RESEARCH CENTRE

**Inria Center**
**at Université Grenoble Alpes**

2022
# 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

*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, Researcher, HDR]
- Karteek Alahari [Inria, Researcher, HDR]
- Michael Arbel [Inria, Researcher]
- Pierre Gaillard [Inria, Researcher]

## Faculty Member

- Jocelyn Chanussot [Grenoble INP, until Aug 2022, HDR]

## Post-Doctoral Fellows

- Heeseung Kwon [Inria]
- Huu Dien Khue Le [Inria]
- Romain Menegaux [UGA]
- Margot Selosse [UGA, until Nov 2022]

## PhD Students

- Minttu Alakuijala [CIFRE Google until February, then Inria]
- Florent Bartoccioni [Valeo AI, CIFRE]
- Tariq Berrada Ifriqi [Inria, from Nov 2022]
- Gaspard Beugnot [Inria, joint with Sierra]
- Théo Bodrito [Inria, joint with Willow]
- Jules Bourcier [Preligens, CIFRE]
- Timothee Darcet [Facebook, CIFRE]
- Camila Fernadez Morales [NOKIA BELL LABS, CIFRE]
- Valentin Gabeur [CIFRE Google until February, then UGA, until Jul 2022]
- Ekaterina Iakovleva [UGA, until Aug 2022]
- Zhiqi Kang [Inria]
- Bruno Lecouat [Inria, joint with Willow]
- Hubert Leterme [UGA]
- Paul Liautaud [SORBONNE UNIVERSITE, from Aug 2022]
- Juliette Marrie [NaverLabs, CIFRE]
- Lina Mezghani [Facebook, CIFRE]
- Mert Sariyildiz [NaverLabs, CIFRE]
- Houssam Zenati [CRITEO, CIFRE]
- Alexandre Zouaoui [Inria]

### Technical Staff

- Loic Arbez [Inria, Engineer, from Sep 2022]
- Emmanuel Jehanno [UGA, Engineer, from Aug 2022]
- Thomas Ryckeboer [Inria, Engineer]

### Administrative Assistant

- Nathalie Gillot [Inria]

### Visiting Scientists

- Enrico Fini [University of Trento, until Mar 2022]
- David Jimenez Sierra [UNIV JAVERIANA]

## 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 basis of leveraging the script data which does not have a temporal alignment with the video is to use matches in the narrative in the script and the subtitles (which do have a temporal alignment with the video). We will go beyond simple correspondences between names and verbs relating to self-motion, and match more complex sentences related to interaction with objects and other people. To deal with the limited number of occurrences of such actions in a single movie, we will consider approaches that learn action models across a collection of movies.

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

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

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

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

### 3.3 Large-scale learning and optimization

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

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

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

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

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

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

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

# 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

The Thoth team has recruited two new permanent researchers this year: Michael Arbel and Hadrien Hendrikx.

The start-up Enhance Lab was founded in October 2022, following the PhD work of Bruno Lecouat (joint PhD student between Thoth and Willow).

### 5.1 Awards

- Mathilde Caron (PhD student from Thoth, 2018-2021) has received the best PhD award of the ELLIS society (European AI network) and an accessit for the Gilles Kahn best PhD award.
- Best paper runner up award at ICVGIP 2022 for the work [10] supervised by Karteek Alahari.
- Mert Bulent Sariyildiz was an outstanding reviewer for ECCV 2022.
- D. Khue Le received highlighted reviewer award for ICLR 2022.

## 6 New software and platforms

### 6.1 New software

#### 6.1.1 Cyanure

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

**Keyword:** Machine learning

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

**Release Contributions:** 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

## 7 New results

### 7.1 Visual Recognition

**Self-Supervised Models are Continual Learners**

**Participants:** Enrico Fini, Victor Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, Julien Mairal.

Self-supervised models have been shown to produce comparable or better visual representations than their supervised counterparts when trained offline on unlabeled data at scale. However, their efficacy is catastrophically reduced in a Continual Learning (CL) scenario where data is presented to the model sequentially. In this paper [11], we show that self-supervised loss functions can be seamlessly converted into distillation mechanisms for CL by adding a predictor network that maps the current state of the representations to their past state. This enables us to devise a framework for Continual self-supervised visual representation Learning that (i) significantly improves the quality of the learned representations, (ii) is compatible with several state-of-the-art self-supervised objectives, and (iii) needs little to no hyperparameter tuning. We demonstrate the effectiveness of our approach empirically by training six popular self-supervised models in various CL settings.

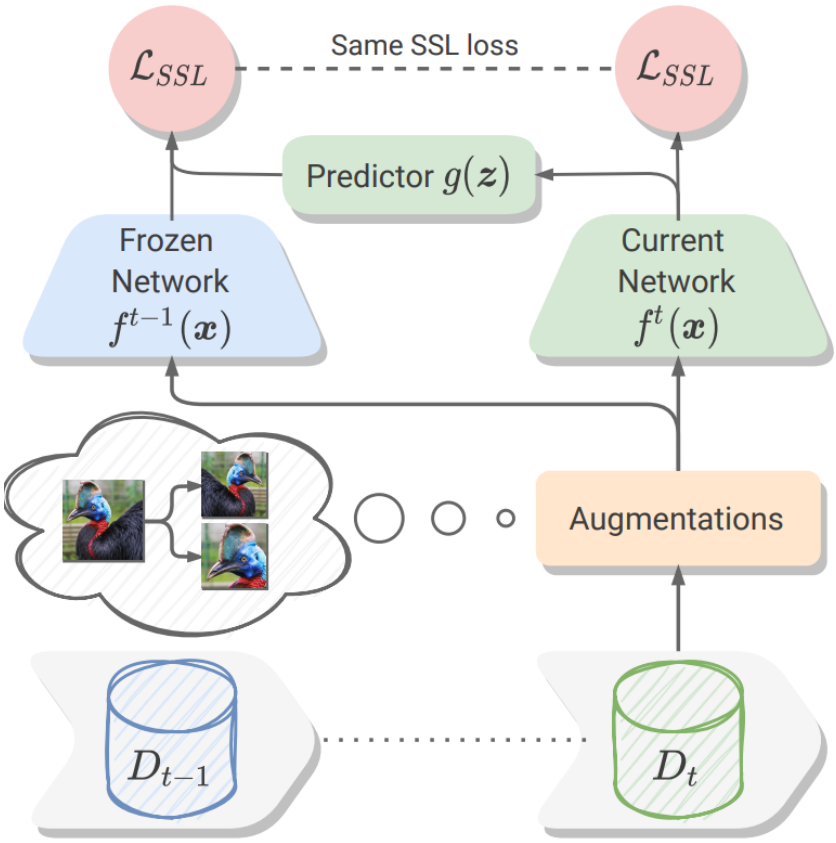



Figure 1: Overview of our framework.

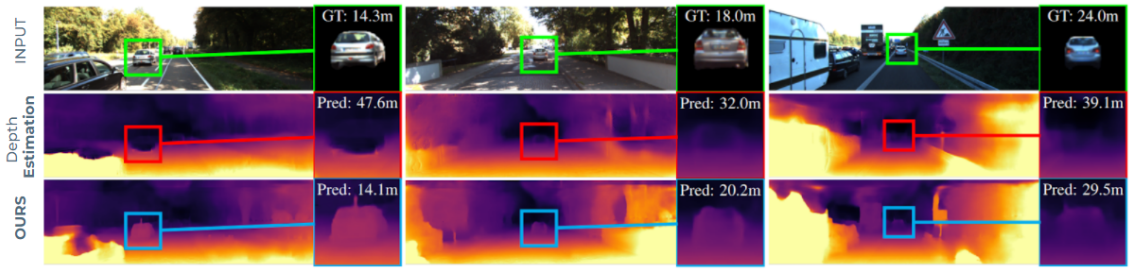



Figure 2: **Mitigation of the infinite-depth problem.** Self-supervised image-only approaches tend to predict objects with no relative-motion at an infinite depth, as indicated by the hole in the depth close-up (red). In contrast, our LiDARTouch framework estimates the depth of these vehicles, as shown in the green close-up

**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 2, 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.

**LaRa: Latents and Rays for Multi-Camera Bird's-Eye-View Semantic Segmentation**

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

Recent works in autonomous driving have widely adopted the bird's-eye-view (BEV) semantic map as an intermediate representation of the world. Online prediction of these BEV maps involves non-trivial operations such as multi-camera data extraction as well as fusion and projection into a common top-view grid. This is usually done with error-prone geometric operations (e.g., homography or back-projection from monocular depth estimation) or expensive direct dense mapping between image pixels and pixels in BEV (e.g., with MLP or attention). In this work [6], we present 'LaRa', an efficient encoder-decoder, transformer-based model for vehicle semantic segmentation from multiple cameras. Our approach uses a system of cross-attention to aggregate information over multiple sensors into a compact, yet rich, collection of latent representations. These latent representations, after being processed by a series of self-attention blocks, are then reprojected with a second cross-attention in the BEV space (an overview is

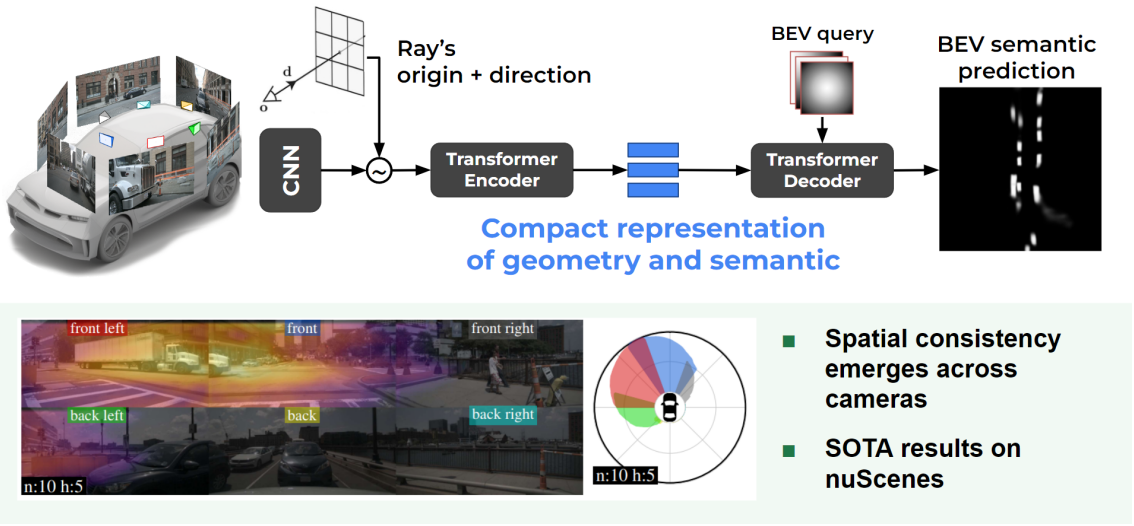



Figure 3: **LaRa overview.** At the top is a representation of our architecture. Our model compresses the scene information from the visual features and the cameras' parameters (ray embedding) into a compact but rich latent representation. The information available through the cameras is merged and efficiently processed in this latent space, then re-projected into the prediction space, here in BEV. At the bottom, the latent-to-camera attention's directional intensity is close to being continuous across cameras, highlighting that our ray embedding helps to stitch information across cameras. In the paper, we show that a spatial partitioning of the scene across latent vectors emerges.

provided in 3). We demonstrate that our model outperforms the best previous works using transformers on nuScenes. The code and trained models are available at https://github.com/valeoai/LaRa.

