RESEARCH CENTRE

Inria Centre
at Université Grenoble Alpes

2023
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

Inria

Contents

Project-Team THOTH **1**

1 Team members, visitors, external collaborators **2**

2 Overall objectives **3**

3 Research program **4**
3.1 Designing and learning structured models . . . 4
3.2 Learning of visual models from minimal supervision . . . 5
3.3 Large-scale learning and optimization . . . 6

4 Application domains **7**
4.1 Visual applications . . . 7
4.2 Pluri-disciplinary research . . . 7

5 Highlights of the year **8**
5.1 Awards . . . 8

6 New software, platforms, open data **8**
6.1 New software . . . 8
6.1.1 Cyanure . . . 8
6.1.2 HySUPP . . . 8
6.1.3 t-ReX . . . 8

7 New results **9**
7.1 Visual Recognition . . . 9
7.2 Statistical Machine Learning . . . 16
7.3 Pluri-disciplinary Research . . . 19
7.4 Optimization . . . 26

8 Bilateral contracts and grants with industry **29**
8.1 Bilateral contracts with industry . . . 29

9 Partnerships and cooperations **30**
9.1 International initiatives . . . 30
9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program . . . 30
9.2 International research visitors . . . 31
9.2.1 Visits of international scientists . . . 31
9.3 European initiatives . . . 31
9.3.1 ERC Project APHELEIA . . . 31
9.4 National initiatives . . . 31
9.4.1 ANR Project AVENUE . . . 31
9.4.2 ANR Project BONSAI . . . 32
9.4.3 MIAI chair: Towards More Data Efficiency in Machine Learning . . . 32
9.4.4 MIAI chair: Learning Visual Representations from Interaction for Robot Manipulation Tasks . . . 32
9.4.5 PEPR project Numpex . . . 33
9.4.6 PEPR project Origins . . . 33

10 Dissemination **33**
10.1 Promoting scientific activities . . . 33
10.1.1 Scientific events: organisation . . . 33
10.1.2 Scientific events: selection . . . 33
10.1.3 Journal . . . 33

10.1.4 Invited talks . . . 34
10.1.5 Scientific expertise . . . 34
10.1.6 Research administration . . . 34
10.2 Teaching - Supervision - Juries . . . 34
10.2.1 Teaching . . . 34
10.2.2 Supervision . . . 35
10.2.3 Juries . . . 35
10.2.4 Internal or external Inria responsibilities . . . 36
10.2.5 Interventions . . . 36

11 Scientific production **36**
11.1 Publications of the year . . . 36

Project-Team THOTH

Creation of the 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, Senior Researcher, DR since Oct 2023, détachement du corps des Mines, HDR]
- Karteek Alahari [Inria, Senior Researcher, DR since Oct 2023, HDR]
- Michael Arbel [Inria, Researcher]
- Pia Bideau [UGA, Chair, from Oct 2023, MIAI junior chair]
- Jocelyn Chanussot [Inria, Senior Researcher, from Sep 2023, détachement Grenoble INP, HDR]
- Pierre Gaillard [Inria, Researcher]
- Hadrien Hendrikx [Inria, Researcher]

Post-Doctoral Fellows

- Heeseung Kwon [Inria until Aug 2023, then UGA, Post-Doctoral Fellow]
- Huu Dien Khue Le [Inria, until May 2023]
- Romain Menegaux [UGA, Post-Doctoral Fellow, until Nov 2023]

PhD Students

- Loic Arbez [GRENOBLE INP, from Dec 2023]
- Florent Bartoccioni [VALEO, until Aug 2023, CIFRE]
- Tariq Berrada Ifriqi [FACEBOOK, CIFRE, from May 2023]
- Jules Bourcier [PRELIGENS]
- Timothee Darcet [NAVER LABS, CIFRE]
- Camila Fernadez Morales [NOKIA BELL LABS, until Nov 2023, CIFRE]
- Renaud Gaucher [ECOLE POLY PALAISEAU, from May 2023]
- Emmanuel Jehanno [Inria, from Oct 2023]
- Zhiqi Kang [Inria]
- Hubert Leterme [UGA, until Jun 2023]
- Paul Liautaud [SORBONNE UNIVERSITE]
- Bianca Marin Moreno [EDF, CIFRE]
- Juliette Marrie [NAVER LABS, CIFRE]
- Lina Mezghani [FACEBOOK, until Aug 2023]
- Ieva Petrulionyte [UGA, from Sep 2023]
- Mert Sariyildiz [NAVER LABS, until Aug 2023]
- Houssam Zenati [CRITEO, until Jun 2023]
- Julien Zhou [CRITEO, CIFRE]
- Alexandre Zouaoui [Inria]

Technical Staff

- Loic Arbez [Inria, Engineer, until Nov 2023]
- Emmanuel Jehanno [UGA, Engineer, until Sep 2023]
- Thomas Ryckeboer [Inria, Engineer, affiliated to SED]
- Romain Seailles [Inria, Engineer, from Dec 2023]

Interns and Apprentices

- Ieva Petrulionyte [UGA, Intern, from Feb 2023 until Aug 2023]

Administrative Assistant

- Nathalie Gillot [Inria]

Visiting Scientist

- Nassim Ait Ali Braham [DLR, from Apr 2023 until Oct 2023]

2 Overall objectives

Thoth is a computer vision and machine learning team. Our initial goal was to develop machine learning models for analyzing the massive amounts of visual data that are currently available on the web. Then, the focus of the team has become more diverse. More precisely, we share a common objective of developing machine learning models that are robust and efficient (in terms of computational cost and data requirements).

Our main research directions are the following ones:

- **visual understanding from limited annotations and data:** Many state-of-the-art computer vision models are typically trained on a huge corpus of fully annotated data. We want to reduce the cost by developing new algorithms for unsupervised, self-supervised, continual, or incremental learning.
- **efficient deep learning models, from theory to applications**: We want to invent a new generation of machine learning models (in particular deep learning) with theoretical guarantees, efficient algorithms, and a wide range of applications. We develop for instance models for images, videos, graphs, or sequences.
- **statistical machine learning and optimization:** we are also developing efficient machine learning methods, with a focus on stochastic optimization for processing large-scale data, and online learning.
- **pluri-disciplinary collaborations:** Machine learning being at the crossing of several disciplines, we have successfully conducted collaborations in scientific domains that are relatively far from our domains of expertise. These fields are producing massive amounts of data and are in dire needs of efficient tools to make predictions or interpretations. For example, we have had the chance to collaborate with many colleagues from natural language processing, robotics, neuroimaging, computational biology, genomics, astrophysics for exoplanet detections, and we are currently involved in several remote sensing and hyperspectral imaging projects thanks to Jocelyn Chanussot (hosted by Thoth from 2019 to 2022).

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, recovering scene geometry. 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, (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 two 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 second 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 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.

- **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

[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.

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 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.

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 autonomous driving, to 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

Julien Mairal received the young researcher award from Inria - Academie des Sciences.

6 New software, platforms, open data

6.1 New software

6.1.1 Cyanure

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

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: packaging on conda and pipy + various improvements

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

Contact: Julien Mairal

Participants: Julien Mairal, Thomas Ryckeboer

6.1.2 HySUPP

Keyword: Image processing

Functional Description: Toolbox for hyperspectral unmixing. This is a Python package described in the following publication https://hal.science/hal-04180307v2/document

URL: https://github.com/BehnoodRasti/HySUPP

Contact: Julien Mairal

6.1.3 t-ReX

Name: Software package for improving generalization in supervised models

Keywords: Generalization, Supervised models

Scientific Description: We consider the problem of training a deep neural network on a given classification task, e.g., ImageNet-1K (IN1K), so that it excels at both the training task as well as at other (future) transfer tasks. These two seemingly contradictory properties impose a trade-off between improving the model's generalization and maintaining its performance on the original task. Models trained with self-supervised learning tend to generalize better than their supervised counterparts for transfer learning, yet, they still lag behind supervised models on IN1K. In this paper, we propose a supervised learning setup that leverages the best of both worlds. We extensively analyze supervised training using multi-scale crops for data augmentation and an expendable projector head, and reveal that the design of the projector allows us to control the trade-off between performance on the training task and transferability. We further replace the last layer of class weights with class prototypes computed on the fly using a memory bank and derive two models: t-ReX that achieves a new state of the art for transfer learning and outperforms top methods such as DINO and PAWS on IN1K, and t-ReX* that matches the highly optimized RSB-A1 model on IN1K while performing better on transfer tasks. Code and pretrained models: https://europe.naverlabs.com/t-rex

Functional Description: In this repository, we provide: - Several pretrained t-ReX and t-ReX* models (proposed in Sariyildiz et al., ICLR 2023) in PyTorch. - Code for training our t-ReX and t-ReX* models on the ImageNet-1K dataset in PyTorch. - Code for running transfer learning evaluations of pretrained models via linear classification over pre-extracted features on 16 downstream datasets.

URL: https://github.com/naver/trex

Contact: Karteek Alahari

7 New results

7.1 Visual Recognition

Semi-supervised learning made simple with self-supervised clustering

Participants: Enrico Fini, Pietro Astolfi, Karteek Alahari, Xavier Alameda-Pineda, Julien Mairal, Moin Nabi, Elisa Ricci.

Self-supervised learning models have been shown to learn rich visual representations without requiring human annotations. However, in many real-world scenarios, labels are partially available, motivating a recent line of work on semi-supervised methods inspired by self-supervised principles. In [9], we propose a conceptually simple yet empirically powerful approach to turn clusteringbased self-supervised methods such as SwAV or DINO into semi-supervised learners. More precisely, we introduce a multi-task framework merging a supervised objective using ground-truth labels and a self-supervised objective relying on clustering assignments with a single cross-entropy loss. This approach may be interpreted as imposing the cluster centroids to be class prototypes. Despite its simplicity, we provide empirical evidence that our approach is highly effective and achieves state-of-the-art performance on CIFAR100 and ImageNet. This method is illustrated in Figure 1

SLACK: Stable Learning of Augmentations with Cold-start and KL regularization

Participants: Juliette Marrie, Michael Arbel, Diane Larlus, Julien Mairal.

Data augmentation is known to improve the generalization capabilities of neural networks, provided that the set of transformations is chosen with care, a selection often performed manually. Automatic data augmentation aims at automating this process. However, most recent approaches still rely on some prior information; they start from a small pool of manually-selected default transformations that are either used to pretrain the network or forced to be part of the policy learned by the automatic data augmentation algorithm. In [15], we propose to directly learn the augmentation policy without leveraging such prior knowledge. The resulting bilevel optimization problem becomes more challenging due to the larger search space and the inherent instability of bilevel optimization algorithms. To mitigate these issues (i) we follow a successive cold-start strategy with a Kullback-Leibler regularization, and (ii) we parameterize magnitudes as continuous distributions. Our approach leads to competitive results on standard benchmarks despite a more challenging setting, and generalizes beyond natural images. Examples of learned transformations are presented in Figure 2.

Vision Transformers Need Registers

Participants: Timothee Darcet, Maxime Oquab, Julien Mairal, Piotr Bojanowski.

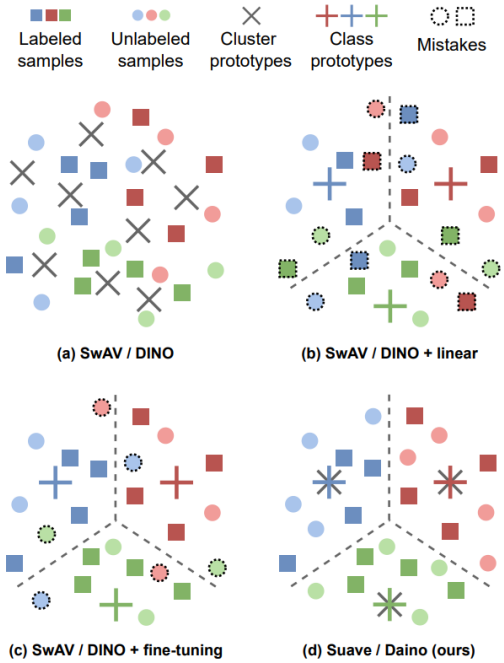


Figure 1: Schematic illustration of the motivation behind the proposed semi-supervised framework. (a) Self-supervised clustering methods like SwAV and DINO compute cluster prototypes that are not necessarily well aligned with semantic categories, but they do not require labeled data. (b) Adding a linear classifier provides class prototypes, but the labeled (and unlabeled) samples are not always correctly separated. (c) Fine-tuning can help separating labeled data. (d) Our framework learns cluster prototypes that are aligned with class prototypes thus correctly separating both labeled and unlabeled data.

