutional Kernel Networks for Graph-Structured Data

Keyword: Machine learning

Functional Description: This is an open-source software package that reproduces the results of the ICML paper "Convolutional Kernel Networks for Graph-Structured Data". It addresses classification and regression problems when data points are graphs.

URL: https://github.com/claying/GCKN

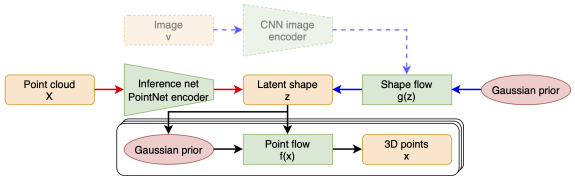


Figure 1: Overview of DPF-Net: arrows indicate data flow to sample new point clouds (blue) and point cloud autoencoding (red), black arrows are used in both processes. During training flow modules are traversed in the reverse direction. For single-view reconstruction the shape prior is conditioned on the image (dashed).

Author: Dexiong Chen

Contact: Julien Mairal

6.1.3 non-local-sparse-coding

Name: Fully Trainable and Interpretable Non-Local Sparse Models for Image Restoration.

Keyword: Image processing

Functional Description: This is an open-source implementation of the ECCV 2020 paper " Fully Trainable and Interpretable Non-Local Sparse Models for Image Restoration.", which reproduces their results.

URL: https://github.com/bruno-31/groupsc

Contact: Julien Mairal

7 New results

7.1 Visual Recognition and Robotics

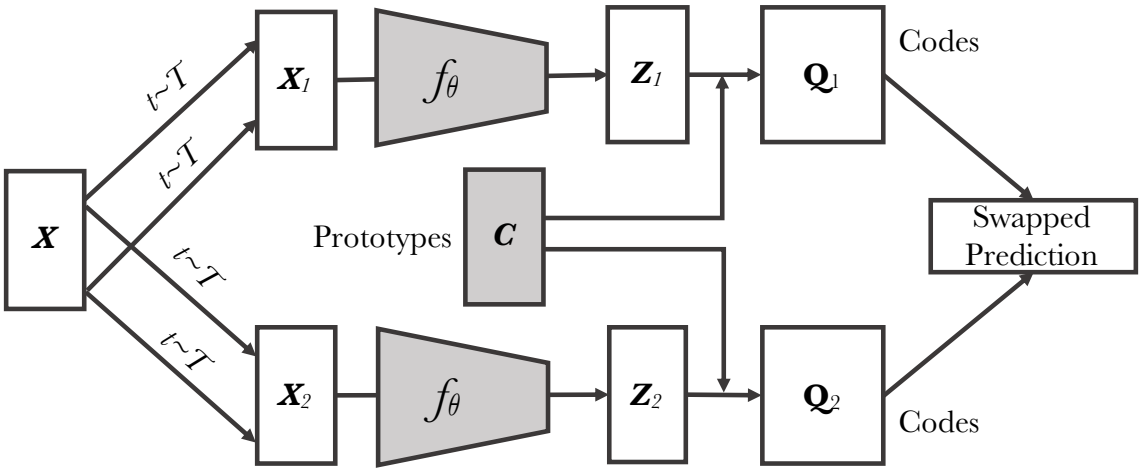
Discrete Point Flow Networks for Efficient Point Cloud Generation

Participants Roman Klokov, Edmond Boyer, Jakob Verbeek.

Generative models have proven effective at modeling 3D shapes and their statistical variations. In this paper [19], we investigate their application to point clouds, a 3D shape representation widely used in computer vision for which, however, only few generative models have yet been proposed. We introduce a latent variable model, depicted in Figure 1, that builds on normalizing flows with affine coupling layers to generate 3D point clouds of an arbitrary size given a latent shape representation. To evaluate its benefits for shape modeling we apply this model for generation, autoencoding, and single-view shape reconstruction tasks. We improve over recent GAN-based models in terms of most metrics that assess generation and autoencoding. Compared to recent work based on continuous flows, our model offers a significant speedup in both training and inference times for similar or better performance. For single-view shape reconstruction we also obtain results on par with state-of-the-art voxel, point cloud, and mesh-based methods.

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

Participants Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin.



Unsupervised image representations have significantly reduced the gap with supervised pretraining, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large number of explicit pairwise feature comparisons, which is computationally challenging. In this paper [7], we propose an online algorithm, SwAV, that takes advantage of contrastive methods without requiring to compute pairwise comparisons. Specifically, our method simultaneously clusters the data while enforcing consistency between cluster assignments produced for different augmentations (or "views") of the same image, instead of comparing features directly as in contrastive learning. Simply put, we use a "swapped" prediction mechanism where we predict the code of a view from the representation of another view. Our method can be trained with large and small batches and can scale to unlimited amounts of data. Compared to previous contrastive methods, our method is more memory efficient since it does not require a large memory bank or a special momentum network. In addition, we also propose a new data augmentation strategy, `multi-crop`, that uses a mix of views with different resolutions in place of two full-resolution views, without increasing the memory or compute requirements. We validate our findings by achieving 75.3% top-1 accuracy on ImageNet with ResNet-50, as well as surpassing supervised pretraining on all the considered transfer tasks.

Synthetic Humans for Action Recognition from Unseen Viewpoints

Participants Cordelia Schmid.

Although synthetic training data has been shown to be beneficial for tasks such as human pose estimation, its use for RGB human action recognition is relatively unexplored. Our goal in this paper [38] is to answer the question whether synthetic humans can improve the performance of human action recognition, with a particular focus on the generalization to unseen viewpoints. We make use of the recent advances in monocular 3D human body reconstruction from real action sequences to automatically render synthetic training videos for the action labels. We make the following contributions: (i) we investigate the extent of variations and augmentations that are beneficial to improving performance at the new viewpoints. We consider changes in body shape and clothing for individuals, as well as more action relevant augmentations such as non-uniform frame sampling, and interpolating between the motion of individuals performing the same action; (ii) We introduce a new data generation methodology, SURREACT, that allows training of spatio-temporal CNNs for action classification; (iii) We substantially improve the state-of-the-art action recognition performance on the NTU RGB+D and UESTC standard human action multi-view benchmarks; Finally, (iv) we extend the augmentation approach to in-the-wild videos from a subset of the Kinetics dataset to investigate the case when only one-shot training data

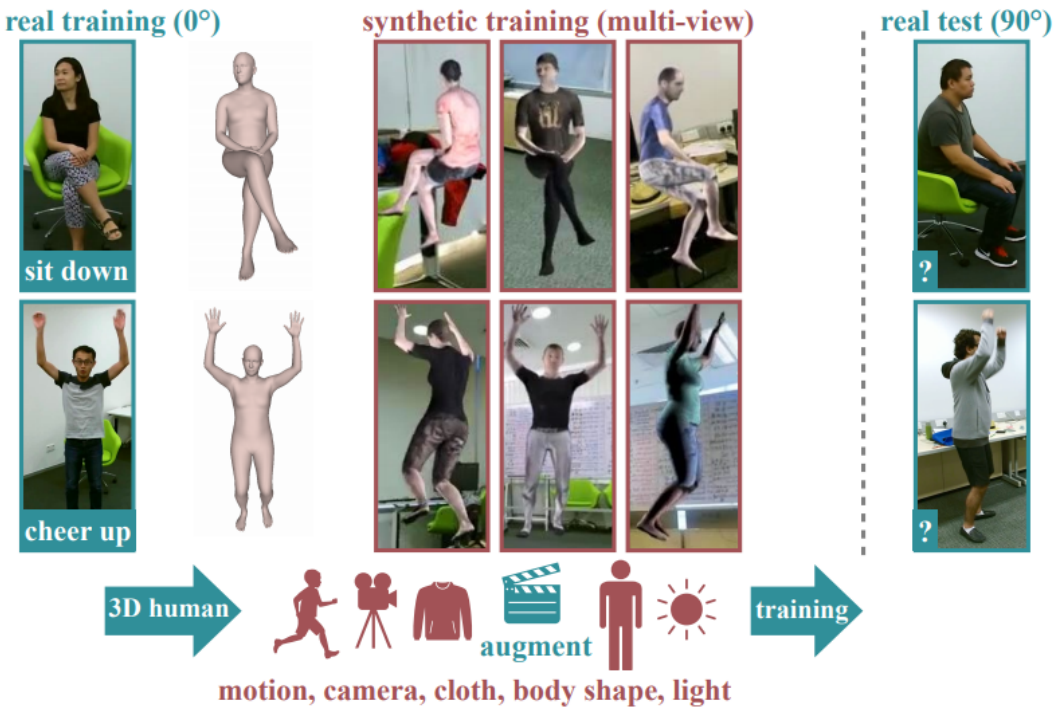


Figure 2: We estimate 3D shape from real videos and automatically render synthetic videos with action labels. We explore various augmentations for motions, viewpoints, and appearance. Training temporal CNNs with this data significantly improves the action recognition from unseen viewpoints.

Distilled Hierarchical Neural Ensembles with Adaptive Inference Cost

Participants Adria Ruiz.

Deep neural networks form the basis of state-of-the-art models across a variety of application domains. Moreover, networks that are able to dynamically adapt the computational cost of inference are important in scenarios where the amount of computation or input data varies over time. In this paper [36], we propose Hierarchical Neural Ensembles (HNE), a novel framework to embed an ensemble of multiple networks by sharing intermediate layers using a hierarchical structure. In HNE we control the inference cost by evaluating only a subset of models, which are organized in a nested manner. Our second contribution is a novel co-distillation method to boost the performance of ensemble predictions with low inference cost. This approach leverages the nested structure of our ensembles, to optimally allocate accuracy and diversity across the ensemble members. Comprehensive experiments over the CIFAR and ImageNet datasets confirm the effectiveness of HNE in building deep networks with adaptive inference cost for image classification.

Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction

Participants Cordelia Schmid.