**Masking Modalities for Cross-modal Video Retrieval**

**Participants:** Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, Cordelia Schmid.

Pre-training on large scale unlabelled datasets has shown impressive performance improvements in the fields of computer vision and natural language processing. Given the advent of large-scale instructional video datasets, a common strategy for pre-training video encoders is to use the accompanying speech as weak supervision. However, as speech is used to supervise the pre-training, it is never seen by the video encoder, which does not learn to process that modality. We address this drawback of current pre-training methods, which fail to exploit the rich cues in spoken language, in [20]. Our proposal is to pre-train a video encoder using all the available video modalities as supervision, namely, appearance, sound, and transcribed speech. We mask an entire modality in the input and predict it using the other two modalities. This encourages each modality to collaborate with the others, and our video encoder learns to process appearance and audio as well as speech 4. We show the superior performance of our "modality masking" pre-training approach for video retrieval on the How2R, YouCook2 and Condensed Movies datasets.

**AVATAR: Unconstrained Audiovisual Speech Recognition**

**Participants:** Valentin Gabeur, Paul Hongsuck Seo, Arsha Nagrani, Chen Sun, Karteek Alahari, Cordelia Schmid.

Audiovisual automatic speech recognition (AV-ASR) is an extension of ASR that incorporates visual cues, often from the movements of a speaker's mouth. Unlike works that simply focus on the lip motion,

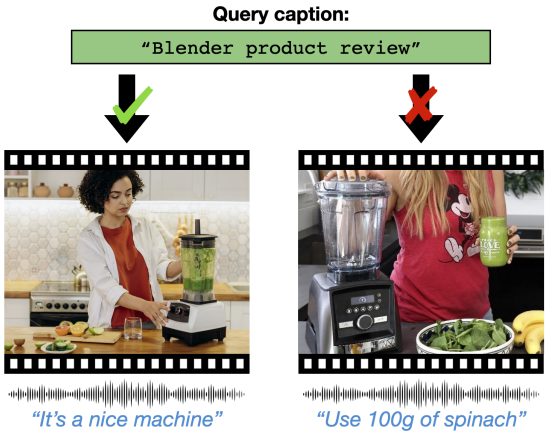



Figure 4: Speech is part of the story! Video retrieval methods that focus on visual inputs alone are likely to miss out on key information (e.g., while both the examples above contain a blender, the speech (in blue) helps identify the one for a product review). In this work, we focus on learning a video encoder to effectively process RGB and audio features, as well as transcribed speech from instructional videos online, through a novel modality masking method. Our approach learns from unlabelled videos, without the need for expensive manual captions.

we investigate the contribution of entire visual frames (visual actions, objects, background etc.). This is particularly useful for unconstrained videos, where the speaker is not necessarily visible. To solve this task, we propose a new sequence-to-sequence AudioVisual ASR TrAnsformeR (AVATAR) [19] which is trained end-to-end from spectrograms and full-frame RGB (see Figure 5). To prevent the audio stream from dominating training, we propose different word-masking strategies, thereby encouraging our model to pay attention to the visual stream. We demonstrate the contribution of the visual modality on the How2 AV-ASR benchmark, especially in the presence of simulated noise, and show that our model outperforms all other prior work by a large margin. Finally, we also create a new, realworld test bed for AV-ASR called VisSpeech, which demonstrates the contribution of the visual modality under challenging audio conditions.

**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 [26], 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 6. 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).

**Improving the Generalization of Supervised Models**

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

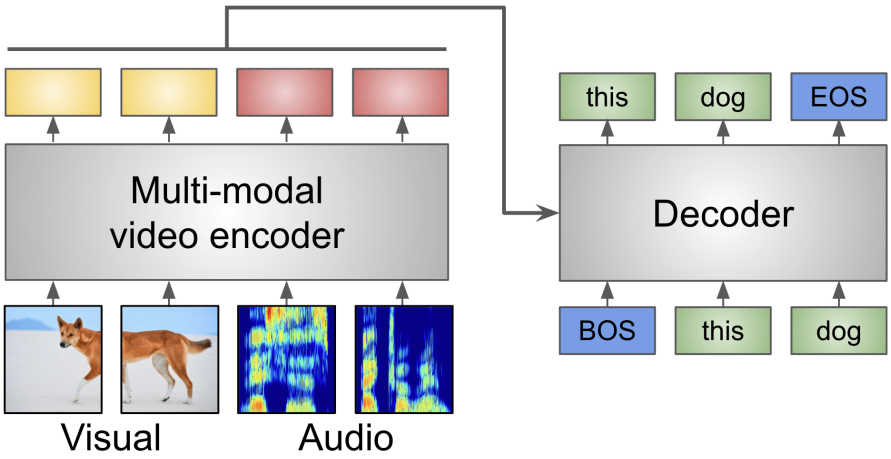



Figure 5: AVATAR: We propose a Seq2Seq architecture for audio-visual speech recognition. Our model is trained end-to-end from RGB pixels and spectrograms.

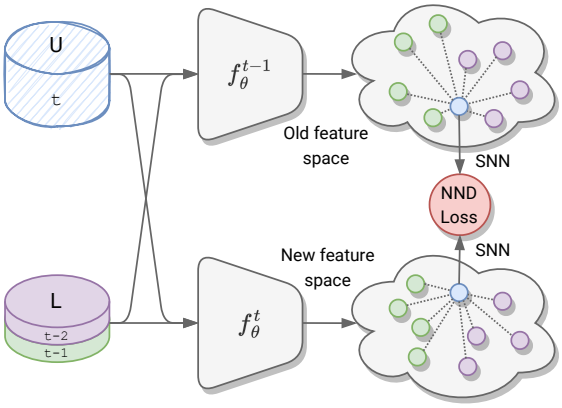



Figure 6: Illustration of our soft nearest-neighbor distillation loss.

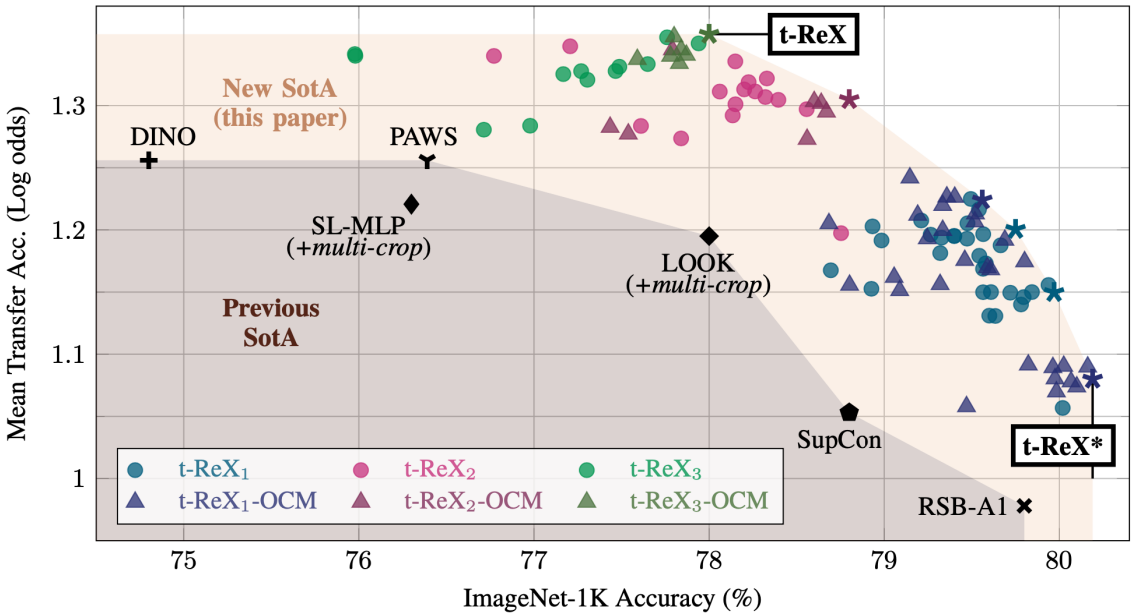


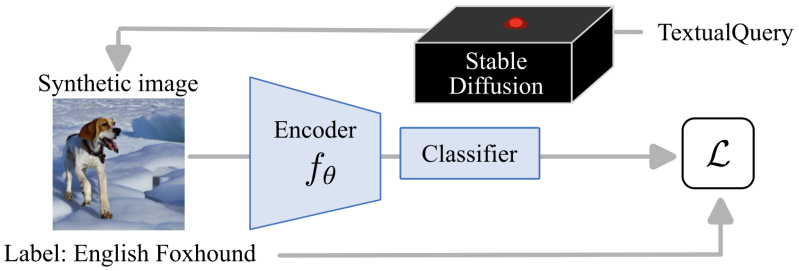

Figure 7: Comparison on the training task vs. transfer task performance for ResNet50 encoders. We report IN1K (Top-1 accuracy) and transfer performance (log odds) averaged over 13 datasets (5 ImageNet-CoG levels, Aircraft, Cars196, DTD, EuroSAT, Flowers, Pets, Food101 and SUN397) for a large number of 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 [31], 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 7) 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.

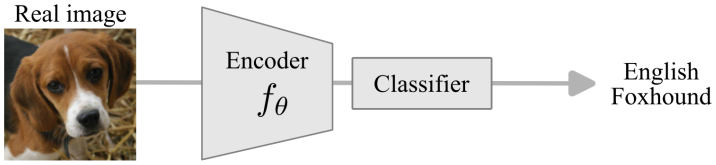
**Fake it till you make it: Learning(s) from a synthetic ImageNet clone**

**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 [30], 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, we explore the ability of Stable Diffusion to generate synthetic clones of ImageNet and measure how




(a) Training a model on synthetic images.




(b) Testing the frozen model on real images.

Figure 8: 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.

useful they are for training classification models from scratch. An overview of our experimental protocol is shown in Figure 8. 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.