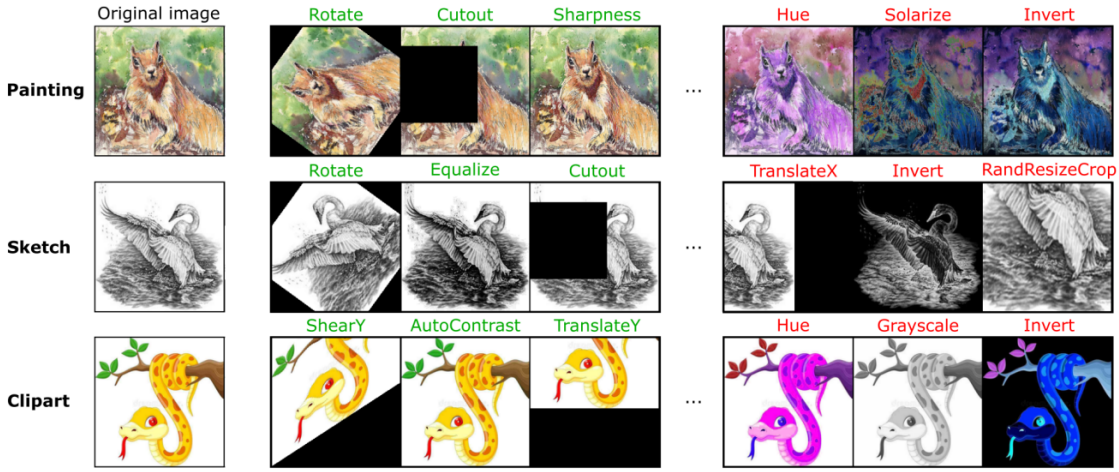


Figure 2: For different domains of the DomainNet dataset (one per line), we show an image from that domain (left) and that image transformed using the three most likely (middle) and the three least likely (right) augmentations for that domain, as estimated by SLACK.

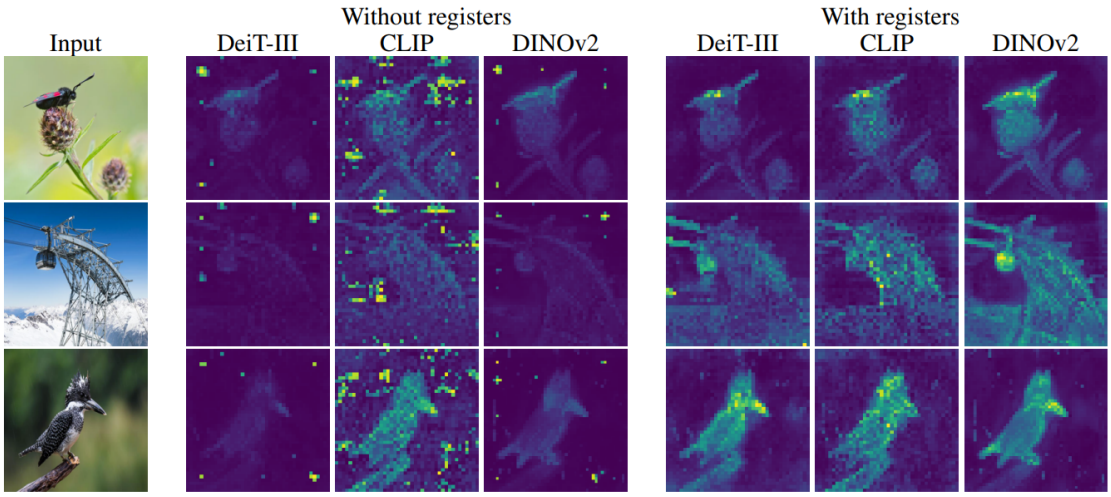


Figure 3: Register tokens enable interpretable attention maps in all vision transformers, similar to the original DINO method (Caron et al., 2021). Attention maps are calculated in high resolution for better visualisation.

Transformers have recently emerged as a powerful tool for learning visual representations. In this paper, we identify and characterize artifacts in feature maps of both supervised and self-supervised ViT networks. The artifacts correspond to high-norm tokens appearing during inference primarily in low-informative background areas of images, that are repurposed for internal computations. In [24], we propose a simple yet effective solution based on providing additional tokens to the input sequence of the Vision Transformer to fill that role. We show that this solution fixes that problem entirely for both supervised and self-supervised models (see Fig 3), sets a new state of the art for self-supervised visual models on dense visual prediction tasks, enables object discovery methods with larger models, and most importantly leads to smoother feature maps and attention maps for downstream visual processing.

LiDARTouch: Monocular metric depth estimation with a few-beam LiDAR

Participants: Florent Bartoccioni, Eloi Zablocki, Patrick Pérez, Matthieu Cord, Karteek Alahari.

In this paper [1], we address the task of monocular depth prediction, a key component of many autonomous systems, by self-supervised deep learning. Existing methods are either fully-supervised with an additional expensive LiDAR (32 or 64 beams) as input or self-supervised with camera-only methods, much cheaper, but suffering from scale ambiguity and infinite depth problems. In contrast, we introduce LiDARTouch, a novel method combining a monocular camera with a cheap minimal 4-beam LiDAR input, typical of laser scanners currently used in the automotive industry. We introduce a new self-supervision scheme to leverage this very sparse LiDAR input at three complementary levels. While being extremely sparse, we show that the use of a few-beam LiDAR alleviate the scaling ambiguity and infinite depth problems that camera-only methods suffer from. We also reach competitive performances with respect to fully-supervised depth completion methods while being significantly cheaper and more annotation friendly. Our method can be trained on any domain with no modification, and it can thus bring accurate and metric depth estimation at a vehicle fleet scale. In Figure 4, we present three examples along with selected close-ups highlighting the infinite-depth problem. For example, on the leftmost column, we observe a typical 'hole' in the depth map where previous 'Depth Estimation' method estimates a vehicle three times as far as its true distance. In contrast, by leveraging small touches of LIDAR we disambiguate the prediction and can accurately and safely handle moving objects with no relative motion, typical of cars in fluid traffic.

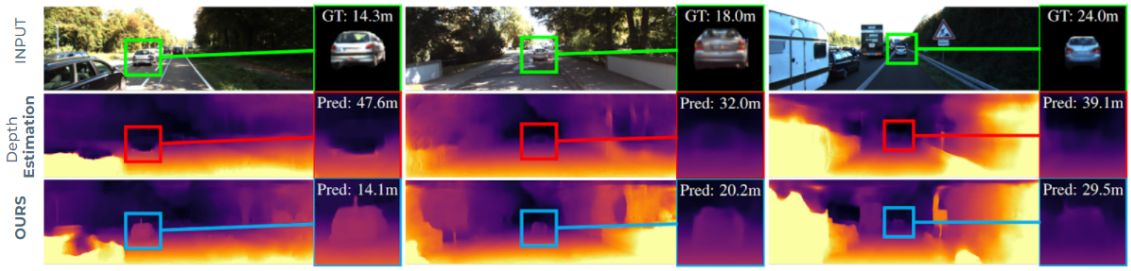


Figure 4: **Mitigation of the infinite-depth problem.** Self-supervised image-only approaches tend to predict objects with no relative-motion at an infinite depth, as indicated by the hole in the depth close-up (red). In contrast, our LiDARTouch framework estimates the depth of these vehicles, as shown in the green close-up

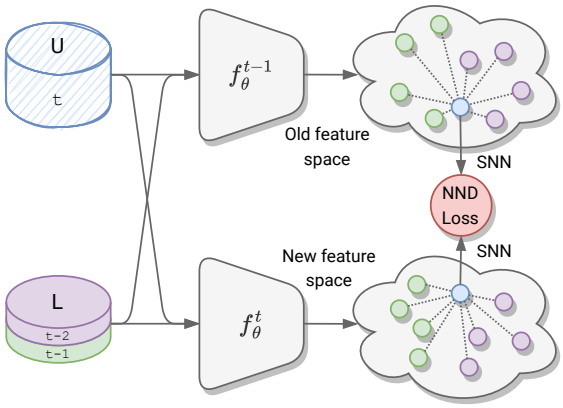


Figure 5: Illustration of our soft nearest-neighbor distillation loss.

A soft nearest-neighbor framework for continual semi-supervised learning

Participants: Zhiqi Kang, Enrico Fini, Moin Nabi, Elisa Ricci, Karteek Alahari.

Despite significant advances, the performance of state-of-the-art continual learning approaches hinges on the unrealistic scenario of fully labeled data. In [13], we tackle this challenge and propose an approach for continual semi-supervised learning—a setting where not all the data samples are labeled. An underlying issue in this scenario is the model forgetting representations of unlabeled data and overfitting the labeled ones. We leverage the power of nearest-neighbor classifiers to non-linearly partition the feature space and learn a strong representation for the current task, as well as distill relevant information from previous tasks, shown in Figure 5. We perform a thorough experimental evaluation and show that our method outperforms all the existing approaches by large margins, setting a strong state of the art on the continual semi-supervised learning paradigm. For example, on CIFAR100 we surpass several others even when using at least 30 times less supervision (0.8% vs. 25% of annotations).

No Reason for No Supervision: Improved Generalization in Supervised Models

Participants: Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, Diane Larlus.

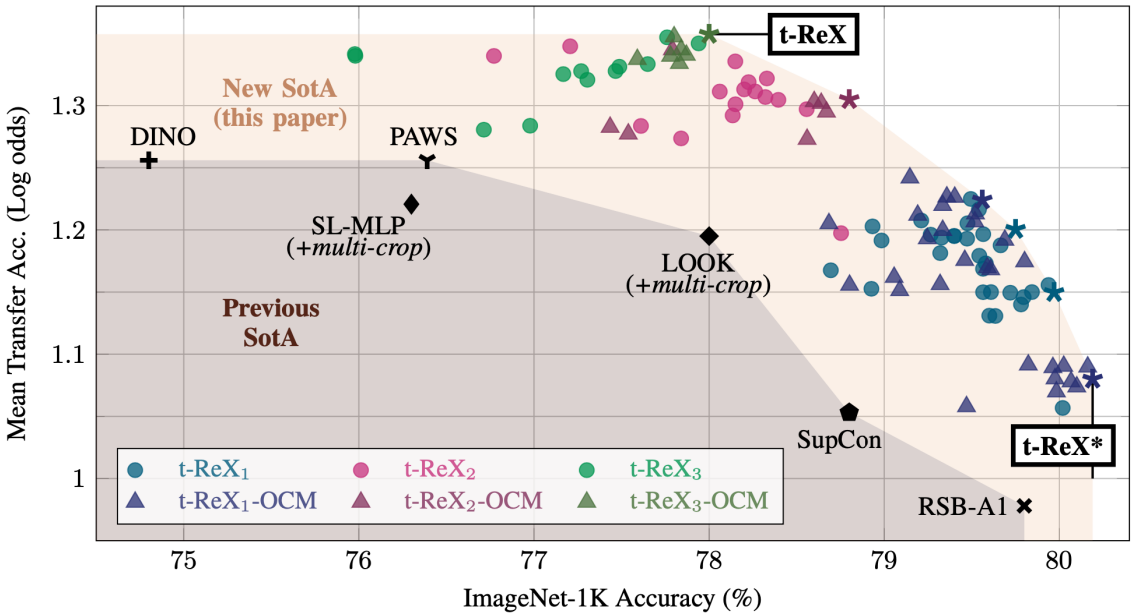


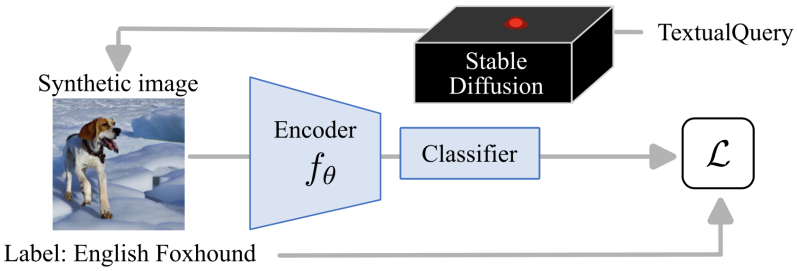
Figure 6: Comparison on the training task vs. transfer task performance for ResNet50 encoders. We report IN1K (Top-1 accuracy) and transfer performance (log odds) averaged over 13 datasets (5 ImageNet-CoG levels, Aircraft, Cars196, DTD, EuroSAT, Flowers, Pets, Food101 and SUN397) for a large number of our models trained with the supervised training setup presented in this work on the convex hull are denoted by stars. We compare to the following state-of-the-art (SotA) models: Supervised: RSB-A1, SupCon, SL-MLP and LOOK with multi-crop; self-supervised: DINO; semi-supervised: PAWS.

We consider the problem of training a deep neural network on a given classification task, e.g., ImageNet-1K (IN-1K), so that it excels at that task as well as at other (future) transfer tasks. These two seemingly contradictory properties impose a trade-off between improving the model's generalization while maintaining its performance on the original task. Models trained with self-supervised learning (SSL) tend to generalize better than their supervised counterparts for transfer learning; yet, they still lag behind supervised models on IN-1K. In this work [18], we propose a supervised learning setup that leverages the best of both worlds. We enrich the common supervised training framework using two key components of recent SSL models: Multi-scale crops for data augmentation and the use of an expendable projector head. We replace the last layer of class weights with class *prototypes* computed on the fly using a memory bank. We show in our experiments (see Figure 6) that these three improvements lead to a more favorable trade-off between the IN-1K training task and 13 transfer tasks. Over all the explored configurations, we single out two models: t-ReX that achieves a new state of the art for transfer learning and outperforms top methods such as DINO and PAWS on IN-1K, and t-ReX* that matches the highly optimized RSB model on IN-1K while performing better on transfer tasks.