Modeling hand-object manipulations is essential for understanding how humans interact with their environment. While of practical importance, estimating the pose of hands and objects during interactions is challenging due to the large mutual occlusions that occur during manipulation. Recent efforts have

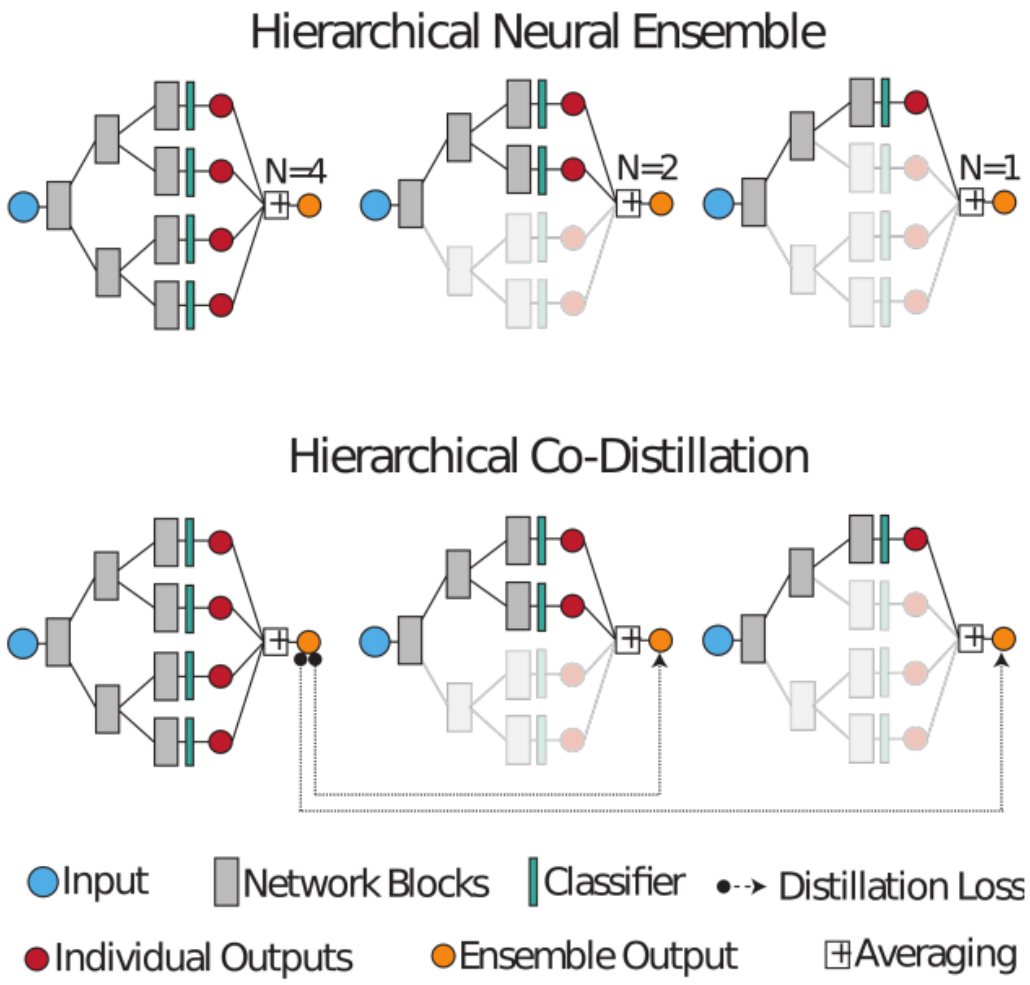


Figure 3: HNE uses a hierarchical parameter-sharing scheme generating a binary tree, where each leaf produces a separate model output. The amount of computation during inference is controlled by determining which part of the tree is evaluated for the ensemble prediction. (Bottom) Our hierarchical distillation approach leverages the full ensemble to supervise parts of the tree that are used in small ensembles.

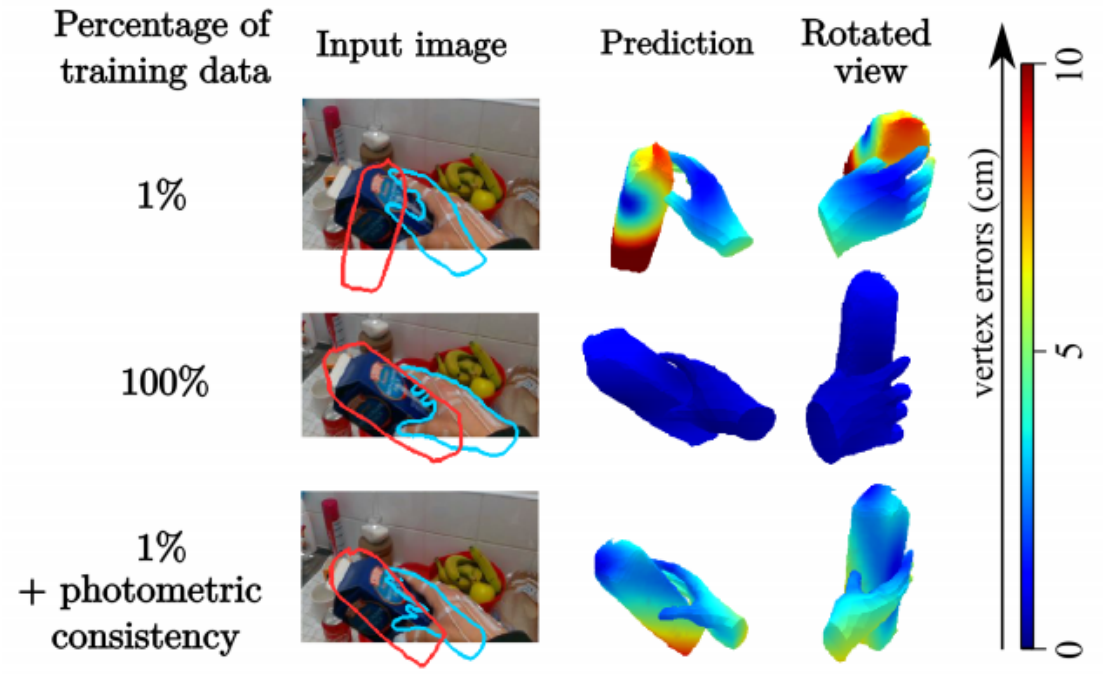


Figure 4: Our method provides accurate 3D hand-object reconstructions from monocular, sparsely annotated RGB videos. We introduce a loss which exploits photometric consistency between neighboring frames. The loss effectively propagates information from a few annotated frames to the rest of the video.

been directed towards fully-supervised methods that require large amounts of labeled training samples. Collecting 3D ground-truth data for hand-object interactions, however, is costly, tedious, and error-prone. To overcome this challenge we present in [17] a method to leverage photometric consistency across time when annotations are only available for a sparse subset of frames in a video. Our model is trained end-to-end on color images to jointly reconstruct hands and objects in 3D by inferring their poses. Given our estimated reconstructions, we differentiably render the optical flow between pairs of adjacent images and use it within the network to warp one frame to another. We then apply a self-supervised photometric loss that relies on the visual consistency between nearby images. We achieve state-of-the-art results on 3D hand-object reconstruction benchmarks and demonstrate that our approach allows us to improve the pose estimation accuracy by leveraging information from neighboring frames in low-data regimes.

Selecting Relevant Features from a Multi-domain Representation for Few-shot Classification

Participants Nikita Dvornik, Cordelia Schmid, Julien Mairal.

Popular approaches for few-shot classification consist of first learning a generic data representation based on a large annotated dataset, before adapting the representation to new classes given only a few labeled samples. In [11], we propose a new strategy based on feature selection, which is both simpler and more effective than previous feature adaptation approaches. First, we obtain a multi-domain representation by training a set of semantically different feature extractors. Then, given a few-shot learning task, we use our multi-domain feature bank to automatically select the most relevant representations. We show that a simple nonparametric classifier built on top of such features produces high accuracy and generalizes to domains never seen during training, leading to state-of-the-art results on MetaDataset and improved accuracy on mini-ImageNet.

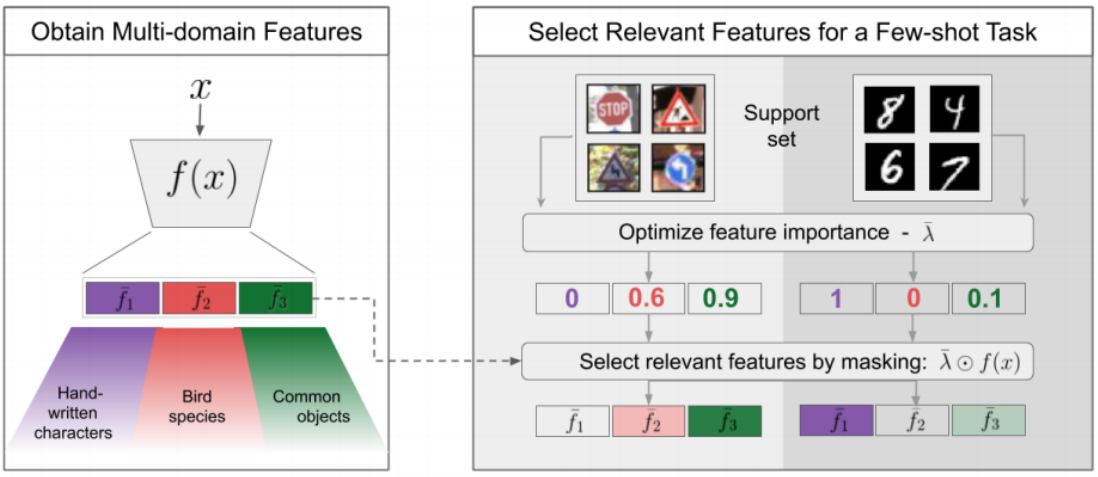


Figure 5: (Left) First, we obtain a multi-domain feature representation, consisting of feature blocks with different semantics. (Right) Given a few-shot task, we select only the relevant feature blocks from the multi-domain representation, by optimizing masking parameters λ on the support set.

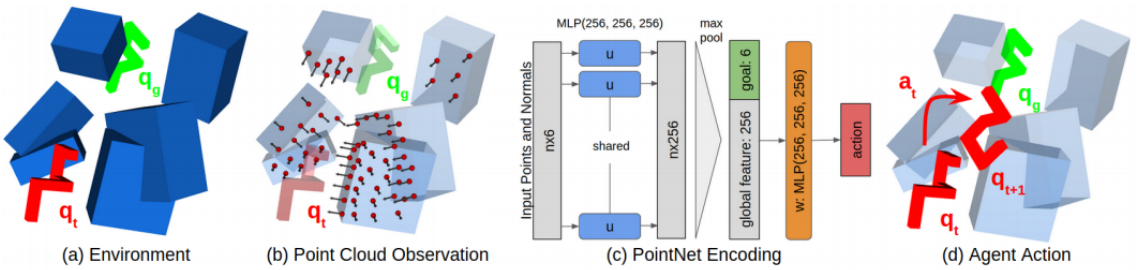


Figure 6: Overview of our approach. (a) We aim to find a collision-free path for a rigid body from its current configuration qt to the goal configuration qg. (b) We assume no prior knowledge about the scene and represent obstacles by points and normals sampled on object surfaces. (c) Our neural network learns the PointNet encoding of observed points and normals together with the motion policy. (d) The learned network generates actions that move the body towards the goal configuration along a collision-free path.

Learning Obstacle Representations for Neural Motion Planning

Participants Cordelia Schmid, Ricardo Garcia.

Motion planning and obstacle avoidance is a key challenge in robotics applications. While previous work succeeds to provide excellent solutions for known environments, sensor-based motion planning in new and dynamic environments remains difficult. In this work [26] we address sensor-based motion planning from a learning perspective. Motivated by recent advances in visual recognition, we argue the importance of learning appropriate representations for motion planning. We propose a new obstacle representation based on the PointNet architecture and train it jointly with policies for obstacle avoidance. We experimentally evaluate our approach for rigid body motion planning in challenging environments and demonstrate significant improvements of the state of the art in terms of accuracy and efficiency.

TAO: A Large-Scale Benchmark for Tracking Any Object

Participants Cordelia Schmid.

For many years, multi-object tracking benchmarks have focused on a handful of categories. Motivated primarily by surveillance and self-driving applications, these datasets provide tracks for people, vehicles, and animals, ignoring the vast majority of objects in the world. By contrast, in the related field of object detection, the introduction of large-scale, diverse datasets (e.g., COCO) have fostered significant progress in developing highly robust solutions. To bridge this gap, we introduce in [10] a similarly diverse dataset for Tracking Any Object (TAO)4 . It consists of 2,907 high resolution videos, captured in diverse environments, which are half a minute long on average. Importantly, we adopt a bottom-up approach for discovering a large vocabulary of 833 categories, an order of magnitude more than prior tracking benchmarks. To this end, we ask annotators to label objects that move at any point in the video, and give names to them post factum. Our vocabulary is both significantly larger and qualitatively different from existing tracking datasets. To ensure scalability of annotation, we employ a federated approach that focuses manual effort on labeling tracks for those relevant objects in a video (e.g., those that move). We perform an extensive evaluation of state-of-the-art trackers and make a number of important discoveries regarding large-vocabulary tracking in an open world. In particular, we show that existing single- and multi-object trackers struggle when applied to this scenario in the wild, and that detection-based, multi-object trackers are in fact competitive with user-initialized ones. We hope that our dataset and analysis will boost further progress in the tracking community.