**Participants:** Heeseung Kwon, Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Karteek Alahari.

Vision Transformers (ViTs) have become a dominant paradigm for visual representation learning with self-attention operators. Although these operators provide flexibility to the model with their adjustable attention kernels, they suffer from inherent limitations: (1) the attention kernel is not discriminative enough, resulting in high redundancy of the ViT layers, and (2) the complexity in computation and memory is quadratic in the sequence length. In this paper [27], we propose a novel attention operator, called lightweight structure-aware attention (LiSA), which has a better representation power with log-linear complexity (see Figure 9). Our operator learns structural patterns by using a set of relative position embeddings (RPEs). To achieve log-linear complexity, the RPEs are approximated with fast Fourier transforms. Our experiments and ablation studies demonstrate that ViTs based on the proposed operator outperform self-attention and other existing operators, achieving state-of-the-art results on ImageNet, and competitive results on other visual understanding benchmarks such as COCO and Something-Something-V2. The source code of our approach will be released online.

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

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

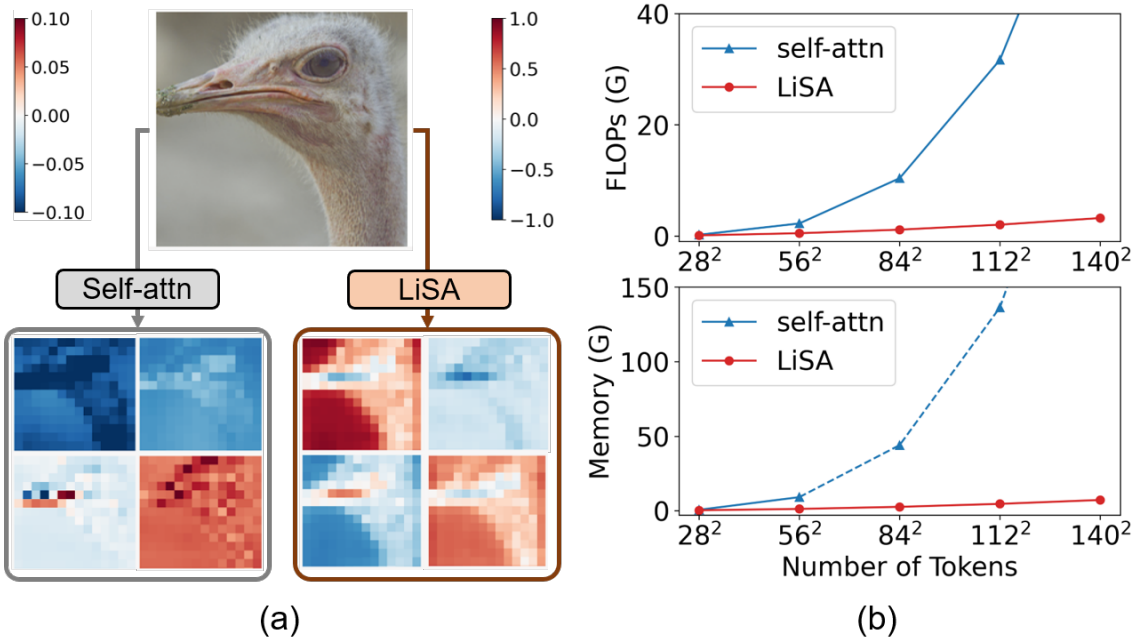



Figure 9: **Self-attention vs. LiSA.** (a) Feature visualization of self-attention & LiSA: compared to self-attention, LiSA learns better discriminative features by capturing geometric structural patterns. (b) Computation (FLOPs) & memory cost: LiSA is significantly more efficient than self-attention when the sequence length increases, due to its log-linear complexity.

In this paper [29], 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 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 10.

**From CNNs to Shift-Invariant Twin Wavelet Models**

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

In this paper [28], 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 [29], 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

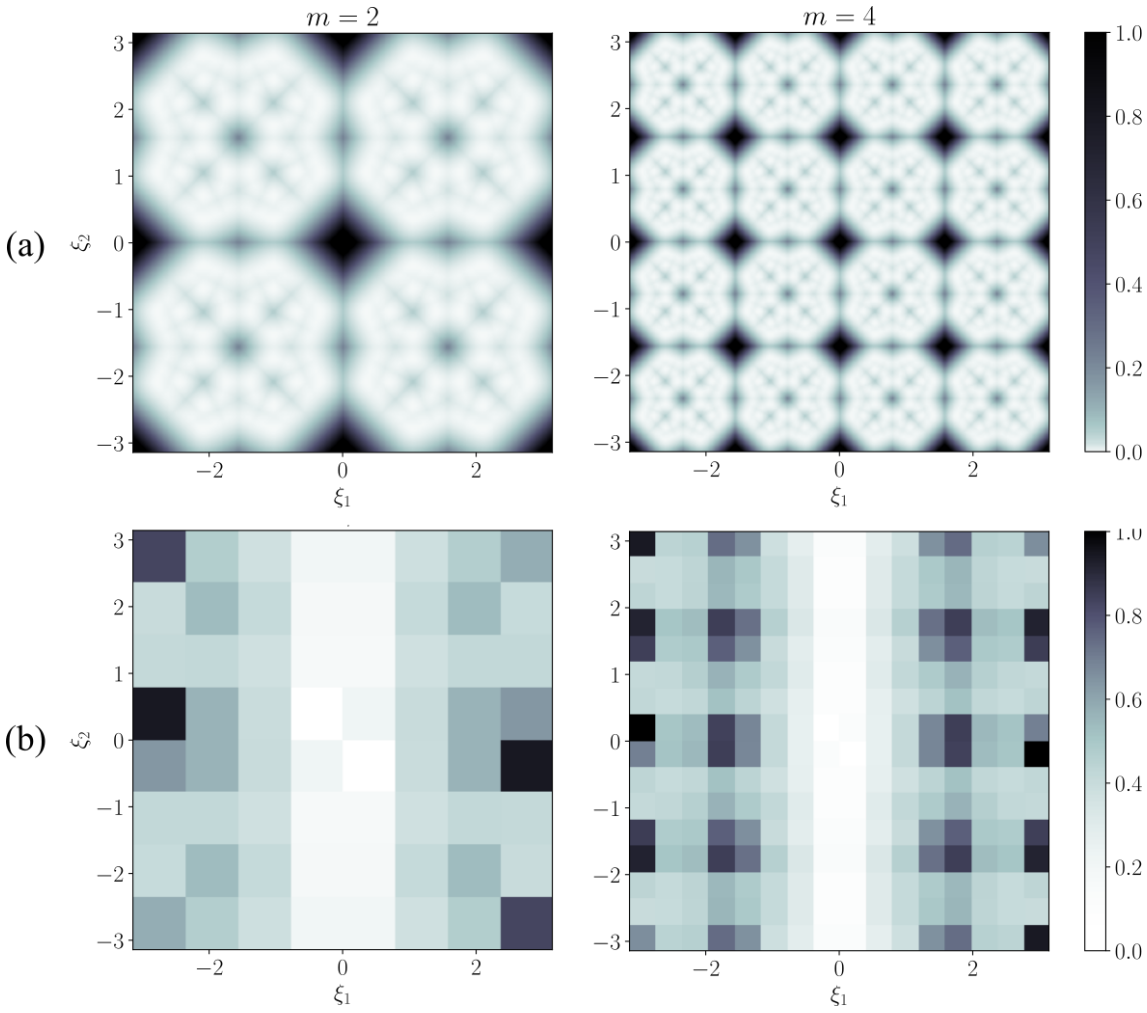



Figure 10: Top (a): expected discrepancy between complex modulus feature maps and max pooling feature maps (theoretical result). Input images are assumed to be filtered using oriented, band-pass kernels with characteristic frequencies $\boldsymbol{\xi} = (\xi_1, \xi_2) \in [-\pi, \pi]^2$, and subsampled 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).

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 11 compares the accuracy and shift invariance of the various methods.

**The Right Spin: Learning Object Motion from Rotation-Compensated Flow Fields**

**Participants:** Pia Bideau, Erik Learned-Miller, Cordelia Schmid, Karteek Alahari.

Both a good understanding of geometrical concepts and a broad familiarity with objects lead to our excellent perception of moving objects. The human ability to detect and segment moving objects works in the presence of multiple objects, complex background geometry, motion of the observer and even camouflage. How humans perceive moving objects so reliably is a longstanding research question in computer vision and borrows findings from related areas such as psychology, cognitive science and physics. One approach to the problem is to teach a deep network to model all of these effects. This contrasts with the strategy used by human vision, where cognitive processes and body design are tightly coupled and each is responsible for certain aspects of correctly identifying moving objects. Similarly from the computer vision perspective, there is evidence that classical, geometry-based techniques are better suited to the "motion-based" parts of the problem, while deep networks are more suitable for modeling appearance. In this work [21], we argue that the coupling of camera rotation and camera translation can create complex motion fields that are difficult for a deep network to untangle directly. We present a novel probabilistic model to estimate the camera's rotation given the motion field. We then rectify the flow field to obtain a rotation-compensated motion field for subsequent segmentation. This strategy of first estimating camera motion, and then allowing a network to learn the remaining parts of the problem, yields improved results on the widely used DAVIS benchmark as well as the recently published motion segmentation data set MoCA (Moving Camouflaged Animals).

**Overcoming Label Noise for Source-free Unsupervised Video Domain Adaptation**

**Participants:** Avijit Dasgupta, C. V. Jawahar, Karteek Alahari.

Despite the progress seen in classification methods, current approaches for handling videos with distribution shifts in source and target domains remain source-dependent as they require access to the source data during the adaptation stage. In this paper [10], we present a self-training based source-free video domain adaptation approach (without bells and whistles) to address this challenge by bridging the gap between the source and the target domains. We use the source pre-trained model to generate pseudo-labels for the target domain samples, which are inevitably noisy. We treat the problem of source-free video domain adaptation as learning from noisy labels and argue that the samples with correct pseudo-labels can help in the adaptation stage. To this end, we leverage the cross-entropy loss as an indicator of the correctness of pseudo-labels, and use the resulting small-loss samples from the target domain for fine-tuning the model. Extensive experimental evaluations show that our method termed as CleanAdapt achieves about 7% gain over the source-only model and outperforms the state-of-the-art approaches on various open datasets.