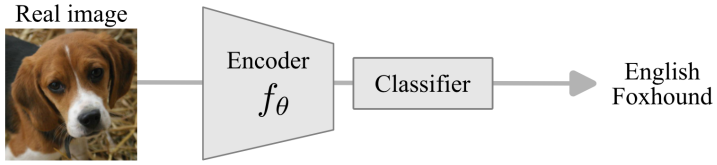
Fake it till you make it: Learning(s) from a synthetic ImageNet clone

Participants: Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, Yannis Kalantidis.

Recent large-scale image generation models such as Stable Diffusion have exhibited an impressive ability to generate fairly realistic images starting from a very simple text prompt. Could such models render real images obsolete for training image prediction models? In this work [17], we answer part of this provocative question by questioning the need for real images when training models for ImageNet classification. More precisely, provided only with the class names that have been used to build the dataset,



(a) Training a model on synthetic images.



(b) Testing the frozen model on real images.

Figure 7: Overview of our experimental protocol. During training, the model has access to synthetic images generated by the Stable Diffusion model, provided with a set of prompts per class. During evaluation, real images are classified by the frozen model.

we explore the ability of Stable Diffusion to generate synthetic clones of ImageNet and measure how useful they are for training classification models from scratch. An overview of our experimental protocol is shown in Figure 7. We show that with minimal and class-agnostic prompt engineering those ImageNet clones we denote as ImageNet-SD are able to close a large part of the gap between models produced by synthetic images and models trained with real images for the several standard classification benchmarks that we consider in this study. More importantly, we show that models trained on synthetic images exhibit strong generalization properties and perform on par with models trained on real data.

Think Before You Act: Unified Policy for Interleaving Language Reasoning with Actions

Participants: Lina Mezghani, Piotr Bojanowski, Karteek Alahari, Sainbayar Sukhbaatar.

The success of transformer models trained with a language modeling objective brings a promising opportunity to the reinforcement learning framework. Decision Transformer is a step towards this direction, showing how to train transformers with a similar next-step prediction objective on offline data. Another important development in this area is the recent emergence of large-scale datasets collected from the internet, such as the ones composed of tutorial videos with captions where people talk about what they are doing. To take advantage of this language component, we propose a novel method [16] for unifying language reasoning with actions in a single policy. Specifically, we augment a transformer policy with word outputs, so it can generate textual captions interleaved with actions (see Figure 8). When tested on the most challenging task in BabyAI, with captions describing next subgoals, our reasoning policy consistently outperforms the caption-free baseline.

Unlocking Pre-trained Image Backbones for Semantic Image Synthesis

Participants: Tariq Berrada, Jakob Verbeek, Camille Couprie, Karteek Alahari.

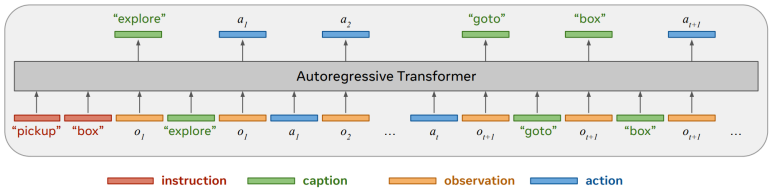


Figure 8: Given an instruction, our transformer policy can generate language based reasoning tokens interleaved with sequence of actions in the environment.

Figure 9: Images generated with models trained on COCO-Stuff, comparing our approach to state-of-the-art methods OASIS, SDM, and PITI, along with inference times to generate a single image. Our approach combines high-quality samples with low-latency sampling.

Semantic image synthesis, i.e., generating images from user-provided semantic label maps, is an important conditional image generation task as it allows to control both the content as well as the spatial layout of generated images. Although diffusion models have pushed the state of the art in generative image modeling, the iterative nature of their inference process makes them computationally demanding. Other approaches such as GANs are more efficient as they only need a single feed-forward pass for generation, but the image quality tends to suffer on large and diverse datasets. In this work [23], we propose a new class of GAN discriminators for semantic image synthesis that generates highly realistic images by exploiting feature backbone networks pre-trained for tasks such as image classification. We also introduce a new generator architecture with better context modeling and using cross-attention to inject noise into latent variables, leading to more diverse generated images. Our model, which we dub DP-SIMS, achieves state-of-the-art results (see examples in Figure 9) in terms of image quality and consistency with the input label maps on ADE-20K, COCO-Stuff, and Cityscapes, surpassing recent diffusion models while requiring two orders of magnitude less compute for inference.

Guided Distillation for Semi-Supervised Instance Segmentation

Participants: Tariq Berrada, Camille Couprie, Karteek Alahari, Jakob Verbeek.

Although instance segmentation methods have improved considerably, the dominant paradigm is to rely on fully-annotated training images, which are tedious to obtain. To alleviate this reliance, and boost results, semi-supervised approaches leverage unlabeled data as an additional training signal that limits overfitting to the labeled samples. In this context, we present novel design choices [7] to significantly improve teacher-student distillation models. In particular, we (i) improve the distillation approach by introducing a novel "guided burn-in" stage, and (ii) evaluate different instance segmentation architectures, as well as backbone networks and pre-training strategies. Contrary to previous work which uses only supervised data for the burn-in period of the student model, we also use guidance of the teacher model to exploit unlabeled data in the burn-in period. Our improved distillation approach leads to substantial

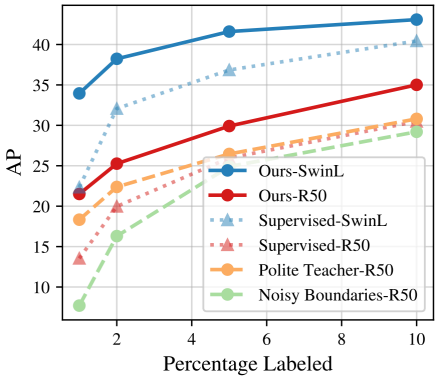


Figure 10: Compared with the state-of-the-art Polite Teacher method, we achieve +15.7 mask-AP when using 1% of labels, for an AP of 34.0, which is more than what Polite Teacher achieved using 10 times more labels (30.8) on the COCO dataset.

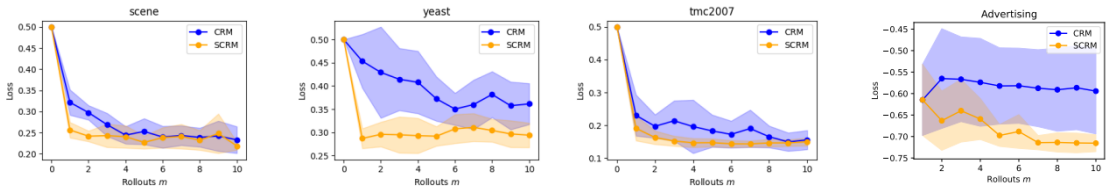


Figure 11: Test loss as a function of sample size on Scene, Yeast, TMC2007, Advertising, (from left to right). SCRM (in orange) converges faster and with less variance than CRM (in blue).

improvements over previous state-of-the-art results. For example, on the Cityscapes dataset we improve mask-AP from 23.7 to 33.9 when using labels for 10% of images, and on the COCO dataset we improve mask-AP from 18.3 to 34.1 when using labels for only 1% of the training data (see Figure 10).

7.2 Statistical Machine Learning

Sequential Counterfactual Risk Minimization

Participants: Houssam Zenati, Eustache Diemert, Matthieu Martin, Julien Mairal, Pierre Gaillard.

Counterfactual Risk Minimization (CRM) is a framework for dealing with the logged bandit feedback problem, where the goal is to improve a logging policy using offline data. In [19], we explore the case where it is possible to deploy learned policies multiple times and acquire new data. We extend the CRM principle and its theory to this scenario, which we call "Sequential Counterfactual Risk Minimization (SCRM)". We introduce a novel counterfactual estimator and identify conditions that can improve the performance of CRM in terms of excess risk and regret rates, by using an analysis similar to restart strategies in accelerated optimization methods. We also provide an empirical evaluation of our method in both discrete and continuous action settings, and demonstrate the benefits of multiple deployments of CRM, as illustrated in Figure 11

Self-Attention in Colors: Another Take on Encoding Graph Structure in Transformers

Participants: Romain Menegaux, Emmanuel Jehanno, Margot Selosse, Julien Mairal.

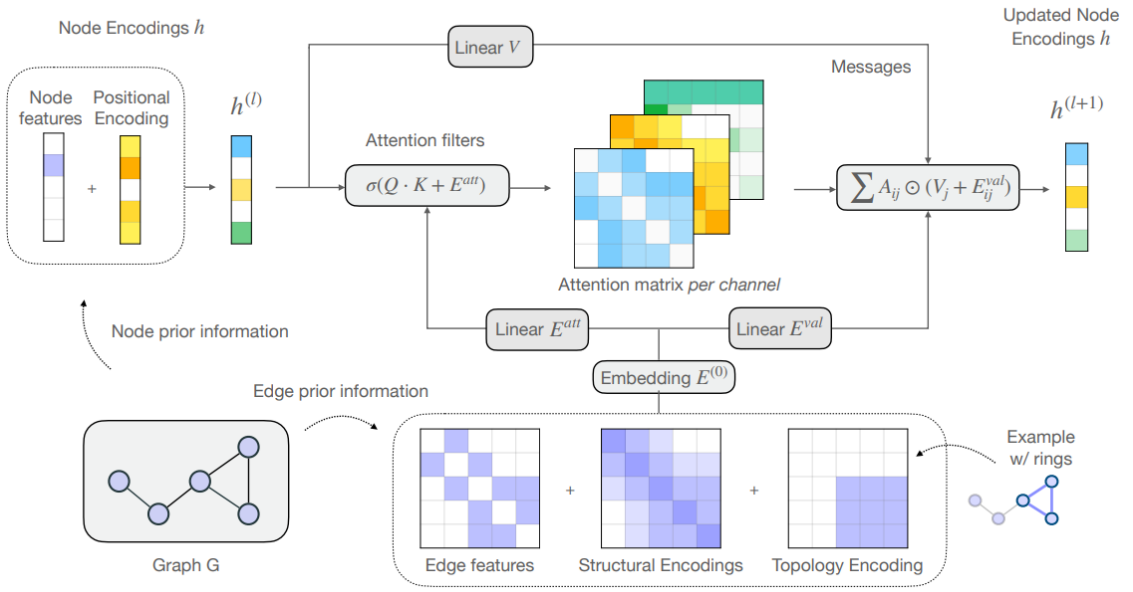


Figure 12: Chromatic Graph Transformer. (a) Graph features are preprocessed into node features h (top-left) and two edge feature matrices (bottom). (b) These are input to the CSA layers, which iteratively update the node representations h. (c) These representations are fed to a classification or regression head (not shown here).

In [2], we introduce a novel self-attention mechanism, which we call CSA (Chromatic Self-Attention), which extends the notion of attention scores to attention filters, independently modulating the feature channels. We showcase CSA in a fully-attentional graph Transformer CGT (Chromatic Graph Transformer) which integrates both graph structural information and edge features, completely bypassing the need for local message-passing components. Our method flexibly encodes graph structure through node-node interactions, by enriching the original edge features with a relative positional encoding scheme. We propose a new scheme based on random walks that encodes both structural and positional information, and show how to incorporate higher-order topological information, such as rings in molecular graphs. Our approach achieves state-of-the-art results on the ZINC benchmark dataset, while providing a flexible framework for encoding graph structure and incorporating higher-order topology. This approach is illustrated in Figure 12

GloptiNets: Scalable Non-Convex Optimization with Certificates

Participants: Gaspard Beugnot, Julien Mairal, Alessandro Rudi.

In [8], we present a novel approach to non-convex optimization with certificates, which handles smooth functions on the hypercube or on the torus. Unlike traditional methods that rely on algebraic properties, our algorithm exploits the regularity of the target function intrinsic in the decay of its Fourier spectrum. By defining a tractable family of models, we allow at the same time to obtain precise certificates and to leverage the advanced and powerful computational techniques developed to optimize neural networks. In this way the scalability of our approach is naturally enhanced by parallel computing with GPUs. Our approach, when applied to the case of polynomials of moderate dimensions but with thousands of coefficients, outperforms the state-of-the-art optimization methods with certificates, as the ones based on Lasserre's hierarchy, addressing problems intractable for the competitors

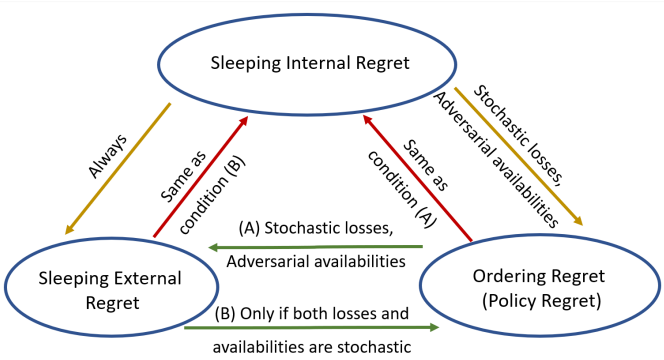


Figure 13: The connections between our proposed notion of Sleeping Internal Regret and different existing notions of regret.