Learning to combine primitive skills: A step towards versatile robotic manipulation

Participants Cordelia Schmid.

Manipulation tasks such as preparing a meal or assembling furniture remain highly challenging for robotics and vision. Traditional task and motion planning (TAMP) methods can solve complex tasks but require full state observability and are not adapted to dynamic scene changes. Recent learning methods can operate directly on visual inputs but typically require many demonstrations and/or task-specific reward engineering. In this work [27], we aim to overcome previous limitations and propose a reinforcement learning (RL) approach to task planning that learns to combine primitive skills. First, compared to previous learning methods, our approach requires neither intermediate rewards nor complete task demonstrations during training. Second, we demonstrate the versatility of our vision-based task planning in challenging settings with temporary occlusions and dynamic scene changes. Third, we propose an efficient training of basic skills from few synthetic demonstrations by exploring recent CNN architectures and data augmentation. Notably, while all of our policies are learned on visual inputs in simulated environments, we demonstrate the successful transfer and high success rates when applying such policies to manipulation tasks on a real UR5 robotic arm.

Radioactive Data: Tracing Through Training

Participants Alexandre Sablayrolles, Cordelia Schmid.

Data tracing determines whether particular data samples have been used to train a model. In [25], we propose a new technique, radioactive data, that makes imperceptible changes to these samples such that any model trained on them will bear an identifiable mark. Given a trained model, our technique detects the use of radioactive data and provides a level of confidence (p-value). Experiments on large-scale benchmarks (Imagenet), with standard architectures (Resnet-18, VGG-16, Densenet-121) and training procedures, show that we detect radioactive data with high confidence ($p < 0.0001$) when only 1% of the data used to train a model is radioactive. Our radioactive mark is resilient to strong data augmentations and variations of the model architecture. As a result, it offers a much higher signal-to-noise ratio than data poisoning and backdoor methods.

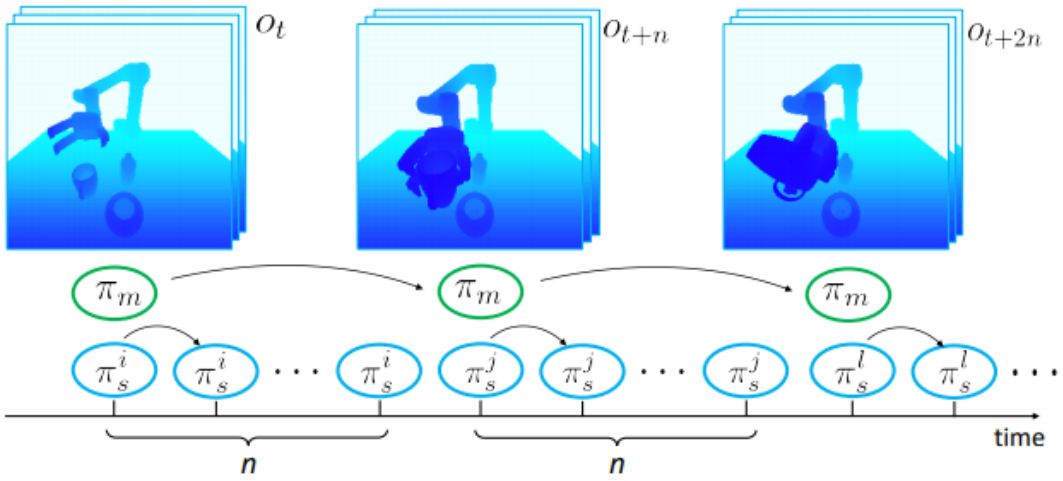


Figure 7: Illustration of our approach. Temporal hierarchy of master and skill policies. Each skill policy generates control for a primitive action such as grasping or pouring.

Multi-modal Transformer for Video Retrieval

Participants Valentin Gabeur, Karteek Alahari, Cordelia Schmid.

The task of retrieving video content relevant to natural language queries plays a critical role in effectively handling internet-scale datasets. Most of the existing methods for this caption-to-video retrieval problem do not fully exploit cross-modal cues present in video. Furthermore, they aggregate per-frame visual features with limited or no temporal information. In this paper [16], we present a multi-modal transformer to jointly encode the different modalities in video, which allows each of them to attend to the others. The transformer architecture is also leveraged to encode and model the temporal information. On the natural language side, we investigate the best practices to jointly optimize the language embedding together with the multi-modal transformer. This novel framework (Fig. 9) allows us to establish state-of-the-art results for video retrieval on three datasets. More details are available at http://thoth.inrialpes.fr/research/MMT

Memory-Augmented Reinforcement Learning for Image-Goal Navigation

Participants Lina Mezghani, Karteek Alahari.

In this work [34], we address the problem of image-goal navigation in the context of visually-realistic 3D environments. This task involves navigating to a location indicated by a target image in a previously unseen environment. Earlier attempts, including RL-based and SLAM-based approaches, have either shown poor generalization performance, or are heavily reliant on pose/depth sensors. We present a novel method, shown in Figure 10, that leverages a cross-episode memory to learn to navigate. We first train a state-embedding network in a self-supervised fashion, and then use it to embed previously-visited states into an agent's memory. In order to avoid overfitting, we propose to use data augmentation on the RGB input during training. We validate our approach through extensive evaluations, showing that our data-augmented memory-based model establishes a new state of the art on the image-goal navigation task in the challenging Gibson dataset. We obtain this competitive performance from RGB input only, without access to additional sensors such as position or depth.

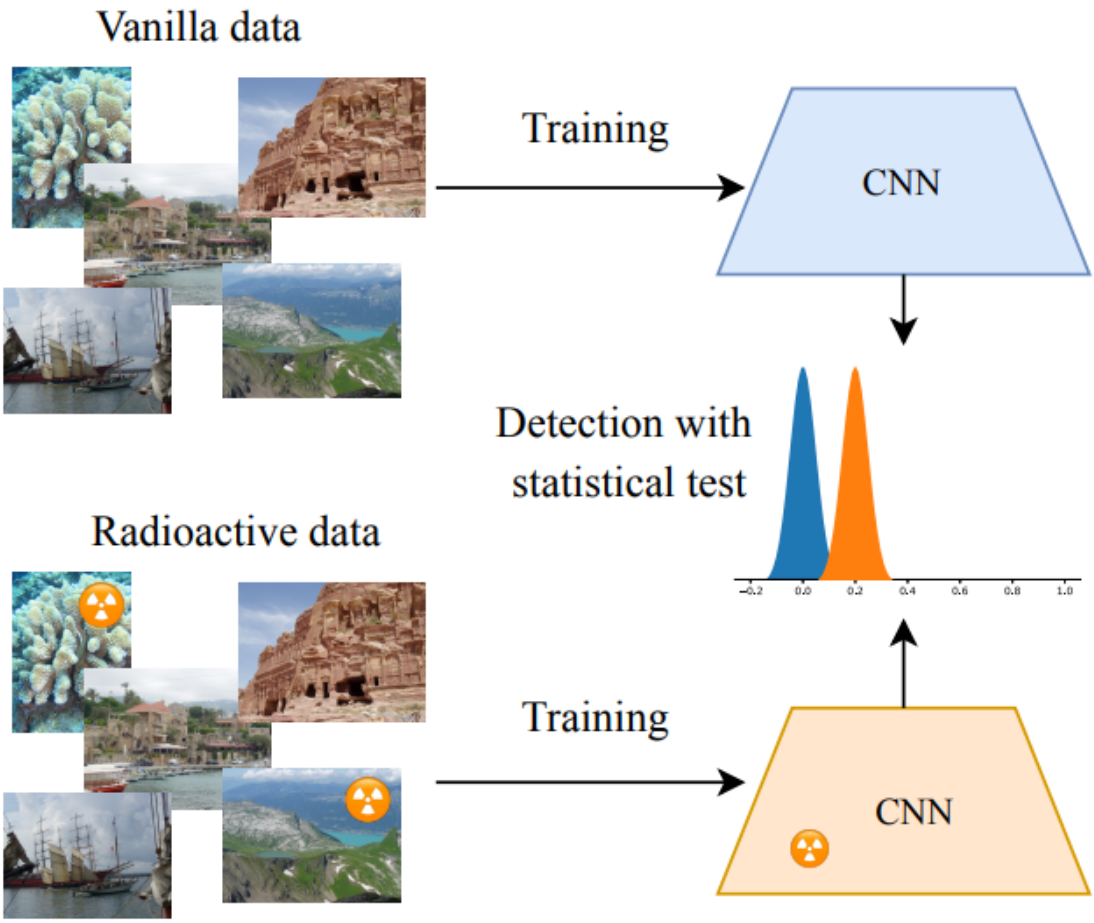


Figure 8: Illustration of our approach: we want to determine through a statistical test (p-value) whether a network has seen a marked dataset or not. The distribution (shown on the histograms) of a statistic on the network weights is clearly separated between the vanilla and radioactive CNNs. Our method works in the cases of both white-box and black-box access to the network.

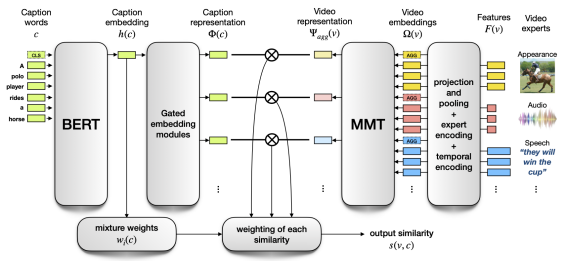


Figure 9: Our cross-modal framework for similarity estimation. We use our Multi-modal Transformer (MMT, right) to encode video, and BERT (left) for text.

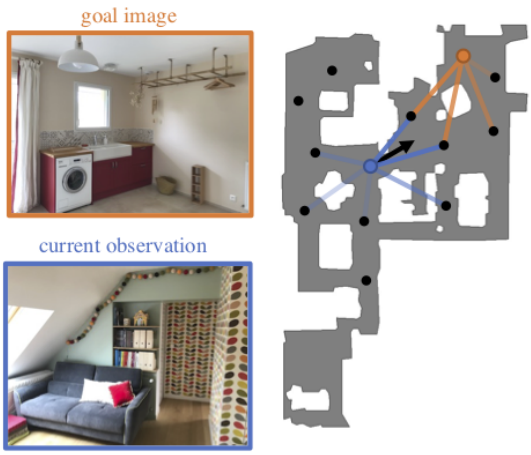


Figure 10: We tackle the problem of image-goal navigation. The agent (shown as the blue dot) is given an image from a goal location (orange dot) which it must navigate to. To address this task, our agent stores a cross-episode memory of previously visited states (black dots), and uses a navigation policy that puts attention (lines) on this memory.

Meta-Learning with Shared Amortized Variational Inference

Participants Ekaterina Iakovleva, Jakob Verbeek, Karteek Alahari.

In this paper [18], we propose a novel amortized variational inference scheme for an empirical Bayes meta-learning model, where model parameters are treated as latent variables. We learn the prior distribution over model parameters conditioned on limited training data using a variational autoencoder approach. Our framework, presented on Figure 11, proposes sharing the same amortized inference network between the conditional prior and variational posterior distributions over the model parameters. While the posterior leverages both the labeled support and query data, the conditional prior is based only on the labeled support data. We show that in earlier work, relying on Monte-Carlo approximation, the conditional prior collapses to a Dirac delta function. In contrast, our variational approach prevents this collapse and preserves uncertainty over the model parameters. We evaluate our approach on the miniImageNet, CIFAR-FS and FC100 datasets, and present results demonstrating its advantages over previous work.