## 7.2 Statistical Machine Learning

**On the Benefits of Large Learning Rates for Kernel Methods**

**Participants:** Gaspard Beugnot, Julien Mairal, Alessandro Rudi.

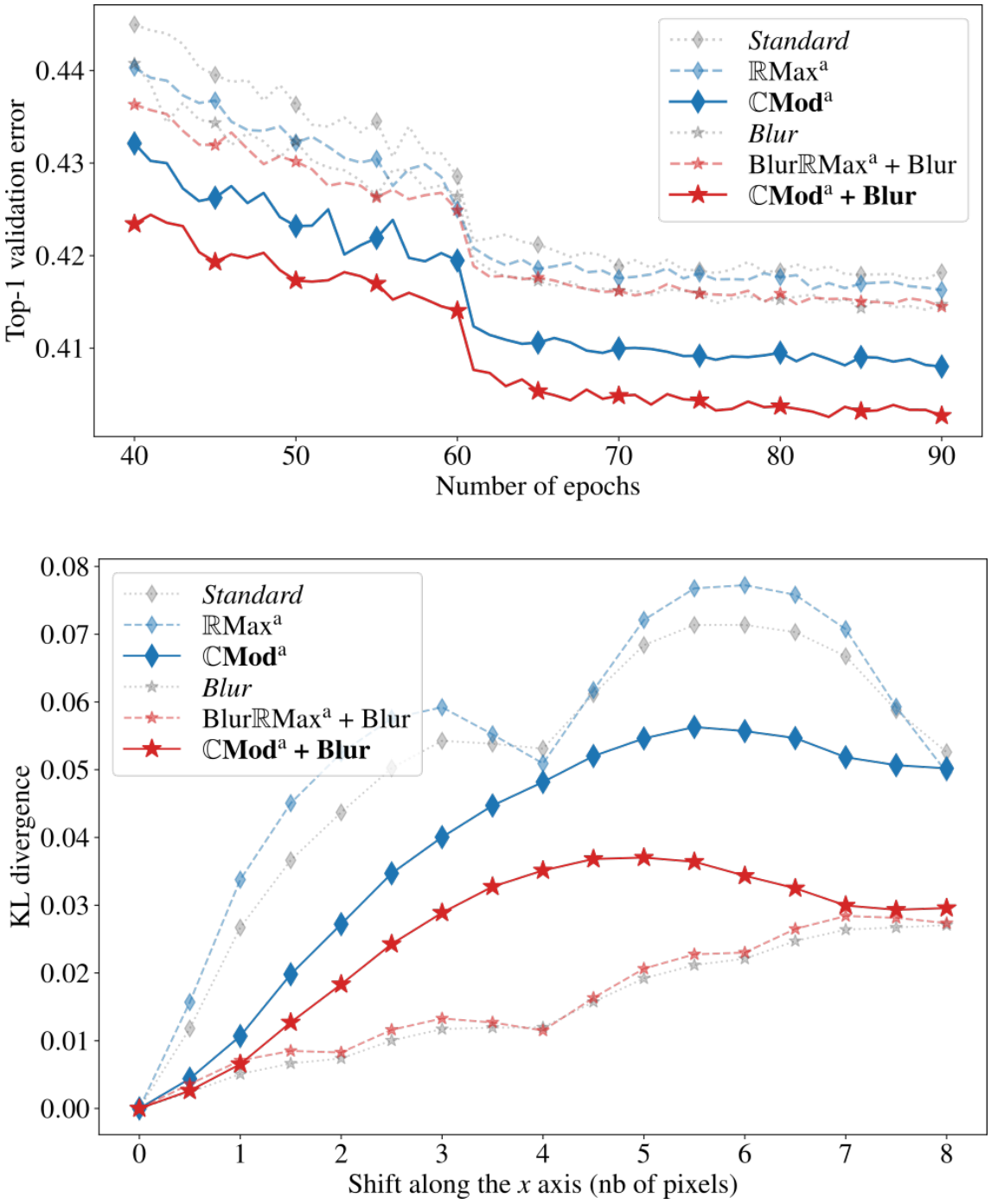



Figure 11: AlexNet and antialiased variants. Top: top-1 validation error along training with ImageNet 2012. Bottom: mean KL divergence between the outputs of a reference image versus shifted images, measuring stability with respect to small input shifts. The solid curves represent the twin models modified 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.

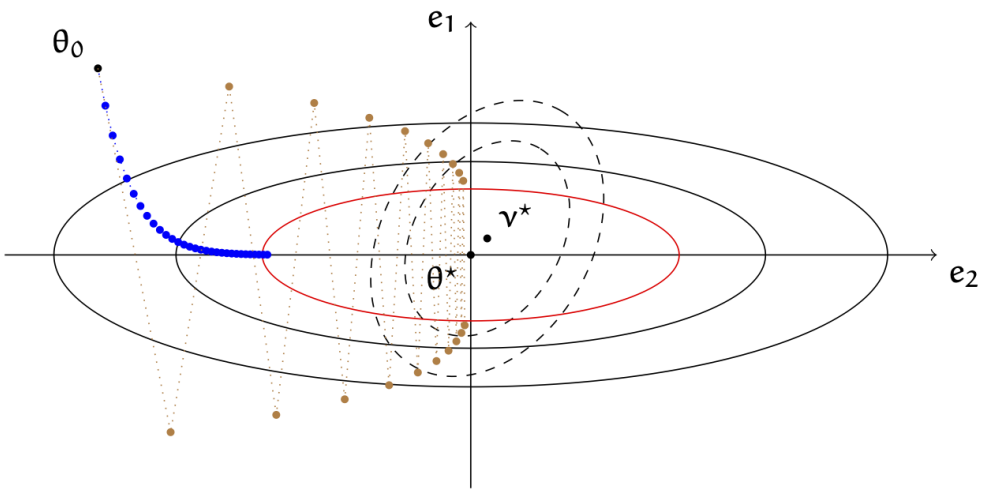



Figure 12: We optimize the quadratic training loss *(level sets are filled lines, centered in $\theta^\star$)* with gradient descent, starting from $\theta_0$ until we reach some target optimization error $\alpha$ *(filled line, red)*. However, we evaluate the quality of the estimator through the test loss *(level sets are dashed lines, centered in $v^\star$)*. Doing small step size *(blue dots)* optimizes the direction of the biggest eigenvector of the train loss first, and yields an estimate which is far from $v^\star$ in the norm induced by the test loss; doing big step size *(brown dots)* oscillates in the direction of the biggest eigenvector, but ultimately yields an estimator which is close to $v^\star$ in the norm induced by the test loss.

In this work [7], we study an intriguing phenomenon related to the good generalization performance of estimators obtained by using large learning rates within gradient descent algorithms. First observed in the deep learning literature, we show that such a phenomenon can be precisely characterized in the context of kernel methods, even though the resulting optimization problem is convex. Specifically, we consider the minimization of a quadratic objective in a separable Hilbert space, and show that with early stopping, the choice of learning rate influences the spectral decomposition of the obtained solution on the Hessian's eigenvectors, see Figure 12 for a piece of intuition. This extends previous work to realistic learning scenarios such as kernel ridge regression. While large learning rates may be proven beneficial as soon as there is a mismatch between the train and test objectives, we further explain why it already occurs in classification tasks without assuming any particular mismatch between train and test data distributions.

**Efficient Kernel UCB for Contextual Bandits**

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

In our paper [18], we tackle the computational efficiency of kernelized UCB algorithms (EK-UCB) in contextual bandits. While standard methods require a $O(CT^3)$ complexity where T is the horizon and the constant C is related to optimizing the UCB rule, we propose an efficient contextual algorithm for large-scale problems. Specifically, our method relies on incremental Nyström approximations of the joint kernel embedding of contexts and actions. This allows us to achieve a complexity of $O(CTm^2)$ where m is the number of Nyström points. To recover the same regret as the standard kernelized UCB algorithm, m needs to be of order of the effective dimension of the problem, which is at most $O(\sqrt{(T)})$ and nearly constant in some cases. Our method also works in practice as shown in Figure 13.

**Nested Bandits**

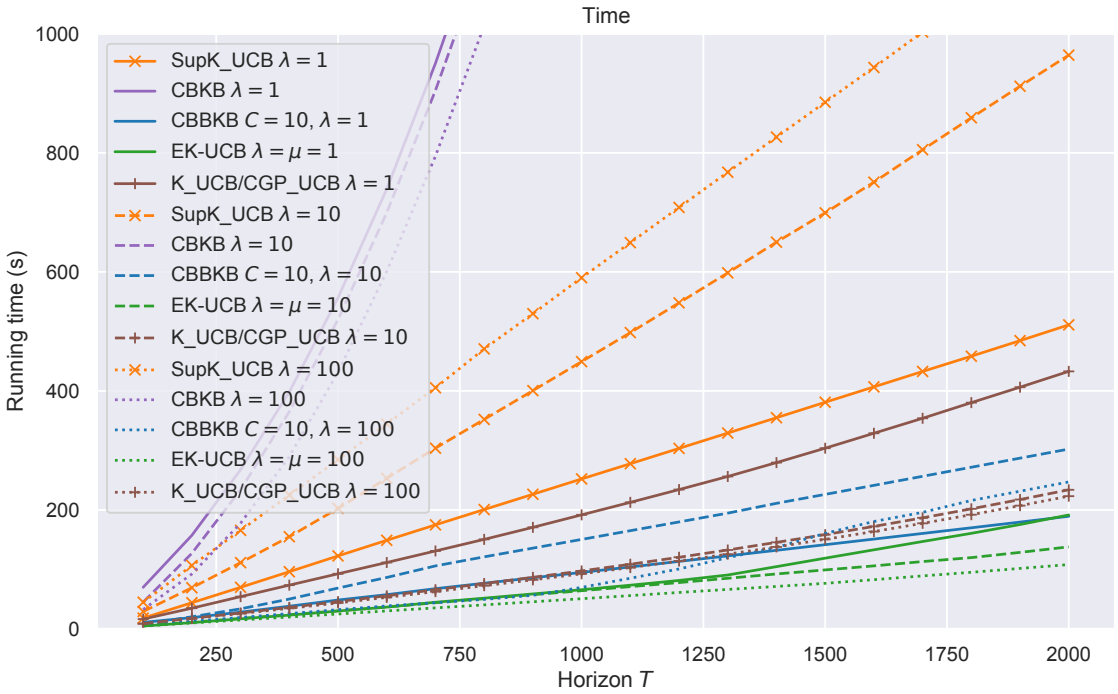



Figure 13: Example of computational times in the 'Chessboard' synthetic setting: Regret and running times of EK-UCB ($\lambda = \mu$), and baselines CBBKB ($C = 10$) and CBKB, with $T = 2000$ and with varying $\lambda$.EK-UCB achieves the lowest computational times

**Participants:** Matthieu Martin, Panagiotis Mertikopoulos, Thibaud Rahier, Houssam Zenati.

In many online decision processes, the optimizing agent is called to choose between large numbers of alternatives with many inherent similarities; in turn, these similarities imply closely correlated losses that may confound standard discrete choice models and bandit algorithms. We study this question in the context of nested bandits in [12], a class of adversarial multi-armed bandit problems where the learner seeks to minimize their regret in the presence of a large number of distinct alternatives with a hierarchy of embedded (non-combinatorial) similarities, as illustrated in Figure 14. In this setting, optimal algorithms based on the exponential weights blueprint (like Hedge, EXP3, and their variants) may incur significant regret because they tend to spend excessive amounts of time exploring irrelevant alternatives with similar, suboptimal costs. To account for this, we propose a nested exponential weights (NEW) algorithm that performs a layered exploration of the learner's set of alternatives based on a nested, step-by-step selection method. In so doing, we obtain a series of tight bounds for the learner's regret showing that online learning problems with a high degree of similarity between alternatives can be resolved efficiently, without a red bus / blue bus paradox occurring.