One Arrow, Two Kills: An Unified Framework for Achieving Optimal Regret Guarantees in Sleeping Bandits

Participants: Pierre Gaillard, Aadirupa Saha, Soham Dan.

In [11], we address the problem of 'Internal Regret' in Sleeping Bandits within a fully adversarial setup. The key contribution lies in unifying different regret notions in sleeping bandits and understanding their interplay illustrated in Figure 13. The paper extends its results to Dueling Bandits, proposing a reduction to multiarmed bandits approach to design a low regret algorithm for sleeping dueling bandits with stochastic preferences and adversarial availabilities. The efficacy of the algorithms is supported by empirical evaluations.

Efficient Model-Based Concave Utility Reinforcement Learning through Greedy Mirror Descent

Participants: Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard, Nadia Oudjane.

The paper [32] introduces MD-CURL, an algorithm for solving the Concave Utility Reinforcement Learning (CURL) problem in finite horizon Markov decision processes. CURL, a general paradigm encompassing reinforcement learning and imitation learning, challenges classical Bellman equations, necessitating innovative algorithms. MD-CURL, inspired by mirror descent, employs a non-standard regularization for convergence guarantees and a closed-form solution, eliminating the need for computationally expensive projection steps. We extend CURL to an online learning scenario with Greedy MD-CURL, adapting it to an episode-based setting with partially unknown dynamics. Both MD-CURL and Greedy MD-CURL offer low computational complexity, ensuring sub-linear or logarithmic regret based on the available information about the underlying dynamics.

On the Shift Invariance of Max Pooling Feature Maps in Convolutional Neural Networks

Participants: Hubert Leterme, Kévin Polisano, Valérie Perrier, Karteek Alahari.

In this paper [30], we aim to improve the mathematical interpretability of convolutional neural networks for image classification. When trained on natural image datasets, such networks tend to learn parameters in the first layer that closely resemble oriented Gabor filters. By leveraging the properties of discrete Gabor-like convolutions, we prove that, under specific conditions, feature maps computed by the

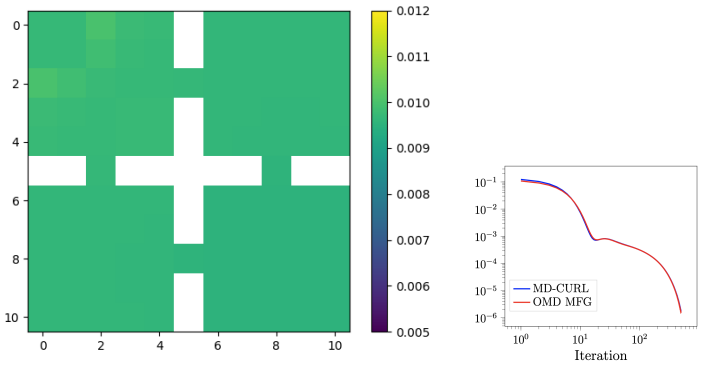


Figure 14: Convergence of MD-CURL on an entropy maximization problem.

subsequent max pooling operator tend to approximate the modulus of complex Gabor-like coefficients, and as such, are stable with respect to certain input shifts. We then compute a probabilistic measure of shift invariance for these layers. More precisely, we show that some filters, depending on their frequency and orientation, are more likely than others to produce stable image representations. We experimentally validate our theory by considering a deterministic feature extractor based on the dual-tree wavelet packet transform, a particular case of discrete Gabor-like decomposition. We demonstrate a strong correlation between shift invariance on the one hand and similarity with complex modulus on the other hand, as illustrated in Figure 15.

From CNNs to Shift-Invariant Twin Wavelet Models

Participants: Hubert Leterme, Kévin Polisano, Valérie Perrier, Karteek Alahari.

In this paper [29], we propose a novel antialiasing method to increase shift invariance in convolutional neural networks (CNNs). More precisely, we replace the conventional combination "real-valued convolutions + max pooling" ($\mathbb{R}$Max) by "complex-valued convolutions + modulus" ($\mathbb{C}$Mod), which produce stable feature representations for band-pass filters with well-defined orientations. In a recent work [30], we proved that, for such filters, the two operators yield similar outputs. Therefore, $\mathbb{C}$Mod can be viewed as a stable alternative to $\mathbb{R}$Max. To separate band-pass filters from other freely-trained kernels, in this paper, we designed a "twin" architecture based on the dual-tree complex wavelet packet transform (DT-$\mathbb{C}$WPT), which generates similar outputs as standard CNNs with fewer trainable parameters. In addition to improving stability to small shifts, our experiments on AlexNet and ResNet showed increased prediction accuracy on natural image datasets such as ImageNet and CIFAR10. Furthermore, our approach outperformed recent antialiasing methods based on low-pass filtering by preserving high-frequency information, while reducing memory usage. Figure 16 compares the accuracy and shift invariance of the various methods.

7.3 Pluri-disciplinary Research

Learning Reward Functions for Robotic Manipulation by Observing Humans

Participants: Minttu Alakuijala, Julien Mairal, Jean Ponce, Cordelia Schmid.

Observing a human demonstrator manipulate objects provides a rich, scalable and inexpensive source of data for learning robotic policies. However, transferring skills from human videos to a robotic manipulator poses several challenges, not least a difference in action and observation spaces. In [5], we use unlabeled videos of humans solving a wide range of manipulation tasks to learn a task-agnostic reward function for robotic manipulation policies. Thanks to the diversity of this training data, the

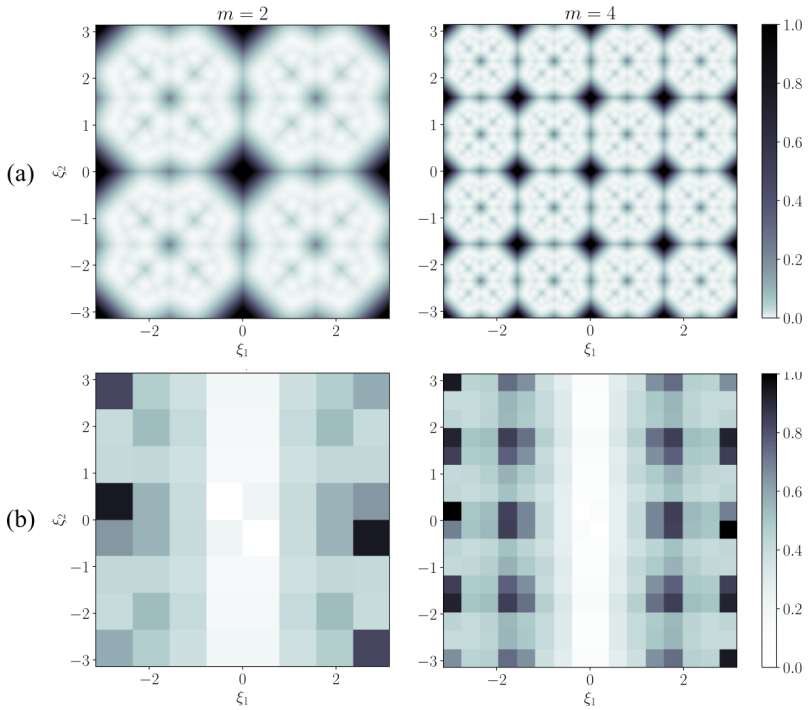


Figure 15: Top (a): expected discrepancy between complex modulus feature maps and max pooling feature maps (theoretical result). Input images are assumed to be filtered using oriented, band-pass kernels with characteristic frequencies $\boldsymbol{\xi} = (\xi_1, \xi_2) \in [-\pi, \pi]^2$, and subsampled by a factor 2 (left) or 4 (right). Bottom (b): stability of max pooling outputs with respect to small shifts along the x-axis, averaged over 50K images from the ImageNet dataset (experimental result). For the sake of visual comparison with (a), several band-pass convolution filters have been tested, with characteristic frequencies $\boldsymbol{\xi}$ varying within $[-\pi, \pi]^2$, and the same subsampling factors as above. We observe regular patterns of dark spots; more precisely, shift instabilities seem to occur when the filter's frequency is located in a dark region of (a).

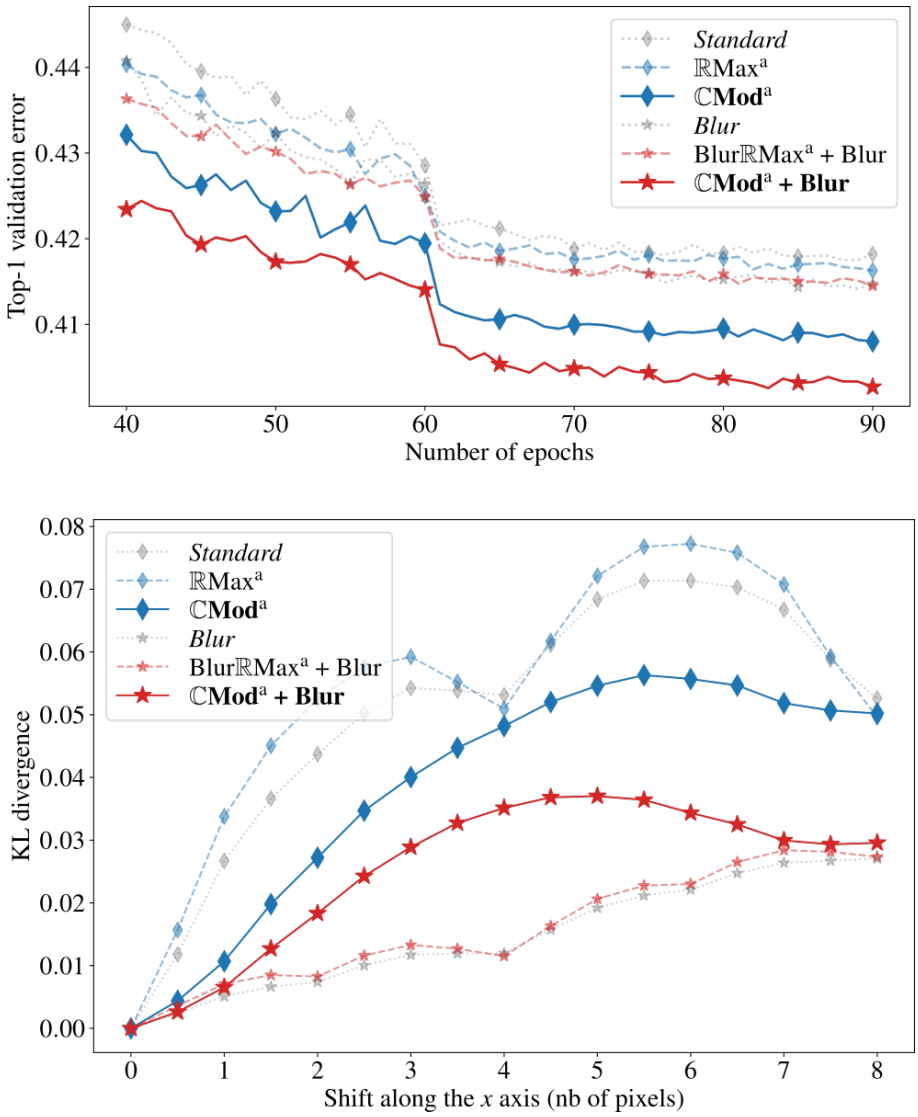


Figure 16: AlexNet and antialiased variants. Top: top-1 validation error along training with ImageNet 2012. Bottom: mean KL divergence between the outputs of a reference image versus shifted images, measuring stability with respect to small input shifts. The solid curves represent the twin models modified with our antialiasing method. It outperforms the standard, non-antialiased approach (blue dashed curve) as well as the antialiasing approach based on low-pass filtering (red dashed curve) in terms of classification accuracy.