Context Aware Group Activity Recognition

Participants Karteek Alahari.

This paper [9] addresses the task of group activity recognition in multi-person videos. Existing approaches decompose this task into feature learning and relational reasoning. Despite showing progress, these methods only rely on appearance features for people and overlook the available contextual information, which can play an important role in group activity understanding. In this work, we focus on the feature learning aspect and propose a two-stream architecture that not only considers person-level appearance features, but also makes use of contextual information present in videos for group activity recognition. In particular, we propose to use two types of contextual information beneficial for two different scenarios: pose context and scene context that provide crucial cues for group activity understanding. We combine appearance and contextual features to encode each person with an enriched representation. Finally, these combined features are used in relational reasoning for predicting group activities. We

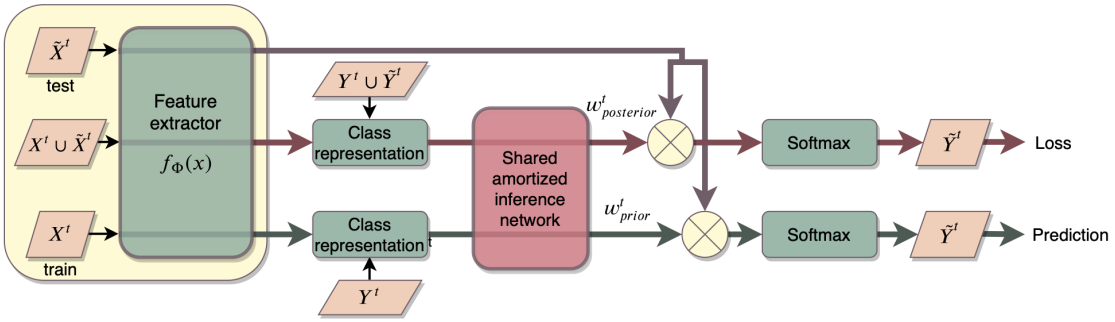


Figure 11: SAMOVAR, our meta-learning model for few-shot image classification. For task t, query data $\tilde{X}^t$ and support data X^t are put through a task-agnostic feature extractor $f_\theta(x)$. The features are then averaged class-wise, and mapped by the shared amortized inference network into prior and posterior over the task-specific classifier weight vectors. Classifiers $w^t_{posterior}$ and w^t_{prior} sampled from these distributions map query features $f_\theta(\tilde{X}^t)$ to predictions on the query labels $\tilde{Y}^t$ used in training and testing, respectively.

evaluate our method on two benchmarks, Volleyball and Collective Activity and show that joint modeling of contextual information with appearance features benefits in group activity understanding.

Concept generalization in visual representation learning

Participants Mert Bulent Sariyildiz, Karteek Alahari.

In this work [37], we study the concept generalization, i.e., the extent to which models trained on a set of (seen) visual concepts can be used to recognize a new set of (unseen) concepts. Although this paradigm is a popular way of evaluating visual representations, the choice of which unseen concepts to use is usually made arbitrarily, and independently from the seen concepts used to train representations, thus ignoring any semantic relationships between the two. We argue that semantic relationships between seen and unseen concepts affect generalization performance and propose ImageNet-CoG (see Figure 12), a novel benchmark on the ImageNet dataset that enables measuring concept generalization in a principled way. Our benchmark leverages expert knowledge that comes from WordNet in order to define a sequence of unseen ImageNet concept sets that are semantically more and more distant from the ImageNet-1K subset, a ubiquitous training set. We analyze a number of such models from supervised, semi-supervised and self-supervised approaches under the prism of concept generalization, and show how our benchmark is able to uncover a number of interesting insights.

7.2 Statistical Machine Learning

Estimate Sequences for Stochastic Composite Optimization: Variance Reduction, Acceleration, and Robustness to Noise

Participants Andrei Kulunchakov, Julien Mairal.

In [3], we propose a unified view of gradient-based algorithms for stochastic convex composite optimization. By extending the concept of estimate sequence introduced by Nesterov, we interpret a large class of stochastic optimization methods as procedures that iteratively minimize a surrogate of the objective. This point of view covers stochastic gradient descent (SGD), the variance-reduction approaches SAGA, SVRG, MISO, their proximal variants, and has several advantages: (i) we provide a simple generic proof of convergence for all of the aforementioned methods; (ii) we naturally obtain new algorithms

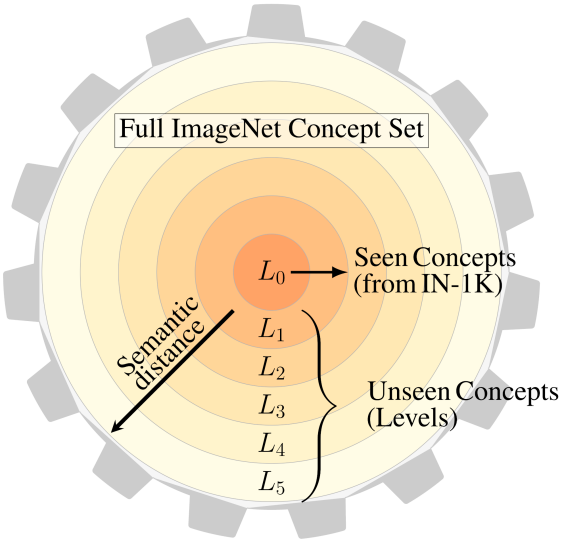


Figure 12: Illustration of concept generalization levels of our proposed ImageNet-CoG benchmark.

with the same guarantees; (iii) we derive generic strategies to make these algorithms robust to stochastic noise, which is useful when data is corrupted by small random perturbations. Finally, we show that this viewpoint is useful to obtain accelerated algorithms. A comparison with different approaches is shown in Figure 13.

Screening Data Points in Empirical Risk Minimization via Ellipsoidal Regions and Safe Loss Functions

Participants Grégoire Mialon, Alexandre d'Aspremont, Julien Mairal.

In this work [24], we design simple screening tests to automatically discard data samples in empirical risk minimization without losing optimization guarantees. We derive loss functions that produce dual objectives with a sparse solution. We also show how to regularize convex losses to ensure such a dual sparsity-inducing property, as can be seen in Figure 14, and propose a general method to design screening tests for classification or regression based on ellipsoidal approximations of the optimal set. In addition to producing computational gains, our approach also allows us to compress a dataset into a subset of representative points.

Counterfactual Learning of Stochastic Policies with Continuous Actions

Participants Houssam Zenati, Alberto Bietti, Matthieu Martin
, Eustache Diemert
, Julien Mairal.

Counterfactual reasoning from logged data has become increasingly important for many applications such as web advertising or healthcare. In this paper, we address the problem of counterfactual learning of stochastic policies with continuous actions, which raises difficult challenges about (i) data modelization, (ii) optimization, and (iii) evaluation on real data.

First, we introduce a modeling strategy based on a joint kernel embedding of contexts and actions, which overcomes the shortcomings of previous discretization strategies as shown in Fig. 15. Second, we empirically show that the optimization aspect of counterfactual learning is more important than previously thought, and we demonstrate the benefits of proximal point algorithms and differentiable

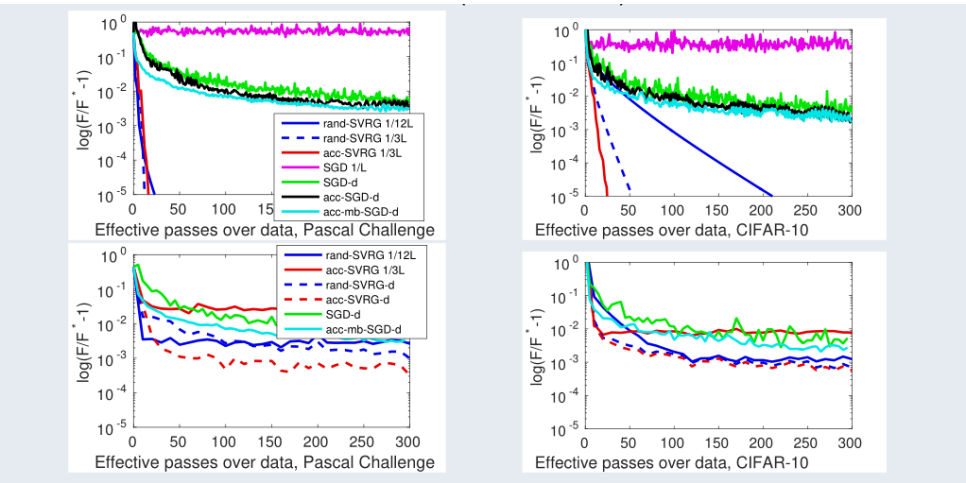


Figure 13: Comparison of different standard approaches with our developed method on two datasets for ℓ_2-logistic regression with mild dropout (bottom) and deterministic case (above). The case of exact gradient computations clearly shows benefits from acceleration, which consist in fast linear convergence. In the stochastic case, we demonstrate either superiority or high competitiveness of the developed method along with its unbiased convergence to the optimum. In both cases, we show that acceleration is able to generically comprise strengths of standard methods and even outperform them.

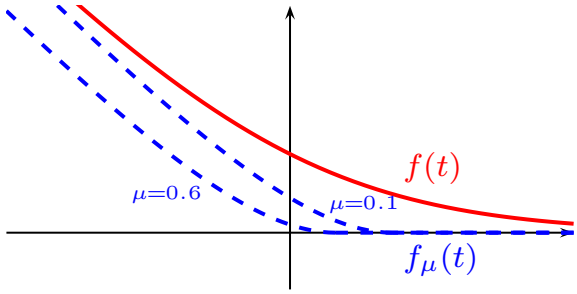


Figure 14: Effect of the dual sparsity inducing regularization on the logistic loss.

estimators. Finally, we propose an evaluation protocol for offline policies in real-world logged systems, which is challenging since policies cannot be replayed on test data, and we release a new large-scale dataset along with multiple synthetic, yet realistic, evaluation setups.

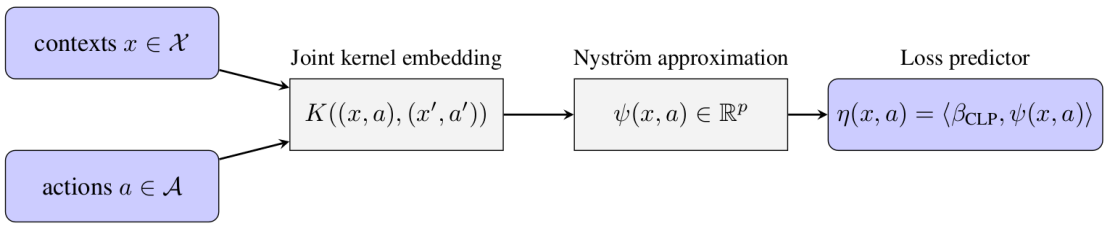


Figure 15: Illustration of the joint kernel embedding for the counterfactual loss predictor (CLP) and loss estimator.

Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model

Participants Raphaël Berthier, Francis Bach, Pierre Gaillard.