**Efficient and Near-Optimal Online Portfolio Selection**

**Participants:** Rémi Jézéquel, Dmitrii Ostrovskii, Pierre Gaillard.

In [25], we address the problem of online portfolio selection as formulated by Cover (1991), where a trader repeatedly distributes her capital over $d$ assets in each of $T > 1$ rounds, with the goal of maximizing the total return. Cover proposed an algorithm, termed Universal Portfolios, that performs nearly as

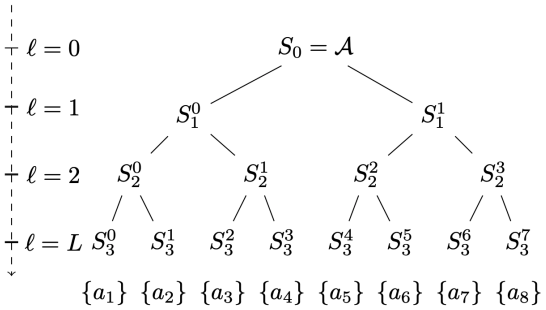



Figure 14: A similarity structure with $L = 3$ attributes on the action set $\mathcal{A} = \{a_1, \ldots, a_8\}$ that are grouped in similarity classes. For example, the class $S_2^1$ that consists of $\{a_3, a_4\}$ belongs to the bigger class$S_1^0$ .

| Algorithm | Regret | Runtime (per round) | Sources |
|---|---|---|---|
| Universal Portfolios | $d\log(T)$ | $d^4T^{14}$ | [12, 22] |
| Online Gradient Descent | $G_2\sqrt{T}$ | $d$ | [50] |
| Exponentiated Gradient | $G_\infty\sqrt{T\log(d)}$ | $d$ | [25, 21] |
| Online Newton Step (ONS) | $G_\infty d\log(T)$ | $d^2$+ generalized projection on $\Delta_d$[4] | [1, 20] |
| Soft-Bayes | $\sqrt{dT\log(d)}$ | $d$ | [39] |
| Ada-BARRONS | $d^2\log^4(T)$ | $d^{2.5}T$ | [29] |
| BISONS | $d^2\log^2(T)$ | poly$(d)$ | [49] |
| AdaMix+DONS | $d^2\log^5(T)$ | $d^3$ | [34] |
| **VB-FTRL** | $d\log(T)$ | $d^2T$ | **our paper** |

Figure 15: Regret guarantees and per-round runtime for various online portfolio selection algorithms.

well as the best (in hindsight) static assignment of a portfolio, with an $O(d\log(T))$ logarithmic regret. Without imposing any restrictions on the market this guarantee is known to be worst-case optimal, and no other algorithm attaining it has been discovered so far. Unfortunately, Cover's algorithm crucially relies on computing certain $d$-dimensional integral, which must be approximated in any implementation; this results in a prohibitive $\tilde{O}(d^4(T+d)^{14})$ per-round runtime for the fastest known implementation due to Kalai and Vempala (2002). We propose an algorithm for online portfolio selection that admits essentially the same regret guarantee as Universal Portfolios—up to a constant factor and replacement of $\log(T)$ with $\log(T+d)$—yet has a drastically reduced runtime of $\tilde{O}(d^2(T+d))$ per round. The selected portfolio minimizes the observed logarithmic loss regularized with the log-determinant of its Hessian—equivalently, the hybrid logarithmic-volumetric barrier of the polytope specified by the asset return vectors. As such, our work reveals surprising connections of online portfolio selection with two classical topics in optimization theory: cutting-plane and interior-point algorithms.

**A Unified Framework for Achieving Optimal Regret Guarantees in Sleeping Bandits**

**Participants:** Pierre Gaillard, Aadirupa Saha, Soham Dan.

We address the problem of *'Internal Regret'* in *Sleeping Bandits* in the fully adversarial setup, as well as draw connections between different existing notions of sleeping regrets in the multiarmed bandits (MAB) literature and consequently analyze the implications: Our first contribution is to propose the new notion of *Internal Regret* for sleeping MAB. We then proposed an algorithm that yields sublinear regret in that measure, even for a completely adversarial sequence of losses and availabilities. We further show that a low sleeping internal regret always implies a low external regret, and as well as a low policy regret for iid sequence of losses. The main contribution of this work precisely lies in unifying different notions of existing regret in sleeping bandits and understand the implication of one to another. Finally, we also extend our results to the setting of *Dueling Bandits* (DB)–a preference feedback variant of MAB, and proposed a reduction to MAB idea to design a low regret algorithm for sleeping dueling bandits with

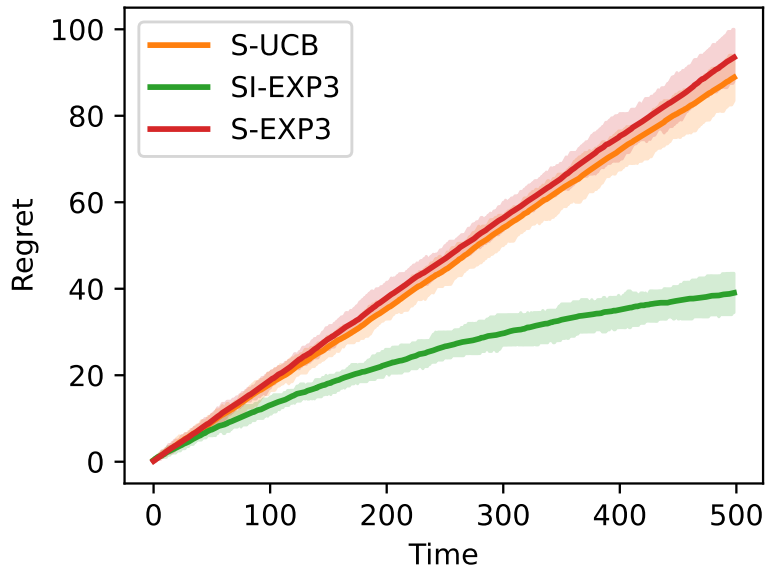



Figure 16: Comparison of the cumulative regret of a few procedures on a two-zero-sum game with sleeping actions.

stochastic preferences and adversarial availabilities. The efficacy of our algorithms is justified through empirical evaluations.

**Versatile Dueling Bandits: Best-of-both-World Analyses for Online Learning from Preferences**

**Participants:** Pierre Gaillard, Aadirupa Saha.

We study the problem of $K$-armed dueling bandit for both stochastic and adversarial environments, where the goal of the learner is to aggregate information through relative preferences of pair of decisions points queried in an online sequential manner. We propose a novel reduction from any (general) dueling bandits to multi-armed bandits and despite the simplicity, it allows us to improve many existing results in dueling bandits. Moreover, our algorithm is also the first to achieve an optimal $O(\sum_{i=1}^{K} \frac{\log T}{\Delta_i})$ regret bound against the Condorcet-winner benchmark, which scales optimally both in terms of the arm-size $K$ and the instance-specific suboptimality gaps $\{\Delta_i\}_{i=1}^{K}$. We further justify the robustness of our proposed algorithm by proving its optimal regret rate under adversarially corrupted preferences. In summary, we believe our reduction idea will find a broader scope in solving a diverse class of dueling bandits setting, which are otherwise studied separately from multi-armed bandits with often more complex solutions and worse guarantees.

**Amortized implicit differentiation for stochastic bilevel optimization**

**Participants:** Michael Arbel, Julien Mairal.

In [4], we study a class of algorithms for solving bilevel optimization problems in both stochastic and deterministic settings when the inner-level objective is strongly convex. Specifically, we consider algorithms based on inexact implicit differentiation and we exploit a warm-start strategy to amortize the estimation of the exact gradient, see Figure 18. We then introduce a unified theoretical framework inspired by the study of singularly perturbed systems (Habets, 1974) to analyze such amortized algorithms. By using

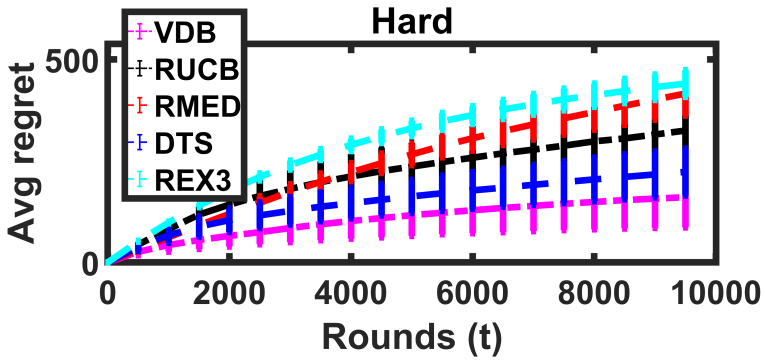



Figure 17: Comparison of the cumulative regret of a few procedures in a corrupted setting.

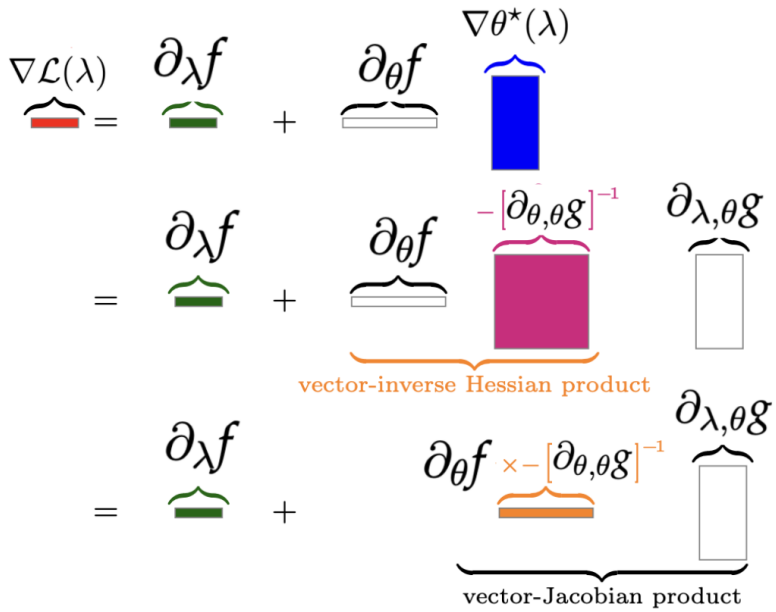



Figure 18: Implicit differentiation for computing the gradient of a bilevel objective.

this framework, our analysis shows these algorithms to match the computational complexity of oracle methods that have access to an unbiased estimate of the gradient, thus outperforming many existing results for bilevel optimization. We illustrate these findings on synthetic experiments and demonstrate the efficiency of these algorithms on hyper-parameter optimization experiments involving several thousands of variables.