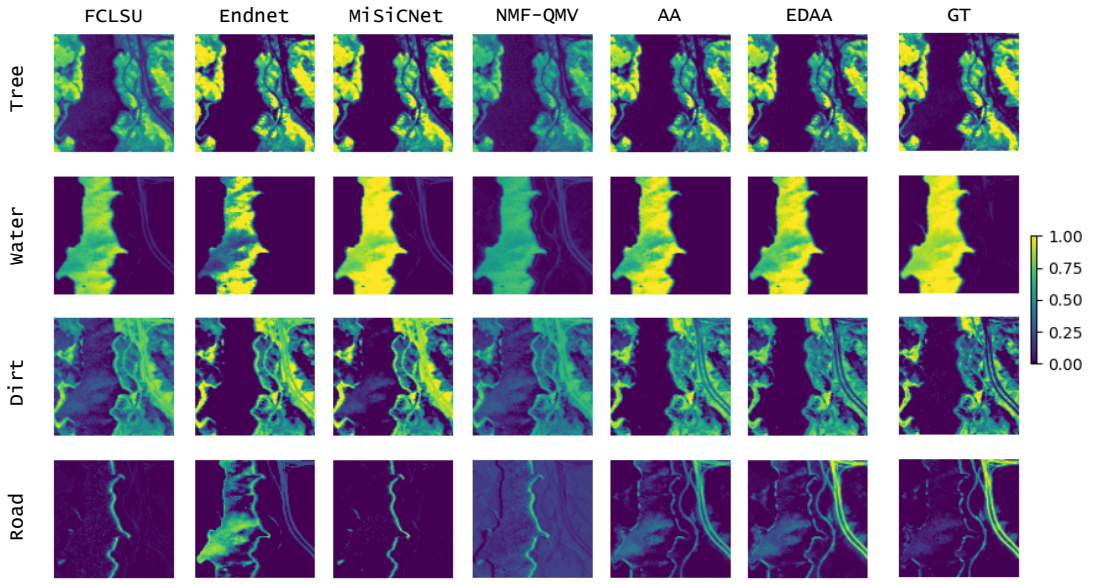


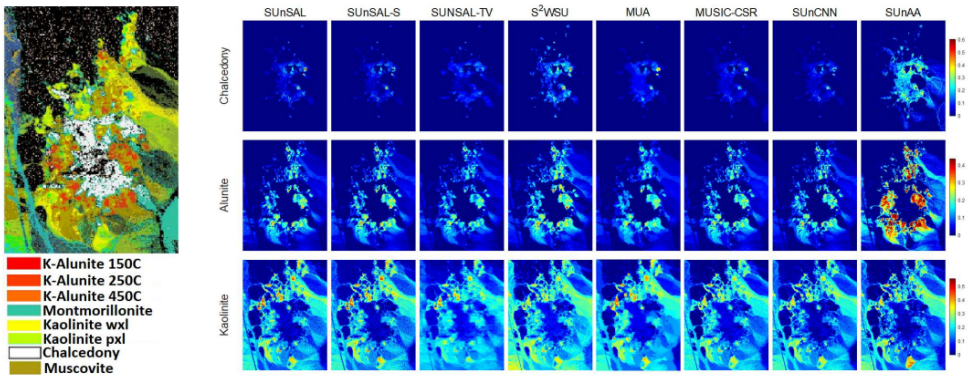
Figure 17: Estimated abundances on the Jasper Ridge dataset. The rows represent the different endmembers. The columns represent competing methods. The ground truth abundance maps are displayed on the rightmost column. Our method, EDAA, captures best the different endmembers, most notably the Road.

learned reward function sufficiently generalizes to image observations from a previously unseen robot embodiment and environment to provide a meaningful prior for directed exploration in reinforcement learning. We propose two methods for scoring states relative to a goal image: through direct temporal regression, and through distances in an embedding space obtained with time-contrastive learning. By conditioning the function on a goal image, we are able to reuse one model across a variety of tasks. Unlike prior work on leveraging human videos to teach robots, our method, Human Offline Learned Distances (HOLD) requires neither a priori data from the robot environment, nor a set of task-specific human demonstrations, nor a predefined notion of correspondence across morphologies, yet it is able to accelerate training of several manipulation tasks on a simulated robot arm compared to using only a sparse reward obtained from task completion.

Entropic Descent Archetypal Analysis for Blind Hyperspectral Unmixing

Participants: Alexandre Zouaoui, Gedeon Muhawenayo, Behnood Rasti, Jocelyn Chanussot, Julien Mairal.

In [4], we introduce a new algorithm based on archetypal analysis for blind hyperspectral unmixing, assuming linear mixing of endmembers. Archetypal analysis is a natural formulation for this task. This method does not require the presence of pure pixels (i.e., pixels containing a single material) but instead represents endmembers as convex combinations of a few pixels present in the original hyperspectral image. Our approach leverages an entropic gradient descent strategy, which (i) provides better solutions for hyperspectral unmixing than traditional archetypal analysis algorithms, and (ii) leads to efficient GPU implementations. Since running a single instance of our algorithm is fast, we also propose an ensembling mechanism along with an appropriate model selection procedure that make our method robust to hyper-parameter choices while keeping the computational complexity reasonable. By using six standard real datasets, we show that our approach outperforms state-of-the-art matrix factorization and recent deep learning methods.



(a) Geological Map (b) Estimated abundance maps

Figure 18: Abundance maps of three dominant minerals estimated using different sparse unmixing techniques applied to the Cuprite dataset.

SUnAA: Sparse Unmixing using Archetypal Analysis

Participants: Behnood Rasti, Alexandre Zouaoui, Julien Mairal, Jocelyn Chanussot.

This paper [3] introduces a new sparse unmixing technique using archetypal analysis (SUnAA). First, we design a new model based on archetypal analysis (AA). We assume that the endmembers of interest are a convex combination of endmembers provided by a spectral library and that the number of endmembers of interest is known. Then, we propose a minimization problem. Unlike most conventional sparse unmixing methods, here the minimization problem is nonconvex. We minimize the optimization objective iteratively using an active set algorithm. Our method is robust to the initialization and only requires the number of endmembers of interest. SUnAA is evaluated using two simulated datasets for which results confirm its better performance over other conventional and advanced techniques in terms of signal-to-reconstruction error (SRE). SUnAA is also applied to Cuprite dataset and the results are compared visually with the available geological map provided for this dataset. The qualitative assessment demonstrates the successful estimation of the minerals abundances and significantly improves the detection of dominant minerals compared to the conventional regression-based sparse unmixing methods. The Python implementation of SUnAA can be found at: github.com/BehnoodRasti/SUnAA. Examples of results are given in Figure 18.

Image Processing and Machine Learning for Hyperspectral Unmixing: An Overview and the HySUPP Python Package

Participants: Behnood Rasti, Alexandre Zouaoui, Julien Mairal, Jocelyn Chanussot.

Spectral pixels are often a mixture of the pure spectra of the materials, called endmembers, due to the low spatial resolution of hyperspectral sensors, double scattering, and intimate mixtures of materials in the scenes. Unmixing estimates the fractional abundances of the endmembers within the pixel. Depending on the prior knowledge of endmembers, linear unmixing can be divided into three main groups: supervised, semi-supervised, and unsupervised (blind) linear unmixing. Advances in Image processing and machine learning substantially affected unmixing. This paper [33] provides an overview of advanced and conventional unmixing approaches. Additionally, we draw a critical comparison between advanced and conventional techniques from the three categories. We compare the performance of the unmixing techniques on three simulated and two real datasets. The experimental results reveal the

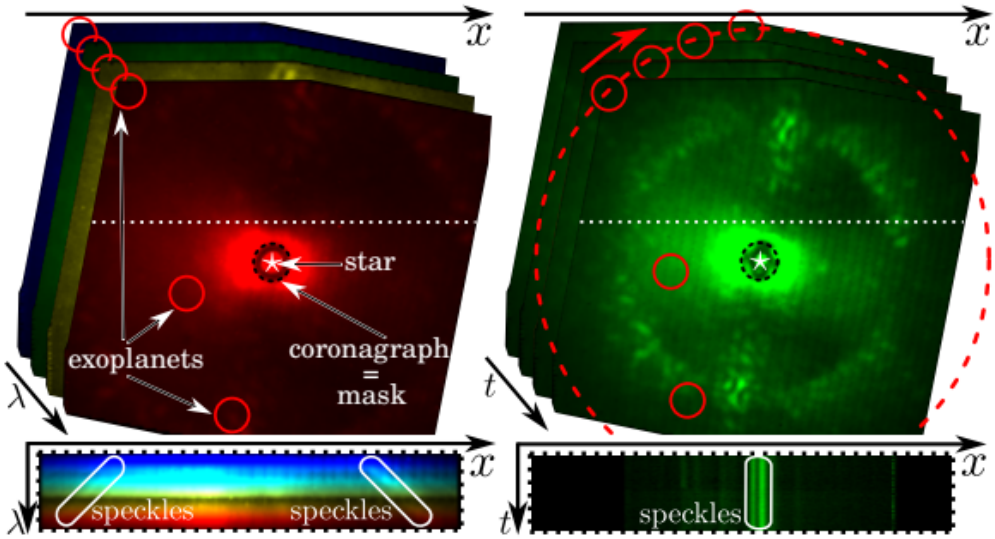


Figure 19: Illustration of a dataset form the SPHERE-IFS instrument. Left: images at different wavelengths. Right: images at different times. Red circles represent the locations of three known exoplanets whose signatures are too faint to be detected without additional processing. Bottom: spatio-spectral and spatio-temporal slice cuts along the white dashed line.

advantages of different unmixing categories for different unmixing scenarios. Moreover, we provide an open-source Python-based package available at github.com/BehnoodRasti/HySUPP to reproduce the results.

Combining multi-spectral data with statistical and deep-learning models for improved exoplanet detection in direct imaging at high contrast

Participants: Olivier Flasseur, Theo Bodrito, Julien Mairal, Jean Ponce, Maud Langlois, Anne-Marie Lagrange.

Exoplanet detection by direct imaging is a difficult task: the faint signals from the objects of interest are buried under a spatially structured nuisance component induced by the host star. The exoplanet signals can only be identified when combining several observations with dedicated detection algorithms. In contrast to most of existing methods, we propose in [10] to learn a model of the spatial, temporal and spectral characteristics of the nuisance, directly from the observations. In a pre-processing step, a statistical model of their correlations is built locally, and the data are centered and whitened to improve both their stationarity and signal-to-noise ratio (SNR). A convolutional neural network (CNN) is then trained in a supervised fashion to detect the residual signature of synthetic sources in the preprocessed images. Our method leads to a better trade-off between precision and recall than standard approaches in the field. It also outperforms a state-of-the-art algorithm based solely on a statistical framework. Besides, the exploitation of the spectral diversity improves the performance compared to a similar model built solely from spatio-temporal data. An example of dataset is illustrated in Figure 19.

deep PACO: Combining statistical models with deep learning for exoplanet detection and characterization in direct imaging at high contrast

Participants: Olivier Flasseur, Theo Bodrito, Julien Mairal, Jean Ponce, Maud Langlois, Anne-Marie Lagrange.

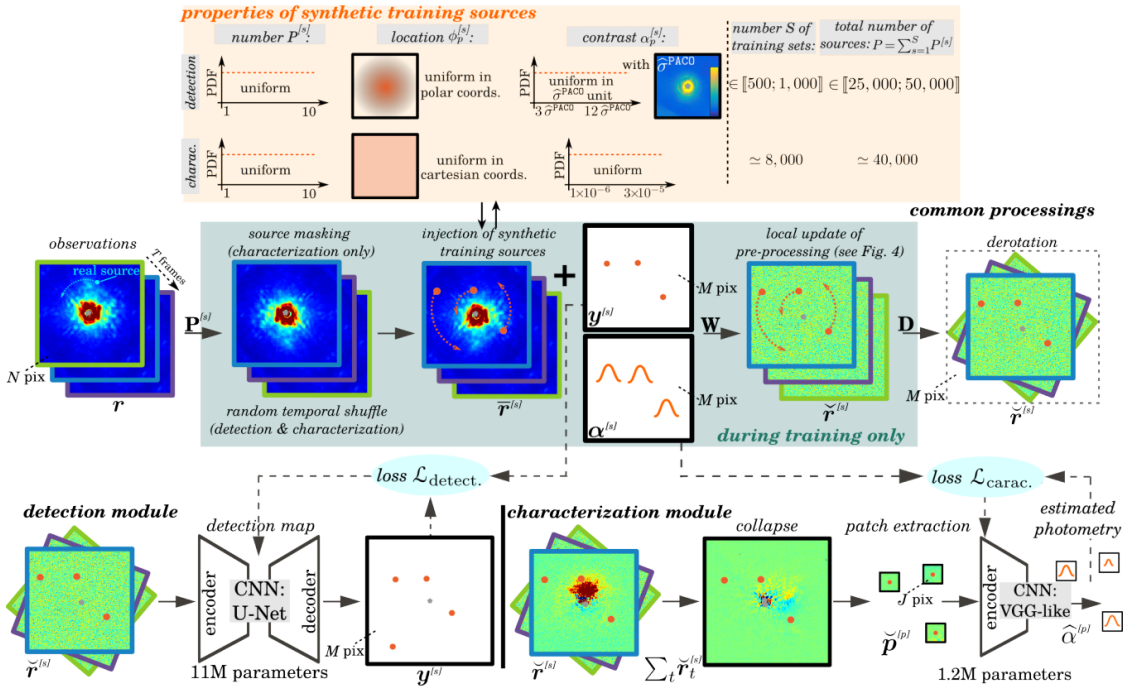


Figure 20: Schematic representation of the main operations performed during the detection and characterization steps of the proposed algorithm by supervised deep learning. The first line displays a view of the main parameters defining synthetic sources injected into the pre-processed observations (see Fig. 1) at training time. The second line shows common operations performed for both the detection and the characterization steps. The left (respectively, the right) part of the third line is for operations applied solely during the detection (respectively, the characterization) step. Throughout this paper, synthetic training sources injected to build our models are highlighted in orange while (possibly unknown) real and synthetic sources that we aim to detect and to characterize at inference time are displayed in light blue in the schematic representations. Dataset: HIP 72192 (2015-05-10).