In the context of statistical supervised learning, the noiseless linear model assumes that there exists a deterministic linear relation $Y = \langle \theta_*, X, \rangle$ between the random output Y and the random feature vector $\phi(U)$, a potentially non-linear transformation of the inputs U. In [5], we analyze the convergence of single-pass, fixed-step-size stochastic gradient descent on the least-square risk under this model. The convergence of the iterates to the optimum θ_* and the decay of the generalization error follow polynomial convergence rates with exponents that both depend on the regularities of the optimum θ_* and of the feature vectors $\phi(u)$. We interpret our result in the reproducing kernel Hilbert space framework. As a special case, we analyze an online algorithm for estimating a real function on the unit interval from the noiseless observation of its value at randomly sampled points; the convergence depends on the Sobolev smoothness of the function and of a chosen kernel. Figure 16 illustrates the convergence of the algorithm to the optimum (blue dots) and the tightness of our analysis (orange line). Finally, we apply our analysis beyond the supervised learning setting to obtain convergence rates for the averaging process (a.k.a. gossip algorithm) on a graph depending on its spectral dimension.

7.3 Theory and Methods for Deep Neural Networks

A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention

Participants Grégoire Mialon*, Dexiong Chen*, Alexandre d'Aspremont, Julien Mairal.

This work [23] addresses the problem of learning on sets of features, motivated by the need of performing pooling operations in long biological sequences of varying sizes, with long-range dependencies, and possibly few labeled data. To address this challenging task, we introduce a parametrized representation of fixed size, which embeds and then aggregates elements from a given input set according to the optimal transport plan between the set and a trainable reference, see Figure 17. Our approach scales to large datasets and allows end-to-end training of the reference, while also providing a simple unsupervised learning mechanism with small computational cost. Our aggregation technique admits two useful interpretations: it may be seen as a mechanism related to attention layers in neural networks, or it may be seen as a scalable surrogate of a classical optimal transport-based kernel. We experimentally demonstrate the effectiveness of our approach on biological sequences, achieving state-of-the-art results for the protein fold recognition task and detection of chromatin profiles, and, as a proof of concept, we show promising

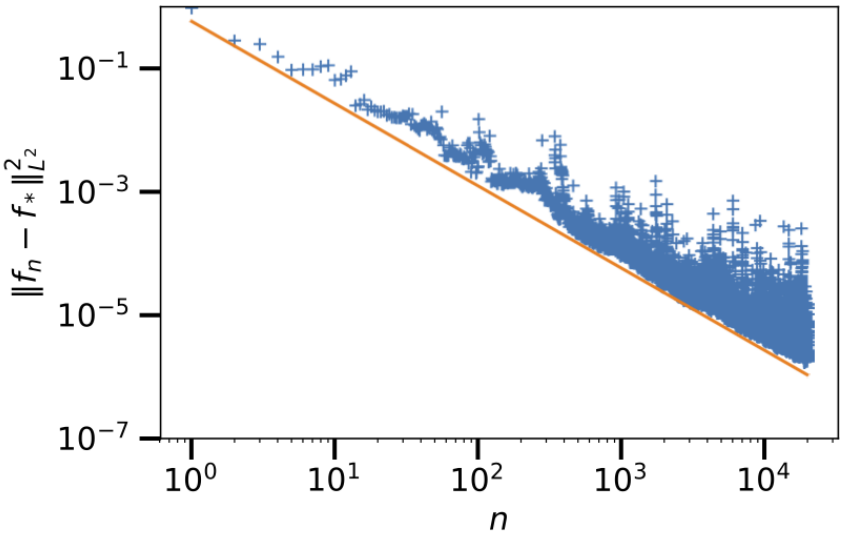


Figure 16: Decay in L2 norm in the interpolation of a function with SGD.

results for processing natural language sequences. We provide an open-source implementation of our embedding that can be used alone or as a module in larger learning models.

Convolutional Kernel Networks for Graph-Structured Data

Participants Dexiong Chen, Laurent Jacob, Julien Mairal.

In this paper [8], we introduce a family of multilayer graph kernels and establish new links between graph convolutional neural networks and kernel methods. Our approach generalizes convolutional kernel networks to graph-structured data, by representing graphs as a sequence of kernel feature maps, where each node carries information about local graph substructures. Figure 18 illustrates the construction of the kernel feature map from layer j to $j+1$. On the one hand, the kernel point of view offers an unsupervised, expressive, and easy-to-regularize data representation, which is useful when limited samples are available. On the other hand, our model can also be trained end-to-end on large-scale data, leading to new types of graph convolutional neural networks. We show that our method achieves competitive performance on several graph classification benchmarks, while offering simple model interpretation.

Fully Trainable and Interpretable Non-Local Sparse Models for Image Restoration

Participants Bruno Lecouat, Jean Ponce, Julien Mairal.

In this work [21] we propose a novel differentiable relaxation of joint sparsity that exploits both principles and leads to a general framework for image restoration which is (1) trainable end to end, (2) fully interpretable, and (3) much more compact than competing deep learning architectures. We apply this approach to denoising, blind denoising, jpeg deblocking, and demosaicking, see Figure 19, and show that, with as few as 100K parameters, its performance on several standard benchmarks is on par or better than state-of-the-art methods that may have an order of magnitude or more parameters.

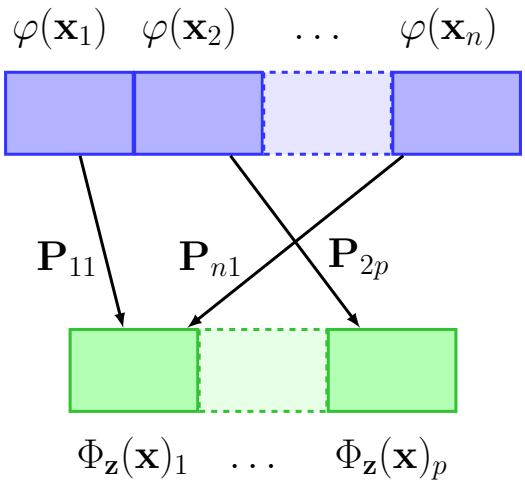


Figure 17: Illustration of our pooling mechanism.

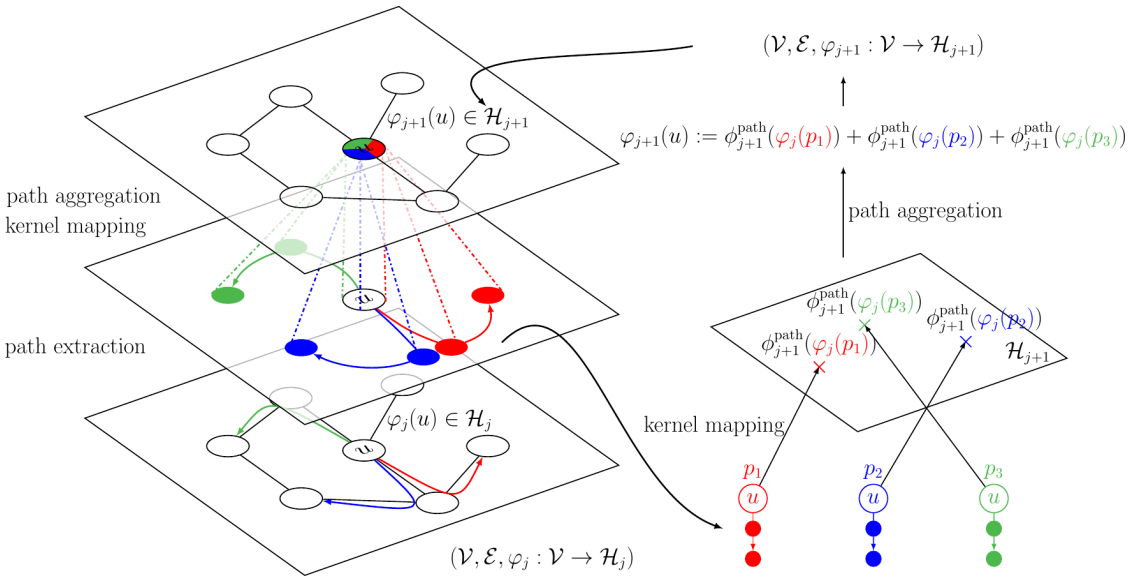


Figure 18: Construction of the graph feature map φ_{j+1} from φ_j given a graph $(\mathcal{V}, \mathcal{E})$. The first step extracts paths of length k (here colored by red, blue and green) from node u, then (on the right panel) maps them to a RKHS $\mathcal{H}_{j+1}$ via the Gaussian kernel mapping. The new map φ_{j+1} at u is obtained by local path aggregation (pooling) of their representations in $\mathcal{H}_{j+1}$. The representations for other nodes can be obtained in the same way. In practice, such a model is implemented by using finite-dimensional embeddings approximating the feature maps.

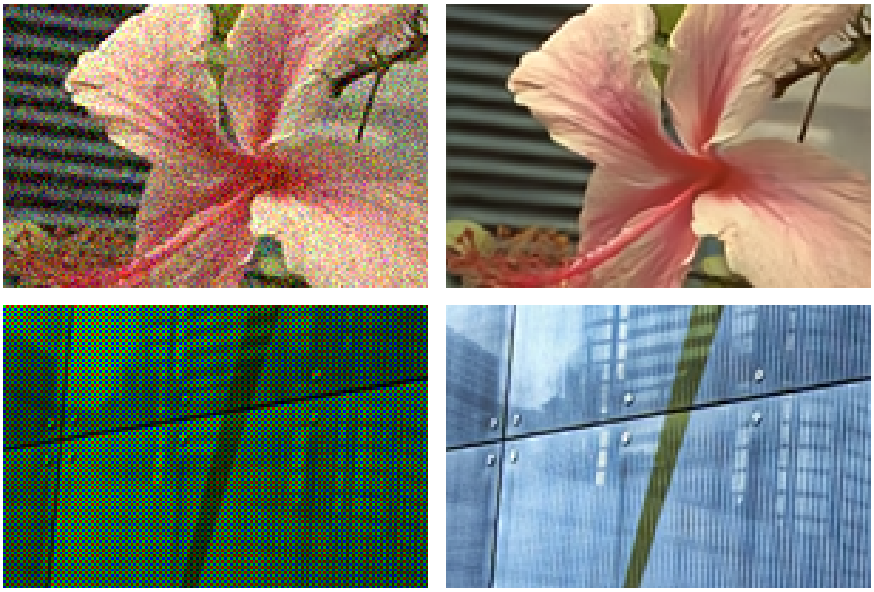


Figure 19: Images on the right are reconstructed from images on the left. Examples of restored images for denoising and demosaicking tasks (reconstructing color images from incomplete measurements made by CCD cameras).

A Flexible Framework for Designing Trainable Priors with Adaptive Smoothing and Game Encoding

Participants Bruno Lecouat, Jean Ponce, Julien Mairal.

In this paper[20], we introduce a general framework for designing and training neural network layers whose forward passes can be interpreted as solving non-smooth convex optimization problems, and whose architectures are derived from an optimization algorithm. We focus on convex games, solved by local agents represented by the nodes of a graph and interacting through regularization functions. This approach is appealing for solving imaging problems, as it allows the use of classical image priors within deep models that are trainable end to end. The priors used in this presentation include variants of total variation, Laplacian regularization, bilateral filtering, sparse coding on learned dictionaries, and non-local self similarities, see Figure 20. Our models are fully interpretable as well as parameter and data efficient. Our experiments demonstrate their effectiveness on a large diversity of tasks ranging from image denoising and compressed sensing for fMRI to dense stereo matching.

7.4 Pluri-disciplinary Research

Depth-adaptive transformer

Participants Maha Elbayad.