**Non-Convex Bilevel Games with Critical Point Selection Maps**

**Participants:** Michael Arbel, Julien Mairal.

In [5], we provide a rigorous framework for non-convex bilevel optimization. Bilevel optimization problems involve two nested objectives, where an upper-level objective depends on a solution to a lower-level problem. When the latter is non-convex, multiple critical points may be present, leading to an ambiguous definition of the problem. In this paper, we introduce a key ingredient for resolving this ambiguity through the concept of a selection map which allows one to choose a particular solution to the lower-level problem, see Figure 19. Using such maps, we define a class of hierarchical games

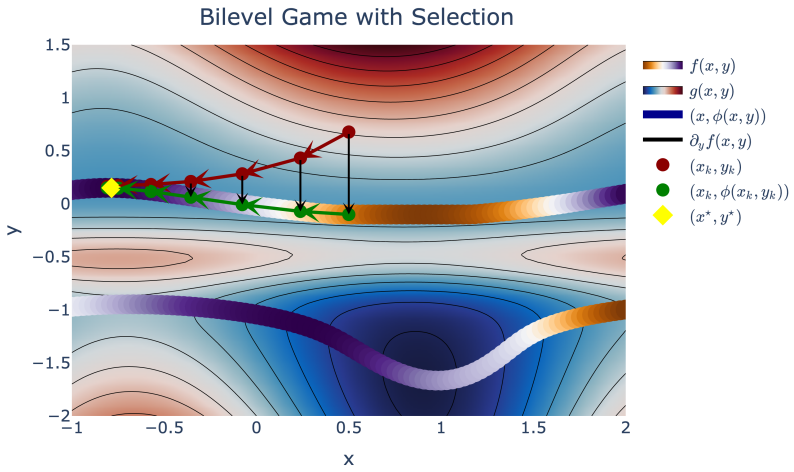



Figure 19: Bilevel game with selection. Iterates (in red ) obtained by playing a Bilevel Game with Selection. The leader finds the next update by optimizing the upper-level objective along a praticular 'critical line' of the follower's objective (iterates in green).

between two agents that resolve the ambiguity in bilevel problems. This new class of games requires introducing new analytical tools in Morse theory to characterize their evolution. In particular, we study the differentiability of the selection, an essential property when analyzing gradient-based algorithms for solving these games. We show that many existing algorithms for bilevel optimization, such as unrolled optimization, solve these games up to approximation errors due to finite computational power. Our analysis allows introducing a simple correction to these algorithms for removing the errors.

**Continual repeated annealed flow transport Monte Carlo**

**Participants:** Alex Matthews, Michael Arbel, Danilo Rezende, Arnaud Doucet.

In [13], we propose Continual Repeated Annealed Flow Transport Monte Carlo (CRAFT), a method that combines a sequential Monte Carlo (SMC) sampler (itself a generalization of Annealed Importance Sampling) with variational inference using normalizing flows. The normalizing flows are directly trained to transport between annealing temperatures using a KL divergence for each transition. This optimization objective is itself estimated using the normalizing flow/SMC approximation. We show conceptually and using multiple empirical examples that CRAFT improves on Annealed Flow Transport Monte Carlo (Arbel et al., 2021), on which it builds and also on Markov chain Monte Carlo (MCMC) based Stochastic Normalizing Flows (Wu et al., 2020). By incorporating CRAFT within particle MCMC, we show that such learnt samplers can achieve impressively accurate results on a challenging lattice field theory example, see Figure 20.

**Maximum Likelihood Learning of Energy-Based Models for Simulation-Based Inference**

**Participants:** Pierre Glaser, Michael Arbel, Arnaud Doucet, Arthur Gretton.

In [24], we introduce two synthetic likelihood methods for Simulation-Based Inference (SBI), to conduct either amortized or targeted inference from experimental observations when a high-fidelity simulator is available. Both methods learn a conditional energy-based model (EBM) of the likelihood using synthetic data generated by the simulator, conditioned on parameters drawn from a proposal distribution. The learned likelihood can then be combined with any prior to obtain a posterior estimate, from which

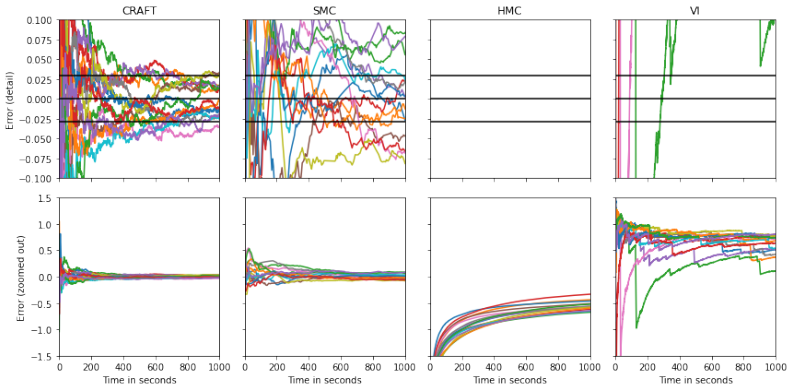



Figure 20: Comparaison between proposed methods CRAFT and State-Of-The-Art baselines: Sequential Monte Carlo Method (SMC) and Hamiltonian Monte Carlo (HMC) for estimating the log-partition function of a lattice field model. CRAFT results in more accurate estimates.

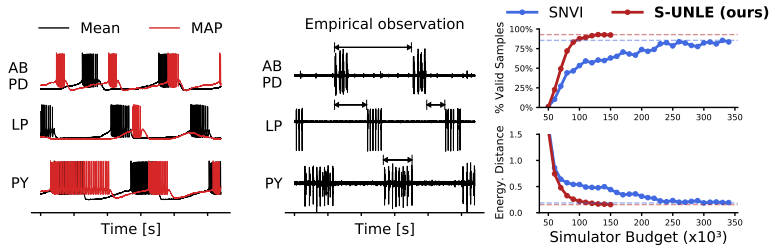



Figure 21: Inference on a model of the crab's pyloric network. The proposed method retrieves accurate posteriors parameters that match empirical observations.

samples can be drawn using MCMC. Our methods uniquely combine a flexible Energy-Based Model and the minimization of a KL loss: this is in contrast to other synthetic likelihood methods, which either rely on normalizing flows, or minimize score-based objectives; choices that come with known pitfalls. Our first method, Amortized Unnormalized Neural Likelihood Estimation (AUNLE), introduces a tilting trick during training that allows to significantly lower the computational cost of inference by enabling the use of efficient MCMC techniques. Our second method, Sequential UNLE (SUNLE), employs a robust doubly intractable approach in order to re-use simulation data and improve posterior accuracy on a specific dataset. We demonstrate the properties of both methods on a range of synthetic datasets, and apply them to a neuroscience model of the pyloric network in the crab Cancer Borealis, matching the performance of other synthetic likelihood methods at a fraction of the simulation budget, see Figure 21.

**Towards an Understanding of Default Policies in Multitask Policy Optimization**

**Participants:** Ted Moskovitz, Michael Arbel, Jack Parker-Holder.

Much of the recent success of deep reinforcement learning has been driven by regularized policy optimization (RPO) algorithms, with strong performance across multiple domains. In this family of methods, agents are trained to maximize cumulative reward while penalizing deviation in behavior from some reference, or default policy. In addition to empirical success, there is a strong theoretical foundation for understanding RPO methods applied to single tasks, with connections to natural gradient, trust region, and variational approaches. However, there is limited formal understanding of desirable properties for default policies in the multitask setting, an increasingly important domain as the field shifts towards training more generally capable agents. In [16], we take a first step towards filling this gap by formally

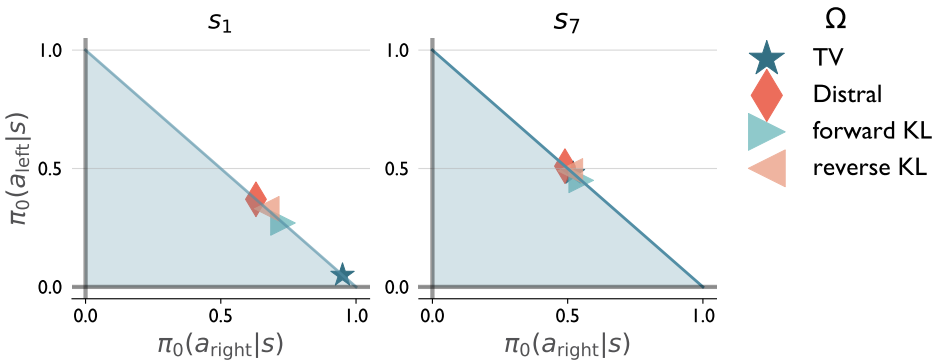



Figure 22: Default policies in two different states at $s_1$ and $s_7$ trained on five tasks with a shared structure.

linking the quality of the default policy to its effect on optimization. Using these results, we then derive a principled RPO algorithm for multitask learning with strong performance guarantees, see Figure 22.

## 7.3 Theory and Methods for Deep Neural Networks

**The spectral bias of polynomial neural networks**

**Participants:** Moulik Choraria, Leelo Dadi, Grigorios Chrysos, Julien Mairal, Volkan Cevher.

Polynomial neural networks (PNNs) have been recently shown to be particularly effective at image generation and face recognition, where high-frequency information is critical. Previous studies have revealed that neural networks demonstrate a spectral bias towards low-frequency functions, which yields faster learning of low-frequency components during training. Inspired by such studies, we conduct in [9] a spectral analysis of the Neural Tangent Kernel (NTK) of PNNs. We find that the Π-Net family, i.e., a recently proposed parametrization of PNNs, speeds up the learning of the higher frequencies. We verify the theoretical bias through extensive experiments. We expect our analysis to provide novel insights into designing architectures and learning frameworks by incorporating multiplicative interactions via polynomials.

## 7.4 Pluri-disciplinary Research and Robotics Applications

**Memory-Augmented Reinforcement Learning for Image-Goal Navigation**

**Participants:** Lina Mezghani, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, Karteek Alahari.

In this work [15], we address the problem of image-goal navigation in the context of visually-realistic 3D environments. This task involves navigating to a location indicated by a target image in a previously unseen environment. Earlier attempts, including RL-based and SLAM-based approaches, have either shown poor generalization performance, or are heavily reliant on pose/depth sensors. We present a novel method, shown in Figure 23, that leverages a cross-episode memory to learn to navigate. We first train a state-embedding network in a self-supervised fashion, and then use it to embed previously-visited states into an agent's memory. In order to avoid overfitting, we propose to use data augmentation on the RGB input during training. We validate our approach through extensive evaluations, showing that our data-augmented memory-based model establishes a new state of the art on the image-goal navigation task in the challenging Gibson dataset. We obtain this competitive performance from RGB input only, without access to additional sensors such as position or depth.