Direct imaging is an active research topic in astronomy for the detection and the characterization of young sub-stellar objects. The very high contrast between the host star and its companions makes the observations particularly challenging. In this context, post-processing methods combining several images recorded with the pupil tracking mode of telescope are needed. In previous works, we have presented a data-driven algorithm, PACO, capturing locally the spatial correlations of the data with a multi-variate Gaussian model. PACO delivers better detection sensitivity and confidence than the standard post-processing methods of the field. However, there is room for improvement due to the approximate fidelity of the PACO statistical model to the time evolving observations. In [25] we propose to combine the statistical model of PACO with supervised deep learning. The data are first pre-processed with the PACO framework to improve the stationarity and the contrast. A convolutional neural network (CNN) is then trained in a supervised fashion to detect the residual signature of synthetic sources. Finally, the trained network delivers a detection map. The photometry of detected sources is estimated by a second CNN. We apply the proposed approach to several datasets from the VLT/SPHERE instrument. Our results show that its detection stage performs significantly better than baseline methods (cADI, PCA), and leads to a contrast improvement up to half a magnitude compared to PACO. The characterization stage of the proposed method performs on average on par with or better than the comparative algorithms (PCA, PACO) for angular separation above 0.5. This approach is illustrated in Figure 20.

Fine Dense Alignment of Image Bursts through Camera Pose and Depth Estimation

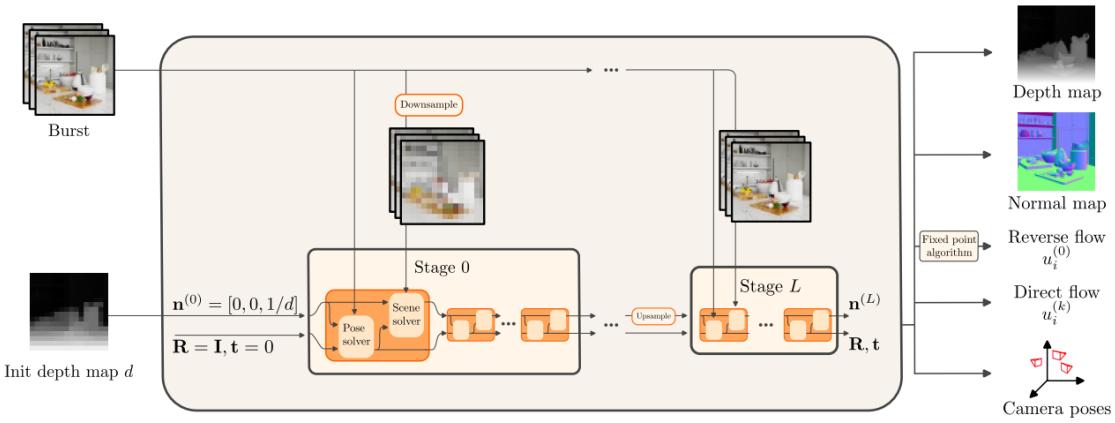


Figure 21: The global pipeline of our optimization-based method. It inputs a burst of images and an initialization depth map and outputs the direct and reverse flow between each image and the first one. Our method estimates the optical flow using the camera's pose and 3D scene structure as optimization variables of the photometric reprojection errors in a reference frame then poses and depth maps can also be retrieved.

Participants: Bruno Lecouat, Yann Dubois de Mont-Marin, Theo Bodrito, Julien Mairal, Jean Ponce.

This paper [28] introduces a novel approach to the fine alignment of images in a burst captured by a handheld camera. In contrast to traditional techniques that estimate twodimensional transformations between frame pairs or rely on discrete correspondences, the proposed algorithm establishes dense correspondences by optimizing both the camera motion and surface depth and orientation at every pixel. This approach improves alignment, particularly in scenarios with parallax challenges. Extensive experiments with synthetic bursts featuring small and even tiny baselines demonstrate that it outperforms the best optical flow methods available today in this setting, without requiring any training. Beyond enhanced alignment, our method opens avenues for tasks beyond simple image restoration, such as depth estimation and 3D reconstruction, as supported by promising preliminary results. This positions our approach as a versatile tool for various burst image processing applications. Our pipeline is illustrated in Figure 21.

7.4 Optimization

Rethinking Gauss-Newton for learning over-parameterized models

Participants: Michael Arbel, Romain Menegaux, Pierre Wolinsky.

Compared to gradient descent, Gauss-Newton's method (GN) and variants are known to converge faster to local optima at the expense of a higher computational cost per iteration. Still, GN is not widely used for optimizing deep neural networks despite a constant effort to reduce their higher computational cost. In [6], we propose to take a step back and re-think the properties of GN in light of recent advances in the dynamics of gradient flows of over-parameterized models and the implicit bias they induce. We first prove a fast global convergence result for the continuous-time limit of the generalized GN in the over-parameterized regime. We then show empirically that GN exhibits both a kernel regime where it generalizes as well as gradient flows, and a feature learning regime where GN induces an implicit bias for selecting global solutions that systematically under-performs those found by a gradient flow. Importantly, we observed this phenomenon even with enough computational budget to perform exact GN steps over

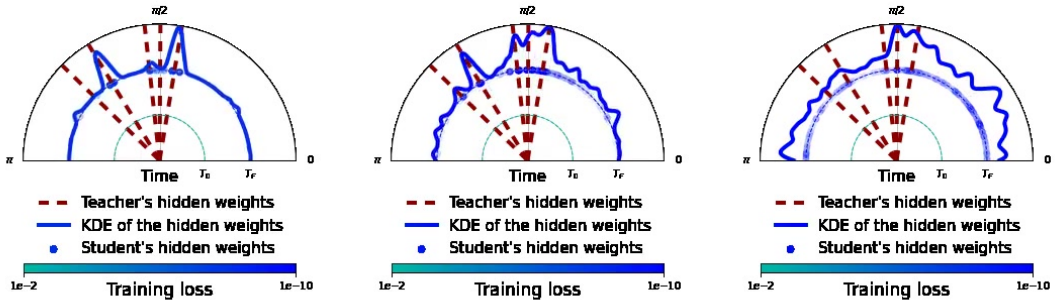


Figure 22: Distribution of the hidden weights directions of a student netwrok (1-hidden layer with 1000 neurons) optimized to match the output of a teacher network (1-hidden layer with 5 neurons). From left to right, result for incresing step sizes. Small step sizes result in hidden weights that concentrate on the teacher's one (feature learning regime). Larger step-sizes results in weights that are span all direction just like at initialization (kernel learning regime).

the total training objective. This study suggests the need to go beyond improving the computational cost of GN for over-parametrized models towards designing new methods that can trade off optimization speed and the quality of their implicit bias.

Beyond spectral gap (extended): the role of the topology in decentralized learning

Participants: Thijs Vogels, Hadrien Hendrikx, Martin Jaggi.

This paper [34] focuses on the problem of decentralized stochastic gradient descent, a method frequently used to train machine learning models when the data is distributed across many computing nodes. While previous analyses predicted an increase in training time with networks of growing size (even if nodes have the same data), our theoretical resuts are actually predictive of what happens in practice: adding computing nodes allows to increase the (initial) learning rate of stochastic gradient descent, and so obtain faster convergence. We link the convergence speed with the topology of the communication graph through a new notion, the effective number of neighbors, which depends on the learning rate. When computing nodes have different data, we finely characterize the solution each converges to, and in particular with respect to the dependence of the other nodes' data. Figure 23 describes the previous gap between theory and practice, that we close in this paper.

Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees

Participants: Anastasia Koloskova, Hadrien Hendrikx, Sebastian Stich.

Gradient clipping is a popular modification to standard (stochastic) gradient descent, at every iteration limiting the gradient norm to a certain value $c > 0$. It is widely used for example for stabilizing the training of deep learning models, or for enforcing differential privacy. Despite popularity and simplicity of the clipping mechanism, its convergence guarantees often require specific values of c and strong noise assumptions.

In this paper [14], we give convergence guarantees that show precise dependence on arbitrary clipping thresholds c and show that our guarantees are tight with both deterministic and stochastic gradients. In

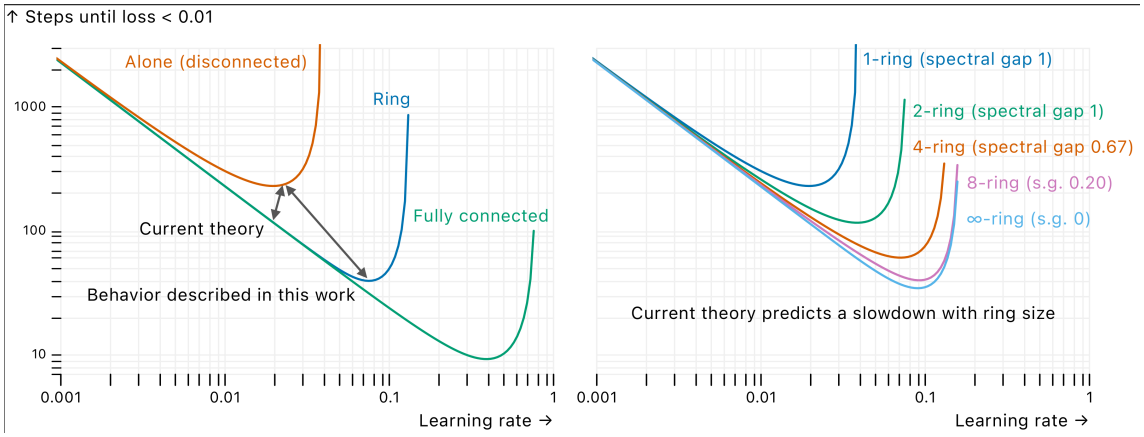


Figure 23: Number of steps required by Decentralized Stochastic Gradient Descent to reach a certain error level depending on the learning rate. We see that, unlike predicted by previous theory, having several nodes that communicate allows to use larger learning rates, which in turns decreases the number of steps required to reach the target error.

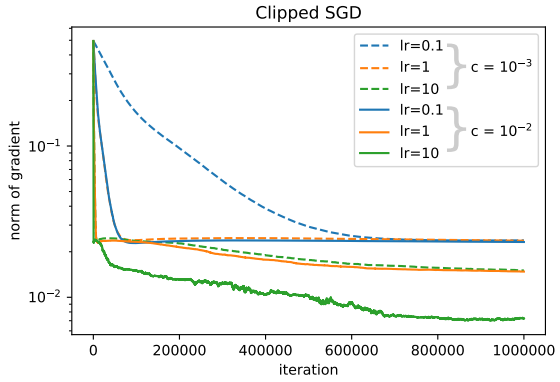


Figure 24: Convergence curves for clipped SGD on a toy function. We observe that small clipping threshold introduce bias in the result, leading to high final gradient norm. Our theory precisely characterizes these thresholds.

particular, we show that (i) for deterministic gradient descent, the clipping threshold only affects the higher-order terms of convergence, (ii) in the stochastic setting convergence to the true optimum cannot be guaranteed under the standard noise assumption, even under arbitrary small step-sizes. We give matching upper and lower bounds for convergence of the gradient norm when running clipped SGD, and illustrate these results with experiments. For instance, Figure 24 shows how different clipping threshold bias the results of clipped SGD.

The Relative Gaussian Mechanism and its Application to Private Gradient Descent

Participants: Hadrien Hendrikx, Paul Mangold, Aurélien Bellet.

The Gaussian Mechanism (GM), which consists in adding Gaussian noise to a vector-valued query before releasing it, is a standard privacy protection mechanism. In particular, given that the query respects some *L2 sensitivity* property (the L2 distance between outputs on any two neighboring inputs is bounded), GM guarantees Rényi Differential Privacy (RDP). Unfortunately, precisely bounding the L2 sensitivity can be hard, thus leading to loose privacy bounds. In this paper [12], we consider a *Relative*

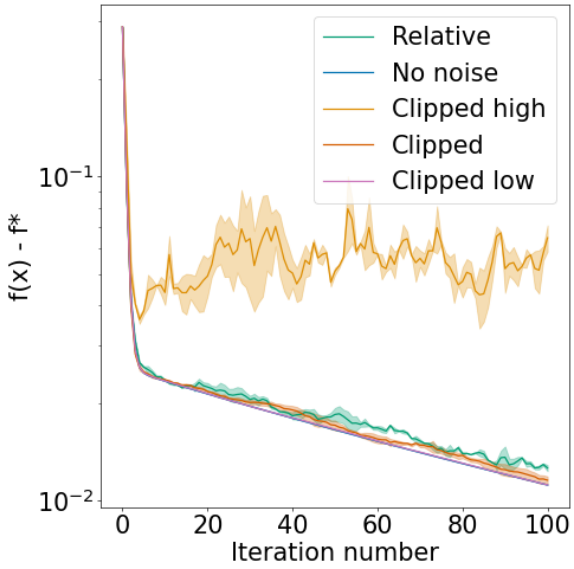


Figure 25: This plot shows convergence curves for differentially private gradient descent, with either gradient clipping or the relative Gaussian mechanism. We see that the relative Gaussian Mechanism is competitive with gradient clipping, while being adaptive to the gradient magnitude and so requiring less tuning.