State of the art sequence-to-sequence models for large scale tasks perform a fixed number of computations for each input sequence regardless of whether it is easy or hard to process. In [13], we train Transformer models which can make output predictions at different stages of the network and we investigate different ways to predict how much computation is required for a particular sequence. Unlike dynamic computation in Universal Transformers, which applies the same set of layers iteratively, we apply different layers at every step to adjust both the amount of computation as well as the model capacity.

Laplacian	$\sum_{k\in\mathcal{N}_j} a_{j-k}\|\mathbf{z}_j - \mathbf{z}_k\|_2^2$
Non-local Laplacian	$\sum_{k\in\mathcal{N}_j} a_{\mathrm{NL}}^{j,k}\|\mathbf{z}_j - \mathbf{z}_k\|_2^2$
Bilateral filter (BF)	$\sum_{k\in\mathcal{N}_j} a_{\mathrm{BL}}^{j-k}\|\mathbf{z}_j - \mathbf{z}_k\|_2^2$
Total variation (TV)	$\sum_{k\in\mathcal{N}_j} a_{j-k}\|\mathbf{z}_j - \mathbf{z}_k\|_1$
Non-local total variation (NLTV)	$\sum_{k\in\mathcal{N}_j} a_{\mathrm{NL}}^{j,k}\|\mathbf{z}_j - \mathbf{z}_k\|_1$
Bilateral TV (BLTV)	$\sum_{k\in\mathcal{N}_j} a_{\mathrm{BL}}^{j-k}\|\mathbf{z}_j - \mathbf{z}_k\|_1$
Weighted ℓ_1-norm (sparse coding)	$\sum_{l=1}^{p} \lambda_l \|\mathbf{z}_j[l]\|$
Non-local group regularization	$\sum_{l=1}^{p} \lambda_l \sqrt{\sum_{k\in\mathcal{N}_j} a_{j,k}\mathbf{z}_k[l]^2}$
Variance reduction	$\|\mathbf{W}\mathbf{z}_j - \mathbf{P}_j\hat{\mathbf{y}}\|^2$

Figure 20: A non-exhaustive list of regularization functions covered by our framework.

On IWSLT German-English translation our approach matches the accuracy of a well-tuned baseline Transformer while using less than a quarter of the decoder layers.

Efficient Wait-k Models for Simultaneous Machine Translation

Participants Maha Elbayad.

Simultaneous machine translation consists in starting output generation before the entire input sequence is available. Wait-k decoders offer a simple but efficient approach for this problem. They first read k source tokens, after which they alternate between producing a target token and reading another source token. In [12], we investigate the behavior of wait-k decoding in low resource settings for spoken corpora using IWSLT datasets. We improve training of these models using unidirectional encoders, and training across multiple values of k. Experiments with Transformer and 2D-convolutional architectures show that our wait-k models generalize well across a wide range of latency levels. We also show that the 2D-convolution architecture is competitive with Transformers for simultaneous translation of spoken language.

Online Versus Offline NMT Quality: An In-depth Analysis on English–German and German–English

Participants Maha Elbayad.

We conduct in this work [15] an evaluation study comparing offline and online neural machine translation architectures. Two sequence-to-sequence models: convolutional Pervasive Attention (Elbayad et al., 2018) and attention-based Transformer (Vaswani et al., 2017) are considered. We investigate, for both architectures, the impact of online decoding constraints on the translation quality through a carefully designed human evaluation on English-German and German-English language pairs, the latter being particularly sensitive to latency constraints. The evaluation results allow us to identify the strengths and shortcomings of each model when we shift to the online setup.

Hyperspectral Images Super-Resolution via Learning High-Order Coupled Tensor Ring Representation

Participants Jocelyn Chanussot.

Hyperspectral image (HSI) super-resolution is a hot topic in remote sensing and computer vision. Recently, tensor analysis has been proven to be an efficient technology for HSI image processing. However, the existing tensor-based methods of HSI super-resolution are not able to capture the high-order correlations in HSI. In this article [4], we propose to learn a high-order coupled tensor ring (TR) representation for HSI super-resolution. The proposed method first tensorizes the HSI to be estimated into a high-order tensor in which multiscale spatial structures and the original spectral structure are represented. Then, a coupled TR representation model is proposed to fuse the low-resolution HSI (LR-HSI) and high-resolution multispectral image (HR-MSI). In the proposed model, some latent core tensors in TR of the LR-HSI and the HR-MSI are shared, and we use the relationship between the spectral core tensors to reconstruct the HSI. In addition, the graph-Laplacian regularization is introduced to the spectral core tensors to preserve the spectral information. To enhance the robustness of the proposed model, Frobenius norm regularizations are introduced to the other core tensors. Experimental results on both synthetic and real data sets show that the proposed method achieves the state-of-the-art super-resolution performance.

Cross-Attention in Coupled Unmixing Nets for Unsupervised Hyperspectral Super-Resolution

Participants Jocelyn Chanussot.

The recent advancement of deep learning techniques has made great progress on hyperspectral image super-resolution (HSI-SR). Yet the development of unsupervised deep networks remains challenging for this task. To this end, we propose in [28] a novel coupled unmixing network with a cross-attention mechanism, CUCaNet for short, to enhance the spatial resolution of HSI by means of higher-spatial-resolution multispectral image (MSI). Inspired by coupled spectral unmixing, a two-stream convolutional autoencoder framework is taken as backbone to jointly decompose MS and HS data into a spectrally meaningful basis and corresponding coefficients. CUCaNet is capable of adaptively learning spectral and spatial response functions from HS-MS correspondences by enforcing reasonable consistency assumptions on the networks. Moreover, a cross-attention module is devised to yield more effective spatial-spectral information transfer in networks. Extensive experiments are conducted on three widely used HS-MS datasets in comparison with state-of-the-art HSI-SR models, demonstrating the superiority of the CUCaNet in the HSI-SR application.

X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data

Participants Jocelyn Chanussot.

This paper [2] addresses the problem of semi-supervised transfer learning with limited cross-modality data in remote sensing. A large amount of multi-modal earth observation images, such as multispectral imagery (MSI) or synthetic aperture radar (SAR) data, are openly available on a global scale, enabling parsing global urban scenes through remote sensing imagery. However, their ability in identifying materials (pixel-wise classification) remains limited, due to the noisy collection environment and poor discriminative information as well as limited number of well-annotated training images. To this end, we propose a novel cross-modal deep-learning framework, called X-ModalNet, with three well-designed modules: the self-adversarial module, the interactive learning module, and the label propagation module, by learning to transfer more discriminative information from a small-scale hyperspectral image (HSI) into the classification task using a large-scale MSI or SAR data. Significantly, X-ModalNet generalizes well, owing to propagating labels on an updatable graph constructed by high-level features on the top of the network, yielding semi-supervised cross-modality learning. We evaluate X-ModalNet on two multi-modal remote sensing datasets (HSI-MSI and HSI-SAR) and achieve a significant improvement in comparison with several state-of-the-art methods.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

We currently have

- one CIFRE PhD student with Criteo (co-advised by J. Mairal)
- two CIFRE PhD students with Facebook: Mathilde Caron (co-advised by J. Mairal) and Lina Mezghani (co-advised by K. Alahari)
- two CIFRE PhD students with Google: Minttu Alakuijala (co-advised by J. Mairal) and Valentin Gabeur (co-advised by K. Alahari)
- one CIFRE PhD student with Valeo AI: Florent Bartoccioni (co-advised by K. Alahari)
- one CIFRE PhD student with NaverLabs Europe: Bulent Sariyildiz (co-advised by K. Alahari)

8.2 Bilateral grants with industry

In 2020, C. Schmid and K. Alahari were part of the Intel Network on Intelligent Systems in Europe, which brings together leading researchers in robotics, computer vision, motor control, and machine learning. We have been receiving financial support of about 20K euros annually since 2017 as part of this network.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Inria International Labs

GAYA

Title: Semantic and Geometric Models for Video Interpretation

Duration: 2019 - 2021

Coordinator: Karteek Alahari

Partners:

- Robotics Institute, Carnegie Mellon University (United States)

Inria contact: Karteek Alahari

Summary: The primary goal of this associate team is interpreting videos in terms of recognizing actions, understanding the human-human and human-object interactions. In the first three years, the team has started addressing the problem of learning an efficient and robust video representation to attack this challenge. GAYA will now focus on building semantic models, wherein we learn incremental, joint audio-visual models, with limited supervision, and also geometric models, where we study the geometric properties of object shapes to better recognize them. The team consists of researchers from two Inria project-teams (Thoth and WILLOW), a US university (Carnegie Mellon University [CMU]) as the main partner team, and another US university (UC Berkeley) as a secondary partner. It will allow the partners to effectively combine their respective strengths in areas such as inference and machine learning approaches for vision tasks, joint audio-visual models, large-scale learning, geometric reasoning. The main expected outcomes of this collaboration are: new machine learning algorithms for handling minimally annotated multi-modal data, large-scale public datasets for benchmarking, theoretical analysis of objects shapes and contours.

4TUNE

Title: Adaptive, Efficient, Provable and Flexible Tuning for Machine Learning

Duration: 2020 - 2022

Coordinator: Francis Bach

Partners:

- Machine Learning group, CWI (Netherlands)

Inria contact: Francis Bach

Website: http://pierre.gaillard.me/4tune/

Summary: The long-term goal of 4TUNE is to push adaptive machine learning to the next level. We aim to develop refined methods, going beyond traditional worst-case analysis, for exploiting structure in the learning problem at hand. We will develop new theory and design sophisticated algorithms for the core tasks of statistical learning and individual sequence prediction. We are especially interested in understanding the connections between these tasks and developing unified methods for both. We will also investigate adaptivity to non-standard patterns encountered in embedded learning tasks, in particular in iterative equilibrium computations.

9.2 European initiatives

9.2.1 FP7 & H2020 Projects

ERC Starting grant Solaris

Participants Julien Mairal, Andrei Kulunchakov, Dexiong Chen, Mikita Dvornik, Gregoire Mialon, Bruno Lecouat, Alexandre Zouaoui, Theo Bodrito, Gedeon Muhawenayo.

The project SOLARIS started in March 2017 for a duration of five years. The goal of the project is to set up methodological and theoretical foundations of deep learning models, in the context of large-scale data processing. The main applications of the tools developed in this project are for processing visual data, such as videos, but also structured data produced in experimental sciences, such as biological sequences.

The main paradigm used in the project is that of kernel methods and consist of building functional spaces where deep learning models live. By doing so, we want to derive theoretical properties of deep learning models that may explain their success, and also obtain new tools with better stability properties. Another work package of the project is focused on large-scale optimization, which is a key to obtain fast learning algorithms.

9.3 National initiatives

9.3.1 ANR Project DeepInFrance

Participants Jakob Verbeek, Adria Ruiz Ovejero.

DeepInFrance (Machine learning with deep neural networks) project also aims at bringing together complementary machine learning, computer vision and machine listening research groups working on deep learning with GPUs in order to provide the community with the knowledge, the visibility and the tools that brings France among the key players in deep learning. The long-term vision of Deep in France is to open new frontiers and foster research towards algorithms capable of discovering sense in data in

an automatic manner, a stepping stone before the more ambitious far-end goal of machine reasoning. The project partners are: INSA Rouen, Univ. Caen, Inria, UPMC, Aix-Marseille Univ., Univ. Nice Sophia Antipolis.

9.3.2 ANR Project AVENUE

Participants Karteek Alahari, D. Khuê Lê-Huu.