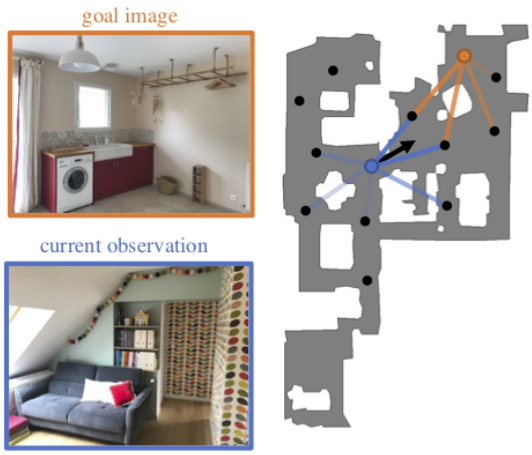



Figure 23: We tackle the problem of image-goal navigation. The agent (shown as the blue dot) is given an image from a goal location (orange dot) which it must navigate to. To address this task, our agent stores a cross-episode memory of previously visited states (black dots), and uses a navigation policy that puts attention (lines) on this memory.

**Entropic Descent Archetypal Analysis for Blind Hyperspectral Unmixing**

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

In [32], 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.

**High Dynamic Range and Super-Resolution from Raw Image Bursts**

**Participants:** Bruno Lecouat, Thomas Eboli, Jean Ponce, Julien Mairal.

Photographs captured by smartphones and mid-range cameras have limited spatial resolution and dynamic range, with noisy response in underexposed regions and color artefacts in saturated areas. This paper introduces the first approach (to the best of our knowledge) to the reconstruction of highresolution, high-dynamic range color images from raw photographic bursts captured by a handheld camera with exposure bracketing. This method uses a physically-accurate model of image formation to combine an iterative optimization algorithm for solving the corresponding inverse problem with a learned image representation for robust alignment and a learned natural image prior. The proposed algorithm is fast, with low memory requirements compared to state-of-the-art learning-based approaches to image restoration, and features that are learned end to end from synthetic yet realistic data. Extensive experiments demonstrate its excellent performance with super-resolution factors of up to ×4 on real photographs

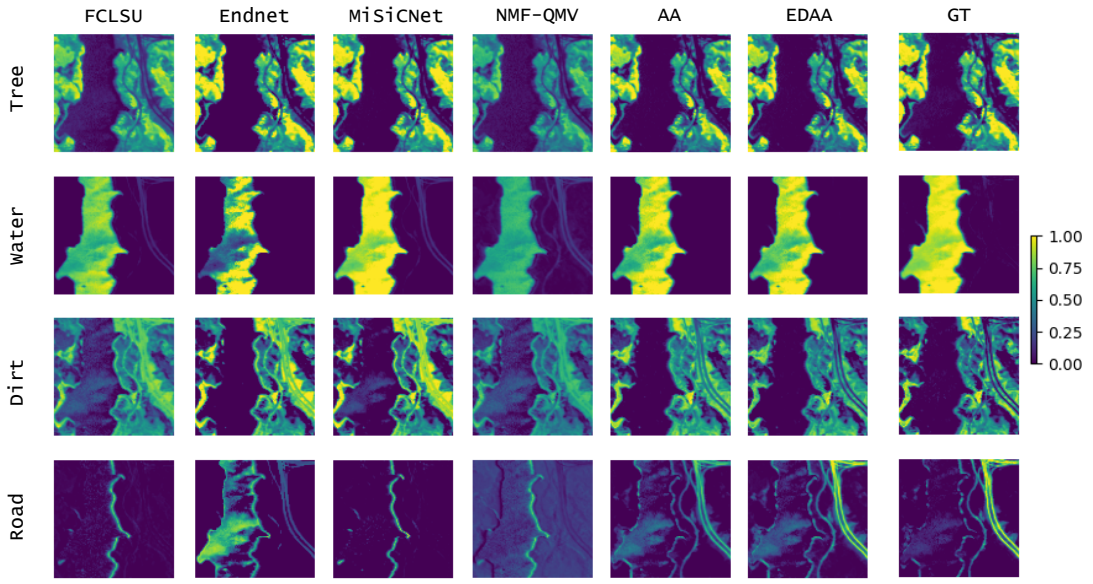



Figure 24: Estimated abundances on the Jasper Ridge dataset. The rows represent the different endmembers. The columns represent competing methods. The ground truth abundance maps are displayed on the rightmost column. Our method, EDAA, captures best the different endmembers, most notably the Road.

taken in the wild with hand-held cameras, and high robustness to low-light conditions, noise, camera shake, and moderate object motion.

**Memory-Augmented Reinforcement Learning for Image-Goal Navigation**

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

Developing agents that can execute multiple skills by learning from pre-collected datasets is an important problem in robotics, where online interaction with the environment is extremely time-consuming. Moreover, manually designing reward functions for every single desired skill is prohibitive. Prior works targeted these challenges by learning goal-conditioned policies from offline datasets without manually specified rewards, through hindsight relabeling. These methods suffer from the issue of sparsity of rewards, and fail at long-horizon tasks. In this work [14], we propose a novel self-supervised learning phase on the pre-collected dataset to understand the structure and the dynamics of the model, and shape a dense reward function, which leverages a graph built as shown in Figure 26, for learning policies offline. We evaluate our method on three continuous control tasks, and show that our model significantly outperforms existing approaches, especially on tasks that involve long-term planning.

**Evaluating the Label Efficiency of Contrastive Self-Supervised Learning for Multi-Resolution Satellite Imagery**

**Participants:** Jules Bourcier, Gohar Dashyan, Jocelyn Chanussot, Karteek Alahari.

The application of deep neural networks to remote sensing imagery is often constrained by the lack of ground-truth annotations. Addressing this issue requires models that generalize efficiently from

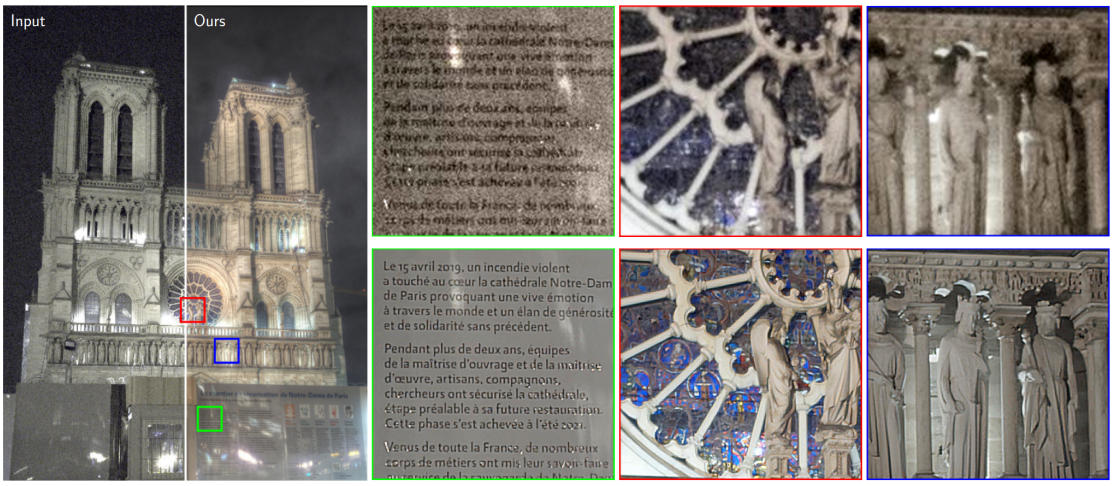



Figure 25: Examples of results.

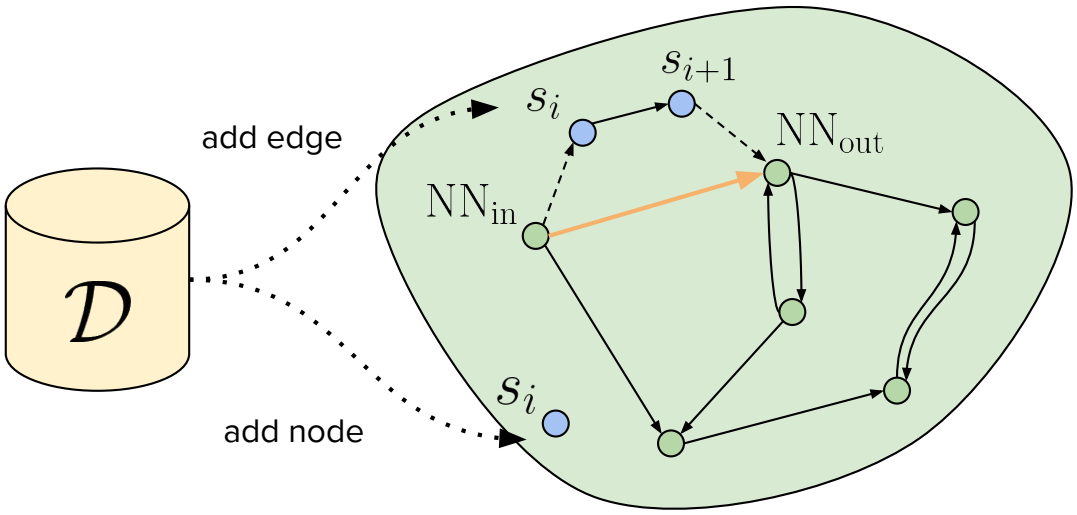



Figure 26: Overview of the graph building algorithm. Given a transition $(s_i, s_{i+1}) \in \mathscr{D}$, we add $s_i$ as node if it is distant enough from existing nodes in the graph. Moreover, we add an edge in the graph between the incoming nearest neighbor of $s_i$ and the outgoing nearest neighbor of $s_{i+1}$.

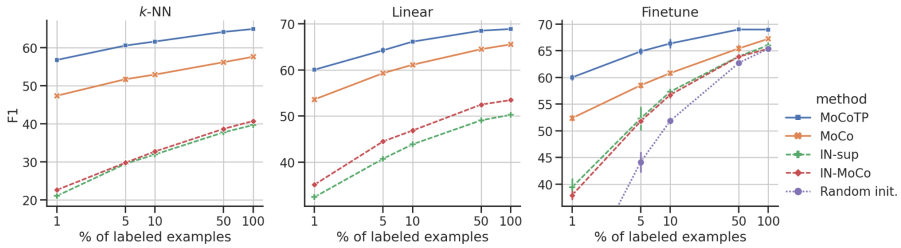



Figure 27: Label-efficient land use classification on the fMoW dataset.

limited amounts of labeled data, allowing us to tackle a wider range of Earth observation tasks. Another challenge in this domain is developing algorithms that operate at variable spatial resolutions, e.g., for the problem of classifying land use at different scales. Recently, self-supervised learning has been applied in the remote sensing domain to exploit readily-available unlabeled data, and was shown to reduce or even close the gap with supervised learning. In this paper [8], we study self-supervised visual representation learning through the lens of label efficiency, for the task of land use classification on multi-resolution/multi-scale satellite images. We benchmark two contrastive self-supervised methods adapted from Momentum Contrast (MoCo) and provide evidence that these methods can be perform effectively given little downstream supervision, where randomly initialized networks fail to generalize. Moreover, they outperform out-of-domain pretraining alternatives. We use the large-scale fMoW dataset to pretrain and evaluate the networks (see Figure 27), and validate our observations with transfer to the RESISC45 dataset.