L2 sensitivity assumption, in which the bound on the distance between two query outputs may also depend on their norm. Leveraging this assumption, we introduce the *Relative Gaussian Mechanism* (RGM), in which the variance of the noise depends on the norm of the output. We prove tight bounds on the RDP parameters under relative L2 sensitivity, and characterize the privacy loss incurred by using output-dependent noise. In particular, we show that RGM naturally adapts to a latent variable that would control the norm of the output. Finally, we instantiate our framework to show tight guarantees for Private Gradient Descent, a problem that naturally fits our relative L2 sensitivity assumption. Figure 25 shows that enforcing privacy through the Relative Gaussian Mechanism is competitive with the standard Gaussian Mechanism applied to clipped gradients.

Adaptive approximation of monotone functions

Participants: Pierre Gaillard, Sébastien Gerchinovitz, Etienne de Montbrun.

In [26], we address the problem of approximating a non-decreasing function through sequential queries. Unlike previous minimax results, our approach provides tight upper and lower bounds specific to each function f. We introduce GreedyBox, a generalized algorithm building on Novak's (1992) proposal for numerical integration, demonstrating its optimality up to logarithmic factors for any function f. Specifically examining monotone functions that are also piecewise-C^2, we explore the error reduction of GreedyBox beyond its predicted performance, uncovering a nuanced relationship between C^1-singularities, monotonicity, and adaptivity in our matching upper and lower bounds.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

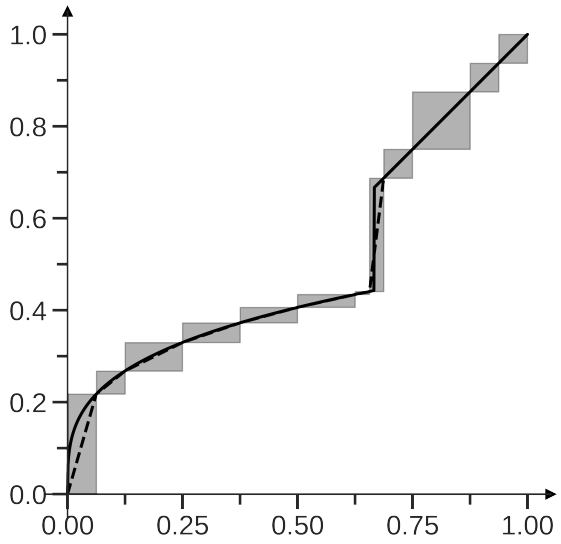


Figure 26: Greedybox approximation of a piecewise-C2 function after 12 iterations.

Participants: Julien Mairal, Karteek Alahari, Pierre Gaillard, Jocelyn Chanussot.

In 2023, we had:

- one CIFRE PhD student with Criteo, Houssam Zenati (co-advised by J. Mairal and P. Gaillard)
- three CIFRE PhD students with Facebook: Timothée Darcet (co-advised by J. Mairal), Lina Mezghani (co-advised by K. Alahari), and Tariq Berrada Ifriqi (co-advised by K. Alahari).
- one CIFRE PhD student with Valeo AI: Florent Bartoccioni (co-advised by K. Alahari)
- two CIFRE PhD student with Naver Labs Europe: Mert Bulent Sariyildiz (co-advised by K. Alahari) and Juliette Marrie (co-advised by J. Mairal and M. Arbel).
- one CIFRE PhD student with Preligens: Jules Bourcier (co-advised by K. Alahari and J. Chanussot)
- one CIFRE PhD student with Nokia Bell Labs: Camila Fernández (co-advised by P. Gaillard)

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

4TUNE

Participants: Pierre Gaillard.

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

Partner Institution: CWI, Pays-Bas

Coordinator: Peter Grünwald (pdg@cwi.nl)

Date/Duration: 2020 ->

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 International research visitors

9.2.1 Visits of international scientists

Other international visits to the team Thomas de Min (Masters student then and now PhD student at University of Trento) visited us from February until May 2023. Nassim Ait Ali Braham, PhD student from DLR, visited us from Apr 2023 until Oct 2023.

9.3 European initiatives

9.3.1 ERC Project APHELEIA

Participants: Julien Mairal, Emmanuel Jehanno, Romain Seailles.

Despite the undeniable success of machine learning in addressing a wide variety of technological and scientific challenges, the current trend of training predictive models with an evergrowing number of parameters from an evergrowing amount of data is not sustainable. These huge models, often engineered by large corporations benefiting from huge computational resources, typically require learning a billion or more of parameters. They have proven to be very effective in solving prediction tasks in computer vision, natural language processing, and computational biology, for example, but they mostly remain black boxes that are hard to interpret, computationally demanding, and not robust to small data perturbations. With a strong emphasis on visual modeling, the grand challenge of APHELEIA is to switch the trajectory of machine learning to a sustainable path, by developing a new generation of models that are robust, interpretable, efficient, and that do not require massive amounts of data to produce accurate predictions. To achieve this objective, we will foster new interactions between classical signal processing, statistics, optimization, and modern deep learning. Our goal is to reduce the need for massive data by enabling scientists and engineers to design trainable machine learning models that directly encode a priori knowledge of the task semantics and data formation process, while automatically prefering simple and stable solutions over complex ones.

9.4 National initiatives

9.4.1 ANR Project AVENUE

Participants: Karteek Alahari.

This ANR project 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. This project ended successfully in September 2023.

9.4.2 ANR Project BONSAI

Participants: Michael Arbel.

Project BONSAI is a multi-disciplinary project aiming at integrating knowledge produced by experts, in the form of simulators, into current machine learning frameworks through bilevel optimization for accurate and efficient inference. We address three challenges. The first one is to develop a deep learning-based approach to simulation-based inference that can adapt to data using bilevel optimization. A second challenge is to depoly the methods to real-world problems which have their specificities. A third challenge is to develop bilevel optimization methods that can handle the non-convexity and over-parameterization arising from using deep learning. The principal investigator is Michael Arbel, and the project involves participants from Toulouse School of Economics, TIMC team at UGA and other INRIA teams (Statify). This project will start in April 2024.

9.4.3 MIAI chair: Towards More Data Efficiency in Machine Learning

Participants: Julien Mairal, Karteek Alahari, Massih-Reza Amini, Margot Selosse, Juliette Marrie, Romain Menegaux, Ieva Petrulyonite.

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.

9.4.4 MIAI chair: Learning Visual Representations from Interaction for Robot Manipulation Tasks

Participants: Pia Bideau, Karteek Alahari.

How to grasp an object has been studied in computer vision and robotics and several approaches to this problem exist - either given a 3D shape of an object contact points are determined that lead to a stable hand object configuration or an other line of work aims at reconstructing stable hand object configurations modelling the reconstruction process of hand pose and object pose jointly. In both cases many solutions are possible, although a majority might not be the natural approach that humans would chose - mainly because the intention behind the grasp is omitted. This project aims at learning visual representations from interaction that encode activity information. Encoding such contextual information appears not only to be relevant to synthesise feasible grasps furthermore this is likely to enhance future generalisation skills facilitating adaptation across the same activity but different objects - grasping a cup to pour something into something shares similar motion pattern as grasping a bottle to pour something into something. Inspired by the effectiveness of human grasping, we aim at finding similarly adaptable representations that are capable of guiding complex manipulation skills. To this end we will fuse ideas relying on classical probabilistic

modeling of distributions over possible motion trajectories and latent action representations from a conditional variational autoencoder (CVAE). Both of these directions come with complementary strengths and thus provide promising capabilities of modulating the degree of action abstractions at test time to enable both coarse and fine-grained control for real world robot manipulation tasks. The chair is taking place in close collaboration with Pierre-Brice Wieber and Karteek Alahari. Currently we have two open PhD positions.

9.4.5 PEPR project Numpex

Participants: Hadrien Hendrikx.

The 'Numpex' programme's objectives are to design and develop the software building blocks required to equip future 'exascale machines' and to prepare the major application domains that aim to fully exploit the capabilities of such machines for scientific research and industry alike. This project is part of France's response to the next EuroHPC call for expressions of interest (Projet Exascale France) in hosting one of the two major exascale machines planned in Europe for 2024. In this way 'Numpex' will contribute to the creation of a set of tools, software, applications and training which will enable France to remain one of the leaders in the field of international competition through its national Exascale ecosystem that is in step with European strategy.

9.4.6 PEPR project Origins

Participants: Julien Mairal.

Thoth is involved in the axis "Direct imaging and exoplanet characterization" of the PEPR Origins. This is an on-going collaboration with astronomers from Observatoire de Paris and Lyon and with the Willow team.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

Member of the organizing committees K. Alahari was co-chair for the doctoral consortium at ICCV 2023.

10.1.2 Scientific events: selection

Member of the conference program committees K. Alahari was an area chair for CVPR 2023, ICCV 2023, NeurIPS 2023, and is an area chair for upcoming CVPR 2024, ECCV 2024.

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 ACM Multimedia, AISTATS, CVPR, ICCV, ICML, ICLR, COLT, ALT, NeurIPS in 2023.

10.1.3 Journal

Member of the editorial boards

- K. Alahari is associate editor of International Journal of Computer Vision (IJCV).

- K. Alahari was associate editor of the Computer Vision and Image Understanding (CVIU) journal.
- J. Mairal is associate editor of the Journal of Machine Learning Research (JMLR).

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), machine learning (JMLR), remote sensing (TGRS), and statistics (AOS).

10.1.4 Invited talks

- K. Alahari: Invited tutorial at Iberian Conf. Pattern Recognition and Image Analysis (IbPRIA), Alicante, Spain.
- J. Mairal: Seminar at Owkin, Paris.
- J. Mairal: Seminar at ISterre laboratory, Paris.
- P. Gaillard: invited talk at the Kick-off collaboration Inria-CWI.
- P. Gaillard: invited talk at the "Machine Learning and Industries" seminar organized by EDF R&D and frENBIS.
- H. Hendrikx: invited talk at CMAP, Polytechnique
- M. Arbel: invited talk at Bayes Comp 2023, Levi, Finland.
- M. Arbel: Plenary talk at Mathematics and Image Analysis conference, Berlin, Germany.

10.1.5 Scientific expertise

- K. Alahari was a reviewer for the Swiss National Science Foundation.
- K. Alahari was a member of the CRCN/ISFP 2023 recruitment committees at Grenoble and Rennes.

10.1.6 Research administration

- K. Alahari is a member of commission prospection postes at LJK.
- K. Alahari is responsible for the Mathematics and Computer Science specialist field at the MSTII doctoral school.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Master: K. Alahari, Advanced Machine Learning, 11.25h eqTD, M2, UGA, Grenoble.
- Master: K. Alahari, Continually Learning Visual Representations, 13.5h eqTD, M2, Grenoble INP.
- Master: K. Alahari, Graphical Models Inference and Learning, 13.5h eqTD, MVA, M2, CentraleSupelec, Paris.
- Master: K. Alahari, Introduction to computer vision, 4.5h eqTD, M1, ENS Paris.
- Master: J. Marrie, Kernel methods for statistical learning, 12h eqTD, M2, African Master on Machine Intelligence (AMMI).
- Master: J. Mairal, Kernel methods for statistical learning, 13.5h eqTD, M2, UGA, Grenoble.
- Master: P. Gaillard, Sequential Learning, 13.5h eqTD, M2, Ecole Normale Supérieure, Cachan, France.

- Master: P. Gaillard, Kernel methods for statistical learning, 18h eqTD, M2, UGA, Grenoble.
- Master: P. Gaillard, Online Convex Opitmization, 18h eqTD, M2, Sorbonne Université, Paris.
- Master: P. Gaillard, Mathematical foundations of machine learning, 13.5eqTD, M2, UGA, Grenoble.
- Master: H. Hendrikx, Generalization Properties of Machine Learning Algorithms, 18h eqTD, M2, Université Paris Saclay, Orsay.
- Master: M. Arbel, Kernel methods for statistical learning, 13.5h eqTD, M2, UGA, Grenoble.
- Master: M. Arbel, Kernel methods for statistical learning, 31h eqTD, M2 MVA, ENS Paris-Saclay, Paris-Sacaly.