This ANR project (started in October 2018) aims to address the perception gap between human and artificial visual systems through a visual memory network for human-like interpretation of scenes. To this end, we address three scientific challenges. The first is to learn a network representation of image, video and text data collections, to leverage their inherent diverse cues. The second is to depart from supervised learning paradigms, without compromising on the performance. The third one is to perform inference with the learnt network, e.g., to estimate physical and functional properties of objects, or give cautionary advice for navigating a scene. The principal investigator is Karteek Alahari, and the project involves participants from CentraleSupelec and Ecole des Ponts in Paris.

9.4 Regional initiatives

9.4.1 3IA MIAI chair: Towards More Data Efficiency in Machine Learning

Participants Julien Mairal, Karteek Alahari, Massih-Reza Amini, Margot Selosse.

Training deep neural networks when the amount of annotated data is small or in the presence of adversarial perturbations is challenging. More precisely, for convolutional neural networks, it is possible to engineer visually imperceptible perturbations that can lead to arbitrarily different model predictions. Such a robustness issue is related to the problem of regularization and to the ability to generalizing with few training examples. Our objective is to develop theoretically-grounded approaches that will solve the data efficiency issues of such huge-dimensional models. The principal investigator is Julien Mairal.

10 Dissemination

General chair, scientific chair

Member of the organizing committees

- J. Mairal: Member of the organizing committee of the SIAM Conference on Imaging Science (IS20).
- J. Mairal: workshop co-chair for NeurIPS 2020 (+20K participants)
- K. Alahari: co-organizer of CVPR 2020 workshop on annotation-efficient learning

10.1 Scientific events: selection

Member of the conference program committees

- J. Mairal: area chair for AISTATS 2020, ECCV 2020, NeurIPS 2020, ICLR 2021, AISTATS 2021 and ICML 2021.
- K. Alahari: area chair for CVPR 2020, CVPR 2021, ECCV 2020, ICCV 2021.
- P. Gaillard: area chair for COLT 2020.

Reviewer The permanent members, postdocs and senior PhD students of the team reviewed numerous papers for international conferences in artificial intelligence, computer vision and machine learning, including AISTATS, CVPR, ECCV, ICML, ICLR, NeurIPS in 2020.

10.1.1 Journal

Member of the editorial boards

- J. Mairal: Associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), since 2021.
- J. Mairal: Associate editor of the Journal of Machine Learning Research (JMLR), since 2019.
- J. Mairal: Associate editor of the International Journal of Computer Vision, since 2015.
- J. Mairal: Associate editor of Journal of Mathematical Imaging and Vision, since 2015.
- J. Mairal: Associate editor of the SIAM Journal of Imaging Science, since 2018, until the end of 2020.
- K. Alahari: Associate editor for International Journal of Computer Vision, since 2019.
- K. Alahari: Associate editor for Computer Vision and Image Understanding, since 2018.
- J. Chanussot: Associate editor of IEEE Transactions on Image Processing
- J. Chanussot: Associate editor of IEEE Transactions on Geoscience and Remote Sening
- J. Chanussot: Associate editor of Proceedings of the IEEE.

Reviewer - reviewing activities The permanent members, postdocs and senior PhD students of the team reviewed numerous papers for international journals in computer vision (IJCV, PAMI, CVIU), machine learning (JMLR).

10.1.2 Invited talks

- J. Mairal: DataSig seminar from the Alan Turing Institute/Oxford University/UCL, 2020 (online).
- J. Mairal: Keynote speaker at ICT Innovation, Skopje, 2020 (online).
- J. Mairal: MalGA seminar at the University of Genova, 2020 (online).
- J. Mairal: seminar given at ATOS, Grenoble, 2020 (online).
- J. Mairal: seminar given at NaverLabs, Grenoble, 2020 (online).
- P. Gaillard: Talk given at the Valpred Workshop, Aussois, 2020
- P. Gaillard: Seminar given at the Potsdamer Research Seminar, Potsdam, 2020 (online)
- P. Gaillard: Seminar given at the Statify Research Seminar, Grenoble, 2020
- D. Chen: seminar given at DFKZ Heidelberg, 2020 (online).
- D. Chen: Seminar given at TUM Munich, 2020 (online).

10.1.3 Leadership within the scientific community

- J. Mairal: panel member of the New in AI workshop at NeurIPS 2020
- J. Mairal: has become ELLIS fellow
- K. Alahari: Panel discussion on computer vision for India, Vaibhav Summit, Govt. of India Initiative, Online, 2020

10.1.4 Scientific expertise

- K. Alahari: Reviewer for ERC grants, 2020

10.1.5 Research administration

- J. Mairal: jury member for assistant professor position at ENS Lyon.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Master: J. Mairal, Kernel methods for statistical learning, 17h eqTD, M2, Ecole Normale Supérieure, Cachan, France.
- Master: J. Mairal, Advanced Learning Models, 27h eqTD, M2, UGA, Grenoble.
- Master: J. Mairal, Kernel methods for statistical learning, African Masters of Machine Intelligence, Kigali, Rwanda. (online)
- Master: K. Alahari, Machine Learning for Computer Vision and Audio Processing, 6.75h eqTD, M2, UGA, Grenoble
- Master: K. Alahari, Understanding Big Visual Data, 13.5h eqTD, M2, Grenoble INP
- Master: K. Alahari, Graphical Models Inference and Learning, 18h eqTD, M2, CentraleSupelec, Paris
- Master: K. Alahari, Introduction to computer vision, 9h eqTD, M1, ENS Paris
- License: Pierre-Louis Guhur, chargé de cours sur un module de 30h d'enseignement à l'UGA auprès des licences pro MMI.
- Master: P. Gaillard, Sequential Learning, 27h eqTD, M2, Ecole Normale Supérieure, Cachan, France.
- Master: P. Gaillard, Advanced Machine Learning, 9h eqTD, M2, Telecom ParisTech, Saclay, France
- License: P. Gaillard, Introduction to Machine Learning, 15h eqTD, L3, Ecole Normale Supérieure, Paris, France
- Master: J. Chanussot, 16h Lectures on hyperspectral remote sensing, M2 level, PHELMA / ENSE3
- License: J. Chanussot, 48h Lectures on signal processing, L3 level, ENSE3
- Master: G. Mialon, Cours de Projets Informatiques, M2 MASH at Université Paris-Dauphine.
- License: H. Leterme, Analyse pour l'ingénieur (1ère année): animation d'un groupe de TD, M1, 18.5h, UGA Grenoble

10.2.2 Supervision (PhD defenses)

- PhD: Dexiong Chen, Modélisation de données structurées avec des machines profondes à noyaux et des applications en biologie computationnelle , Univ. Grenoble Alpes, 15/12/2020, director: Julien Mairal.
- PhD: Daan Wynen,Une représentation archétypale de style artistique : résumer et manipuler des styles artistiques d'une façon interprétable , Univ. Grenoble Alpes, 9/12/2020, directors: Cordelia Schmid and Julien Mairal.
- PhD: Andrei Kulunchakov, Optimisation stochastique pour l'apprentissage machine à grande échelle : réduction de la variance et accélération, Univ. Grenoble Alpes, 3/12/2020, directors: Anatoli Juditsky and Julien Mairal.
- PhD: Thomas Lucas, Modèles génératifs profonds : sur-généralisation et abandon de mode, Univ. Grenoble Alpes, 25/9/2020, directors: Karteek Alahari and Jakob Verbeek.

10.2.3 Juries

- J. Mairal: Jury member for the PhD thesis of Yaroslav Averyanov, Université de Lille
- J. Mairal: Jury member for the PhD thesis of Dmitry Grishcnenko, Université Grenoble-Alpes
- J. Mairal: Jury member for the PhD thesis of Boris Muzellec, ENSAE
- J. Mairal: Reviewer for the PhD thesis of Vincent Prost, Université d'Evry
- J. Mairal: Jury member for the PhD thesis of Pierre Laforgue, Telecom ParisTech
- J. Mairal: Reviewer for the PhD thesis of Leonard Berrada, Oxford University
- J. Mairal: member of the comité de suivi de thèse of Olga PERMIAKOVA, Université Grenoble Alpes
- J. Mairal: member of the comité de suivi de thèse of Tayeb ZARROUK, Université Grenoble Alpes
- K. Alahari: Jury member of the thesis proposal committee of Allison Del Giorno, Carnegie Mellon University
- K. Alahari: member of the comité de suivi de thèse of Abid ALI, Université Côte d'Azur
- K. Alahari: member of the comité de suivi de thèse of Miguel SOLINAS, Université Grenoble Alpes

11 Scientific production

11.1 Publications of the year

International journals

[1] P. Gratier, J. Pety, E. Bron, A. Roueff, J. H. Orkisz, M. Gerin, V. De Souza Magalhaes, M. Gaudel, M. Vono, S. Bardeau, J. Chanussot, P. Chainais, J. R. Goicoechea, V. V. Guzmán, A. Hughes, J. Kainulainen, D. Languignon, J. Le Bourlot, F. Le Petit, F. Levrier, H. Liszt, N. Peretto, E. Roueff and A. Sievers. 'Quantitative inference of the H2 column densities from 3 mm molecular emission: case study towards Orion B'. In: *Astronomy and Astrophysics - A&A* 645.January 2021 (Jan. 2021), A27. DOI: 10.1051/0004-6361/202037871. URL: https://hal.archives-ouvertes.fr/hal-03017404.

[2] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot and X. x. Zhu. 'X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data'. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 167 (Sept. 2020), pp. 12–23. DOI: 10.1016/j.isprsjprs.2020.06.014. URL: https://hal.archives-ouvertes.fr/hal-03142183.

[3] A. Kulunchakov and J. Mairal. 'Estimate Sequences for Stochastic Composite Optimization: Variance Reduction, Acceleration, and Robustness to Noise'. In: *Journal of Machine Learning Research* 21.155 (July 2020), pp. 1–52. URL: https://hal.inria.fr/hal-01993531.

[4] Y. Xu, Z. Wu, J. Chanussot and Z. Wei. 'Hyperspectral Images Super-Resolution via Learning High-Order Coupled Tensor Ring Representation'. In: *IEEE Transactions on Neural Networks and Learning Systems* 31.11 (Nov. 2020), pp. 4747–4760. DOI: 10.1109/TNNLS.2019.2957527. URL: https://hal.archives-ouvertes.fr/hal-03142166.

International peer-reviewed conferences

[5] R. Berthier, F. Bach and P. Gaillard. 'Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model'. In: NeurIPS '20 - 34th International Conference on Neural Information Processing Systems. Vol. 33. Advances in Neural Information Processing Systems. Vancouver, Canada: https://proceedings.neurips.cc/paper/2020/hash/1b33d16fc562464579b7199ca3114982-Abstract.html, 7th Dec. 2020, pp. 2576–2586. URL: https://hal.archives-ouvertes.fr/hal-02866755.

[6] A. Bietti and F. Bach. 'Deep Equals Shallow for ReLU Networks in Kernel Regimes'. In: ICLR 2021 - International Conference on Learning Representations. Virtual, Austria, 3rd May 2021, pp. 1–22. URL: https://hal.inria.fr/hal-02963250.

[7] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski and A. Joulin. 'Unsupervised Learning of Visual Features by Contrasting Cluster Assignments'. In: 34th Conference on Neural Information Processing Systems, NeurIPS'20. Vol. 33. Advances in Neural Information Processing Systems. Virtual-only, United States: https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html, 6th Dec. 2020, pp. 9912–9924. URL: https://hal.archives-ouvertes.fr/hal-02883765.

[8] D. Chen, L. Jacob and J. Mairal. 'Convolutional Kernel Networks for Graph-Structured Data'. In: *Proceedings of the37thInternational Conference on MachineLearning*. ICML 2020 - 37th International Conference on Machine Learning. Vol. 119. Proceedings of Machine Learning Research. Vienna, Austria: http://proceedings.mlr.press/v119/chen20h.html, 6th July 2020, pp. 1576–1586. URL: https://hal.archives-ouvertes.fr/hal-02504965.

[9] A. Dasgupta, C. V. Jawahar and K. Alahari. 'Context Aware Group Activity Recognition'. In: ICPR 2020 - International Conference on Pattern Recognition. Milan (Virtual), Italy, 10th Jan. 2021, pp. 1–8. URL: https://hal.archives-ouvertes.fr/hal-02987414.

[10] A. Dave, T. Khurana, P. Tokmakov, C. Schmid and D. Ramanan. 'TAO: A Large-Scale Benchmark for Tracking Any Object'. In: ECCV 2020 - European Conference on Computer Vision. Vol. 12350. Lecture Notes in Computer Science. Glasgow / Virtual, United Kingdom, 29th Oct. 2020, pp. 436–454. DOI: 10.1007/978-3-030-58558-7_26. URL: https://hal.archives-ouvertes.fr/hal-02951747.

[11] N. Dvornik, C. Schmid and J. Mairal. 'Selecting Relevant Features from a Multi-domain Representation for Few-shot Classification'. In: ECCV 2020 - European Conference on Computer Vision. Vol. 12355. Lecture Notes in Computer Science. Glasgow / Virtual, United Kingdom, 7th Nov. 2020, pp. 769–786. DOI: 10.1007/978-3-030-58607-2_45. URL: https://hal.archives-ouvertes.fr/hal-02513241.

[12] M. Elbayad, L. Besacier and J. Verbeek. 'Efficient Wait-k Models for Simultaneous Machine Translation'. In: Interspeech 2020 - Conference of the International Speech Communication Association. Shangai (Virtual Conf), China, 25th Oct. 2020, pp. 1461–1465. DOI: 10.21437/Interspeech.2020-1241. URL: https://hal.archives-ouvertes.fr/hal-02962195.

[13] M. Elbayad, J. Gu, E. Grave and M. Auli. 'Depth-adaptive Transformer'. In: ICLR 2020 - Eighth International Conference on Learning Representations. Addis Ababa, Ethiopia: https://iclr.cc/, 20th Dec. 2019, pp. 1–14. URL: https://hal.inria.fr/hal-02422914.

[14] M. Elbayad, H. Nguyen, F. Bougares, N. Tomashenko, A. Caubrière, B. Lecouteux, Y. Estève and L. Besacier. 'ON-TRAC Consortium for End-to-End and Simultaneous Speech Translation Challenge Tasks at IWSLT 2020'. In: Proceedings of the 17th International Conference on Spoken Language Translation. Seattle, WA, United States, 9th July 2020, pp. 35–43. DOI: 10.18653/v1/2020.iwslt-1.2. URL: https://hal.archives-ouvertes.fr/hal-02895893.

[15] M. Elbayad, M. Ustaszewski, E. Esperança-Rodier, F. Brunet Manquat, J. Verbeek and L. Besacier. 'Online Versus Offline NMT Quality: An In-depth Analysis on English–German and German–English'. In: COLING 2020 - 28th International Conference on Computational Linguistics. Virtual, Spain, 8th Dec. 2020, pp. 5047–5058. DOI: 10.18653/v1/2020.coling-main.443. URL: https://hal.archives-ouvertes.fr/hal-02991539.

[16] V. Gabeur, C. Sun, K. Alahari and C. Schmid. 'Multi-modal Transformer for Video Retrieval'. In: European Conference on Computer Vision (ECCV). Vol. 12349. Lecture Notes in Computer Science. Glasgow, United Kingdom, 29th Oct. 2020, pp. 214–229. DOI: 10.1007/978-3-030-58548-8_13. URL: https://hal.inria.fr/hal-02903209.

[17] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys and C. Schmid. 'Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction'. In: CVPR 2020 - IEEE Conference on Computer Vision and Pattern Recognition. Seattle / Virtual, United States, 14th June 2020, pp. 568–577. DOI: 10.1109/CVPR42600.2020.00065. URL: https://hal.inria.fr/hal-02557112.

[18] E. Iakovleva, J. Verbeek and K. Alahari. 'Meta-Learning with Shared Amortized Variational Inference'. In: ICML 2020 - 37th International Conference on Machine Learning. Vol. 119. Proceedings of Machine Learning Research. Vienna (Online), Austria: http://proceedings.mlr.press/v119/iakovleva20a.html, 12th July 2020, pp. 4572–4582. URL: https://hal.inria.fr/hal-02925830.

[19] R. Klokov, E. Boyer and J. Verbeek. 'Discrete Point Flow Networks for Efficient Point Cloud Generation'. In: 16th European Conference on Computer Vision - ECCV 2020. Vol. 12368. Lecture Notes in Computer Science. Glasgow, United Kingdom, Aug. 2020, pp. 694–710. DOI: 10.1007/978-3-030-58592-1_41. URL: https://hal.archives-ouvertes.fr/hal-02903163.

[20] B. Lecouat, J. Ponce and J. Mairal. 'A Flexible Framework for Designing Trainable Priors with Adaptive Smoothing and Game Encoding'. In: NeurIPS '20 - 34th International Conference on Neural Information Processing Systems. Vol. 33. Advances in Neural Information Processing Systems. Vancouver, France: https://proceedings.neurips.cc/paper/2020/hash/b4edda67f0f57e218a8e766927e3e5c5-Abstract.html, 6th Oct. 2020, pp. 15664–15675. URL: https://hal.archives-ouvertes.fr/hal-02881924.

[21] B. Lecouat, J. Ponce and J. Mairal. 'Fully Trainable and Interpretable Non-Local Sparse Models for Image Restoration'. In: ECCV 2020 - European Conference on Computer Vision. Vol. 12367. Lecture Notes in Computer Science. Glasgow / Virtual, United Kingdom, 17th Nov. 2020, pp. 238–254. DOI: 10.1007/978-3-030-58542-6_15. URL: https://hal.inria.fr/hal-02414291.

[22] X. Li, S. Wang, Y. Zhao, J. Verbeek and J. Kannala. 'Hierarchical Scene Coordinate Classification and Regression for Visual Localization'. In: CVPR 2020 - IEEE Conference on Computer Vision and Pattern Recognition. Seattle, United States, 5th Aug. 2020, pp. 11980–11989. DOI: 10.1109/CVPR42600.2020.01200. URL: https://hal.inria.fr/hal-02384675.

[23] G. Mialon, D. Chen, A. D'Aspremont and J. Mairal. 'A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention'. In: ICLR 2021 - The Ninth International Conference on Learning Representations. Virtual, France, 4th May 2021. URL: https://hal.archives-ouvertes.fr/hal-02883436.

[24] G. Mialon, A. D'Aspremont and J. Mairal. 'Screening Data Points in Empirical Risk Minimization via Ellipsoidal Regions and Safe Loss Functions'. In: AISTATS 2020 - 23rd International Conference on Artificial Intelligence and Statistics. Vol. 108. Proceedings of Machine Learning Research. Palermo / Virtual, Italy: http://proceedings.mlr.press/v108/mialon20a.html, 26th Aug. 2020, pp. 3610–3620. URL: https://hal.archives-ouvertes.fr/hal-02395624.

[25] A. Sablayrolles, M. Douze, C. Schmid and H. Jégou. 'Radioactive Data: Tracing Through Training'. In: ICML 2020 - Thirty-seventh International Conference on Machine Learning. Vol. 119. Proceedings of Machine Learning Research. Vienna / Virtual, Austria, 12th July 2020, pp. 8326–8335. URL: https://hal.inria.fr/hal-02954159.

[26] R. Strudel, R. Garcia, J. Carpentier, J.-P. Laumond, I. Laptev and C. Schmid. 'Learning Obstacle Representations for Neural Motion Planning'. In: CoRL 2020 - Conference on Robot Learning. Cambridge MA / Virtual, United States, 16th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02944348.

[27] R. Strudel, A. Pashevich, I. Kalevatykh, I. Laptev, J. Sivic and C. Schmid. 'Learning to combine primitive skills: A step towards versatile robotic manipulation'. In: ICRA 2020 - IEEE International Conference on Robotics and Automation. Paris / Virtuel, France, 31st May 2020. DOI: 10.1109/ICRA40945.2020.9196619. URL: https://hal.archives-ouvertes.fr/hal-02274969.

[28] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu and Z. Xu. 'Cross-Attention in Coupled Unmixing Nets for Unsupervised Hyperspectral Super-Resolution'. In: *European Conference on Computer Vision (ECCV)*. ECCV 2020 - 16th European Conference on Computer Vision. Vol. 12374. Lecture Notes i Computer Science. Glasgow, United Kingdom, 7th Oct. 2020, pp. 208–224. DOI: 10.1007/978-3-030-58526-6_13. URL: https://hal.archives-ouvertes.fr/hal-03142195.

Doctoral dissertations and habilitation theses

[29] T. Lucas. 'Deep generative models : over-generalisation and mode-dropping'. Université Grenoble Alpes [2020-....], 25th Sept. 2020. URL: https://tel.archives-ouvertes.fr/tel-03102554.

[30] D. Wynen. 'An Archetypal Representation of Artistic Style : Summarizing and manipulating artistic style in an interpretable manner'. Université Grenoble Alpes [2020-....], 9th Dec. 2020. URL: https://tel.archives-ouvertes.fr/tel-03184810.

Reports & preprints

[31] R. Berthier, F. Bach, N. Flammarion, P. Gaillard and A. Taylor. *A Continuized View on Nesterov Acceleration*. 11th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03138823.

[32] M. Caron, A. Morcos, P. Bojanowski, J. Mairal and A. Joulin. *Pruning Convolutional Neural Networks with Self-Supervision*. 29th June 2020. URL: https://hal.archives-ouvertes.fr/hal-02883772.

[33] A. B. Juditsky, A. Kulunchakov and H. Tsyntseus. *Sparse recovery by reduced variance stochastic approximation*. 30th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03185516.

[34] L. Mezghani, S. Sukhbaatar, T. Lavril, O. Maksymets, D. Batra, P. Bojanowski and K. Alahari. *Memory-Augmented Reinforcement Learning for Image-Goal Navigation*. 14th Jan. 2021. URL: https://hal.inria.fr/hal-03110875.

[35] A. Raj, P. Gaillard and C. Saad. *Non-stationary Online Regression*. 10th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02998781.

[36] A. Ruiz and J. Verbeek. *Distilled Hierarchical Neural Ensembles with Adaptive Inference Cost*. 6th Mar. 2020. URL: https://hal.inria.fr/hal-02500660.

[37] M. B. Sariyildiz, Y. Kalantidis, D. Larlus and K. Alahari. *Concept Generalization in Visual Representation Learning*. 2020. URL: https://hal.inria.fr/hal-03110632.

[38] G. Varol, I. Laptev, C. Schmid and A. Zisserman. *Synthetic Humans for Action Recognition from Unseen Viewpoints*. 11th Jan. 2020. URL: https://hal.inria.fr/hal-02435731.

[39] H. Zenati, A. Bietti, M. Martin, E. Diemert and J. Mairal. *Counterfactual Learning of Continuous Stochastic Policies*. 29th June 2020. URL: https://hal.archives-ouvertes.fr/hal-02883423.