10.2.2 Supervision

- Florent Bartoccioni defended his PhD in May 2023. He was co-advised by K. Alahari and P. Perez (Valeo AI).
- Hubert Leterme defended his PhD in June 2023. He was co-advised by K. Alahari, V. Perrier (Grenoble INP) and K. Polisano (UGA).
- Mert Bulent Sariyildiz defended his PhD in June 2023. He was co-advised by K. Alahari, D. Larlus (NaverLabs Europe) and Y. Kalantidis (NaverLabs Europe).
- Lina Mezghani defended his PhD in July 2023. She was co-advised by K. Alahari and P. Bojanowski (Meta AI).
- Houssam Zenati defended his PhD in September 2023. He was co-advised by J. Mairal, P. Gaillard and E. Diemert (Criteo).
- Bruno Lecouat defended his PhD in November 2023. He was co-advised by J. Mairal and J. Ponce (Inria Willow).

10.2.3 Juries

- K. Alahari: Reviewer for the PhD thesis of Enrico Fini, Univ. Trento.
- K. Alahari: Reviewer for the PhD thesis of Kranti Kumar Parida, IIT Kanpur, India.
- K. Alahari: Reviewer for the PhD thesis of Yuming Du, Ecole des Ponts ParisTech.
- K. Alahari: Reviewer for the PhD thesis of Asya Grechka, Sorbonne Université, Paris.
- K. Alahari: Reviewer for the PhD thesis of Romain Thoreau, ISAE-SUPAERO, Toulouse.
- K. Alahari: member of the CSI of Andrea Basteri (ENS-PSL), Adrien Bardes (ENS-PSL), Nicolas Chahine (ENS-PSL), Deniz Engin (Univ. Rennes), Guillaume Le Moing (ENS), Leopold Maytie (Univ. Toulouse), Ioannis Siglidis (Univ. Paris-Est), Paul Vandame (UGA), Elliot Vincent (Univ. Paris-Est).
- J. Mairal: Reviewer for the PhD thesis of Marine Picot, McGill University.
- J. Mairal: Reviewer for the HdR thesis of Edouard Oyallon, PSL-Sorbonne University.
- J. Mairal: CSI member of Jules Bourcier, UGA.
- J. Mairal: CSI member of Badr Youbi Idrissi, Université Paris-Saclay.

10.2.4 Internal or external Inria responsibilities

- K. Alahari is heading a mission “Recrutement chercheurs” with F. Rastello, Inria Grenoble.
- K. Alahari is a board member of Inria Alumni.
- K. Alahari is a member of commission des emplois scientifiques at Inria Grenoble
- J. Mairal is a member of the scientific committee (COS) of Inria Grenoble.

10.2.5 Interventions

- K. Alahari gave a public talk on Artificial Intelligence at Université Inter-Âges Dauphiné, Grenoble.
- P. Gaillard: invited talk at "Séminaire des cadres du Département de l’Isère".

11 Scientific production

11.1 Publications of the year

International journals

[1] F. Bartoccioni, É. Zablocki, P. Pérez, M. Cord and K. Alahari. ‘LiDARTouch: Monocular metric depth estimation with a few-beam LiDAR’. In: *Computer Vision and Image Understanding* 227 (Jan. 2023), p. 103601. DOI: 10.1016/j.cviu.2022.103601. URL: https://hal.science/hal-03508099.

[2] R. Menegaux, E. Jehanno, M. Selosse and J. Mairal. ‘Self-Attention in Colors: Another Take on Encoding Graph Structure in Transformers’. In: *Transactions on Machine Learning Research Journal* (19th Oct. 2023). URL: https://hal.science/hal-04105101.

[3] B. Rasti, A. Zouaoui, J. Mairal and J. Chanussot. ‘SUnAA: Sparse Unmixing using Archetypal Analysis’. In: *IEEE Geoscience and Remote Sensing Letters* 20 (2023), pp. 1–5. DOI: 10.1109/LGRS.2023.3284221. URL: https://hal.science/hal-04161394.

[4] A. Zouaoui, G. Muhawenayo, B. Rasti, J. Chanussot and J. Mairal. ‘Entropic Descent Archetypal Analysis for Blind Hyperspectral Unmixing’. In: *IEEE Transactions on Image Processing* 32 (8th Aug. 2023), pp. 4649–4663. DOI: 10.1109/TIP.2023.3301769. URL: https://inria.hal.science/hal-03788427.

International peer-reviewed conferences

[5] M. Alakuijala, G. Dulac-Arnold, J. Mairal, J. Ponce and C. Schmid. ‘Learning Reward Functions for Robotic Manipulation by Observing Humans’. In: ICRA 2023 - IEEE International Conference on Robotics and Automation. London, United Kingdom, 16th Nov. 2022, pp. 1–11. URL: https://inria.hal.science/hal-03997549.

[6] M. Arbel, R. Ménégaux and P. Wolinski. ‘Rethinking Gauss-Newton for learning over-parameterized models’. In: NeurIPS 2023 - Thirty-seventh Conference on Neural Information Processing Systems. La Nouvelle-Orléans, United States, 12th Dec. 2023, pp. 1–24. URL: https://hal.science/hal-04362139.

[7] T. Berrada, C. Couprie, K. Alahari and J. Verbeek. ‘Guided Distillation for Semi-Supervised Instance Segmentation’. In: WACV 2024 - IEEE/CVF Winter Conference on Applications of Computer Vision. Waikola, Hawaii, United States, 4th Jan. 2024. URL: https://inria.hal.science/hal-04341872.

[8] G. Beugnot, J. Mairal and A. Rudi. ‘GloptiNets: Scalable Non-Convex Optimization with Certificates’. In: NeurIPS 2023 - 37th Conference on Neural Information Processing Systems. New Orleans, United States, Dec. 2023, pp. 1–21. URL: https://inria.hal.science/hal-04138843.

[9] E. Fini, P. Astolfi, K. Alahari, X. Alameda-Pineda, J. Mairal, M. Nabi and E. Ricci. 'Semi-supervised learning made simple with self-supervised clustering'. In: CVPR 2023 – IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023, pp. 1–11. URL: https://inria.hal.science/hal-04073630.

[10] O. Flasseur, T. Bodrito, J. Mairal, J. Ponce, M. Langlois and A.-M. Lagrange. 'Combining multi-spectral data with statistical and deep-learning models for improved exoplanet detection in direct imaging at high contrast'. In: EUSIPCO 2023 - European Signal Processing Conference. Helsinki, Finland, 21st June 2023, pp. 1–5. URL: https://hal.science/hal-04136362.

[11] P. Gaillard, A. Saha and S. Dan. 'One Arrow, Two Kills: An Unified Framework for Achieving Optimal Regret Guarantees in Sleeping Bandits'. In: *Proceedings of Machine Learning Research*. AISTATS 2023 - 26th International Conference on Artificial Intelligence and Statistics. Vol. 206. Valence (Espagne), Spain: PMLR, 26th Oct. 2022, pp. 7755–7773. URL: https://inria.hal.science/hal-03922350.

[12] H. Hendrikx. 'A principled framework for the design and analysis of token algorithms'. In: *Proceedings of Machine Learning Research*. AISTATS 2023 - 26th International Conference on Artificial Intelligence and Statistics. Valencia, Spain, 27th Apr. 2023, pp. 470–489. URL: https://hal.science/hal-04107248.

[13] Z. Kang, E. Fini, M. Nabi, E. Ricci and K. Alahari. 'A soft nearest-neighbor framework for continual semi-supervised learning'. In: ICCV 2023 - IEEE/CVF International Conference on Computer Vision. Paris, France, 2nd Oct. 2023. URL: https://hal.science/hal-03893056.

[14] A. Koloskova, H. Hendrikx and S. U. Stich. 'Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees'. In: ICML 2023 - 40th International Conference on Machine Learning. Honolulu, Hawaii,, United States, 2nd May 2023, pp. 1–19. URL: https://hal.science/hal-04107297.

[15] J. Marrie, M. Arbel, D. Larlus and J. Mairal. 'SLACK: Stable Learning of Augmentations with Cold-start and KL regularization'. In: CVPR 2023 - IEEE/CVF Conference onComputer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023, pp. 1–17. URL: https://inria.hal.science/hal-04135386.

[16] L. Mezghani, P. Bojanowski, K. Alahari and S. Sukhbaatar. 'Think Before You Act: Unified Policy for Interleaving Language Reasoning with Actions'. In: RRL 2023 - Workshop on Reincarnating Reinforcement Learning. Kigali, Rwanda, May 2023, pp. 1–9. URL: https://inria.hal.science/hal-04107094.

[17] M. B. Sariyildiz, K. Alahari, D. Larlus and Y. Kalantidis. 'Fake it till you make it: Learning transferable representations from synthetic ImageNet clones'. In: CVPR 2023 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023, pp. 1–11. URL: https://inria.hal.science/hal-03916262.

[18] M. B. Sariyildiz, Y. Kalantidis, K. Alahari and D. Larlus. 'No Reason for No Supervision: Improved Generalization in Supervised Models'. In: ICLR 2023 - International Conference on Learning Representations. Kigali, Rwanda, 1st May 2023, pp. 1–26. URL: https://inria.hal.science/hal-03929621.

[19] H. Zenati, J. Mairal, M. Martin, E. Diemert and P. Gaillard. 'Sequential Counterfactual Risk Minimization'. In: ICML 2023 - 40th International Conference on Machine Learning. Honololu, Hawaii, United States, 23rd July 2023, pp. 1–26. URL: https://hal.science/hal-04106246.

Doctoral dissertations and habilitation theses

[20] F. Bartoccioni. 'Driving scene understanding from automotive-grade sensors'. Université Grenoble Alpes [2020-....], 30th May 2023. URL: https://theses.hal.science/tel-04193785.

[21] L. Mezghani. 'Learning goal-oriented agents with limited supervision'. Université Grenoble Alpes [2020-....], 3rd July 2023. URL: https://theses.hal.science/tel-04266339.

[22] M. B. Sariyildiz. 'On the evaluation and generalization of visual representations'. Université Grenoble Alpes [2020-....], 29th June 2023. URL: https://theses.hal.science/tel-04246476.

Reports & preprints

[23] T. Berrada, J. Verbeek, C. Couprie and K. Alahari. *Unlocking Pre-trained Image Backbones for Semantic Image Synthesis*. 9th Jan. 2024. URL: https://inria.hal.science/hal-04381466.

[24] T. Darcet, M. Oquab, J. Mairal and P. Bojanowski. *Vision Transformers Need Registers*. Inria; Meta AI, Sept. 2023, pp. 1–16. URL: https://inria.hal.science/hal-04394066.

[25] O. Flasseur, T. Bodrito, J. Mairal, J. Ponce, M. Langlois and A.-M. Lagrange. *deep PACO: Combining statistical models with deep learning for exoplanet detection and characterization in direct imaging at high contrast*. 25th May 2023. URL: https://hal.science/hal-04106501.

[26] P. Gaillard, S. Gerchinovitz and É. de Montbrun. *Adaptive approximation of monotone functions*. 13th Sept. 2023. URL: https://inria.hal.science/hal-04203136.

[27] H. Hendrikx, P. Mangold and A. Bellet. *The Relative Gaussian Mechanism and its Application to Private Gradient Descent*. 29th Aug. 2023. URL: https://hal.science/hal-04370596.

[28] B. Lecouat, Y. D. de Mont-Marin, T. Bodrito, J. Mairal and J. Ponce. *Fine Dense Alignment of Image Bursts through Camera Pose and Depth Estimation*. 8th Dec. 2023. URL: https://hal.science/hal-04337706.

[29] H. Leterme, K. Polisano, V. Perrier and K. Alahari. *From CNNs to Shift-Invariant Twin Models Based on Complex Wavelets*. 21st Apr. 2023. URL: https://hal.science/hal-03880520.

[30] H. Leterme, K. Polisano, V. Perrier and K. Alahari. *On the Shift Invariance of Max Pooling Feature Maps in Convolutional Neural Networks*. 24th Oct. 2023. URL: https://hal.science/hal-03779434.

[31] B. Marin Moreno, M. Brégère, P. Gaillard and N. Oudjane. *Reimagining Demand-Side Management with Mean Field Learning*. 27th Jan. 2023. URL: https://hal.science/hal-03972660.

[32] B. M. Moreno, M. Brégère, P. Gaillard and N. Oudjane. *Efficient Model-Based Concave Utility Reinforcement Learning through Greedy Mirror Descent*. 23rd Nov. 2023. URL: https://hal.science/hal-04302000.

[33] B. Rasti, A. Zouaoui, J. Mairal and J. Chanussot. *Image Processing and Machine Learning for Hyperspectral Unmixing: An Overview and the HySUPP Python Package*. 2023. URL: https://hal.science/hal-04180307.

[34] T. Vogels, H. Hendrikx and M. Jaggi. *Beyond spectral gap (extended): The role of the topology in decentralized learning*. 5th Jan. 2023. URL: https://hal.science/hal-04107280.