**Self-Supervised Pretraining on Satellite Imagery: A Case Study on Label-Efficient Vehicle Detection**

**Participants:** Jules Bourcier, Thomas Floquet, Gohar Dashyan, Tugdual Ceillier, Karteek Alahari, Jocelyn Chanussot.

In defense-related remote sensing applications, such as vehicle detection on satellite imagery, supervised learning requires a huge number of labeled examples to reach operational performances. Such data are challenging to obtain as it requires military experts, and some observables are intrinsically rare. This limited labeling capability, as well as the large number of unlabeled images available due to the growing number of sensors, make object detection on remote sensing imagery highly relevant for self-supervised learning. In [22], we study in-domain self-supervised representation learning for object detection on very high resolution optical satellite imagery, that is yet poorly explored. For the first time to our knowledge, we study the problem of label efficiency on this task. We use the large land use classification dataset Functional Map of the World to pretrain representations with an extension of the Momentum Contrast framework. We then investigate this model's transferability on a real-world task of fine-grained vehicle detection and classification on Preligens proprietary data, which is designed to be representative of an operational use case of strategic site surveillance, see Figure 28. We show that our in-domain self-supervised learning model is competitive with ImageNet pretraining, and outperforms it in the low-label regime.

# 8 Bilateral contracts and grants with industry

## 8.1 Bilateral contracts with industry

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

We currently have

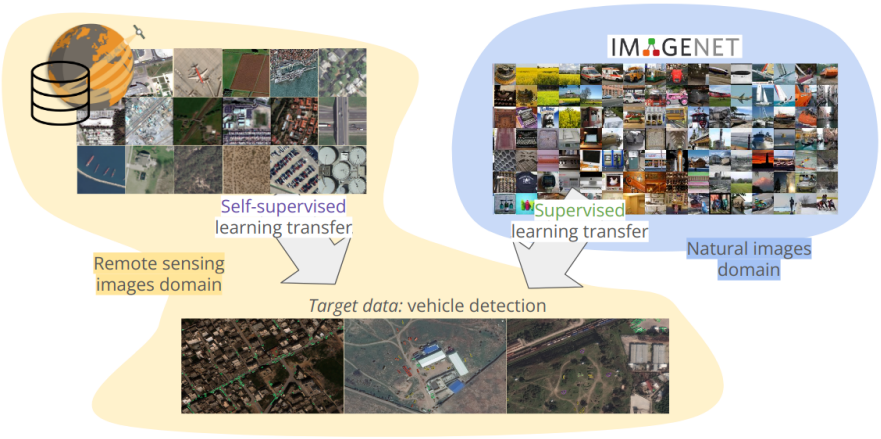



Figure 28: Overview of our study.

- one CIFRE PhD student with Criteo (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).
- two CIFRE PhD students with Google: Minttu Alakuijala (co-advised by J. Mairal) and Valentin Gabeur (co-advised by K. Alahari), both graduated in 2022.
- 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**

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

**Duration:** 2020 ->

**Coordinator:** Peter Grünwald (pdg@cwi.nl)

**Partners:**

- CWI Amsterdam (Pays-Bas)

**Inria contact:** Pierre Gaillard

**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.

**GAYA**

**Title:** Semantic and Geometric Models for Video Interpretation

**Duration:** 2019 -> 2022

**Coordinator:** Katerina Fragkiadaki (katef@cs.cmu.edu)

**Partners:**

- Carnegie Mellon University Pittsburgh (États-Unis)

**Inria contact:** Karteek Alahari

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

## 9.2 International research visitors

### 9.2.1 Visits of international scientists

**Other international visits to the team** Enrico Fini (PhD student from University of Trento) has visited us until March 2022.

## 9.3 European initiatives

### 9.3.1 H2020 projects

**ERC Starting grant SOLARIS**

**Participants:** Julien Mairal.

The project SOLARIS started in March 2017 and ended in February 2022. The goal of the project is to set up methodological and theoretical foundations of deep learning models, in the context of large-scale data processing. The main applications of the tools developed in this project are for processing visual data, such as videos, but also structured data produced in experimental sciences, such as biological sequences. The main paradigm used in the project is that of kernel methods and consist of building functional spaces where deep learning models live. By doing so, we want to derive theoretical properties of deep learning models that may explain their success, and also obtain new tools with better stability properties. Another work package of the project is focused on large-scale optimization, which is a key to obtain fast learning algorithms.

## 9.4 National initiatives

### 9.4.1 ANR Project AVENUE

**Participants:** Karteek Alahari.

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

### 9.4.2 MIAI chair: Towards More Data Efficiency in Machine Learning

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

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

# 10 Dissemination

**Participants:** Julien Mairal, Karteek Alahari, Pierre Gaillard, Michael Arbel.

## 10.1 Promoting scientific activities

### 10.1.1 Scientific events: organisation

- Julien Mairal was tutorial chair for CVPR 2022.
- Julien Mairal was area chair for ICLR 2022, ICML 2022, NeurIPS 2022.
- Julien Mairal was a member of the organizing commmittee of the Continuous Time Perspectives in Machine Learning workshop at ICML 2022.
- Karteek Alahari was area chair for ECCV 2022, WACV 2022, and is currently area chair for CVPR 2023, ICCV 2023.
- Karteek Alahari is doctoral consortium co-chair at ICCV 2023.
- Pierre Gaillard was member of the organizing committee of the CFOL workshop at ICML 2022.
- Michael Arbel is an area chair for ICLR 2023.
- D. Khue Le was area chair for BMVC 2022.

### 10.1.2 Scientific events: selection

**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 2022.

### 10.1.3 Journal

**Member of the editorial boards**

- Julien Mairal was associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence until March 2022.
- Julien Mairal is associate editor of JMLR.
- Julien Mairal is associate editor of JMIV.
- Karteek Alahari is associate editor of IJCV.
- Karteek Alahari is associate editor of CVIU.

**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).

### 10.1.4 Invited talks

- J. Mairal: invited talk at SIAM Imaging Science, mini-symposium, online.
- J. Mairal: invited talk at Mathematics of machine learning workshop, Bilbao.
- J. Mairal: invited talk at the SICO conference, Autrans.
- J. Mairal: seminar at UCL.
- J. Mairal: seminar at Université Paul Sabatier, Toulouse.
- K. Alahari: seminar at University of Bristol
- K. Alahari: seminar at National University of Singapore
- K. Alahari: invited tutorial at Open Data Science Conference Europe.
- P. Gaillard: seminar at Université Grenoble Alpes, Grenoble.
- M. Arbel: seminar at Université Aix Marseille, Luminy.
- M. Arbel: seminar at Eurocom, Sophia-Antipolis.
- H. Leterme: seminar at Rutgers University

### 10.1.5 Scientific expertise

- In 2022, Julien Mairal was a reviewer for ERC, the Swiss national science foundation, the Czech science foundation, and for the PSL Junior Fellows call.
- In 2022, Karteek Alahari was a reviewer for ANR, National Fund for Scientific and Technology Development Chile.
- In 2022, Pierre Gaillard was a reviewer for ANR projects.

#### 10.1.6 Research administration

- Julien Mairal is a member of the scientific committee (COS) of Inria's Grenoble research center.
- Karteek Alahari is a member of *commission des emplois scientifiques* at Inria Grenoble
- Karteek Alahari is a member of *commission prospection postes* at LJK.
- Karteek Alahari is responsible for the Mathematics and Computer Science specialist field at the MSTII doctoral school since September 2022.

### 10.2 Teaching - Supervision - Juries

#### 10.2.1 Teaching

- Master: K. Alahari, Advanced Machine Learning, 11.25h eqTD, M2, UGA, Grenoble.
- Master: K. Alahari, Understanding Big Visual Data, 13.5h eqTD, M2, Grenoble INP.
- Master: K. Alahari, Graphical Models Inference and Learning, 13.5h eqTD, MVA, M2, CentraleSupelec, Paris.
- Master: K. Alahari, Introduction to computer vision, 4.5h eqTD, M1, ENS Paris.
- Doctorat: J. Mairal, Optimization for machine learning, 6H eqTD, Geilo winter school (online).
- Doctorat: J. Mairal, Optimization for machine learning, 6H eqTD, StatLearn, Cargese (online).
- Master: J. Mairal, Kernel methods for statistical learning, 3h eqTD, M2, African Master on Machine Intelligence (AMMI).
- Master: J. Mairal, Kernel methods for statistical learning, 12.5h eqTD, M2, Ecole Normale Supérieure, Cachan, France.
- 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: M. Arbel, Kernel methods for statistical learning, 13.5h eqTD, M2, UGA, Grenoble.
- Master: H. Zenati. Cours de visualisations de donnees (approximativement 10h avec TD inclus) IASD master at PSL Paris Dauphine.

#### 10.2.2 Supervision

- Minttu Alakuijala defended her PhD in December 2022. She was co-advised by Jean Ponce, Julien Mairal and Cordelia Schmid.
- Ekaterina Iakovleva defended her PhD in December 2022. She was advised by Karteek Alahari.
- Valentin Gabeur defended his PhD in October 2022. He was co-advised by Karteek Alahari and Cordelia Schmid.
- Gregoire Mialon defended his PhD in February 2022. He was co-advised by Julien Mairal and Alexandre d'Aspremont.

### 10.2.3 Juries

- PhD: Julien Mairal was reviewer for the PhD thesis of Laurent Meunier, PSL Paris Dauphine.
- PhD: Julien Mairal was a member of the PhD thesis committee of Alfred Laugrois, UGA.
- PhD: Julien Mairal was a member of the CSI of Jules Bourcier, UGA.
- PhD: Karteek Alahari was reviewer for the PhD thesis of Arnaud Deleruyelle, Univ. Lille.
- PhD: Karteek Alahari was reviewer for the PhD thesis of Rui Dai, Univ. Côte d'Azur.
- PhD: Karteek Alahari was a member of the PhD thesis committee of Fabio Pizzati, Mines Paris.
- PhD: Karteek Alahari was a member of the PhD thesis committee of Arthur Douillard, Sorbonne Univ.
- PhD: Karteek Alahari was a member of the PhD thesis committee of Edward Beeching, INSA Lyon.
- PhD: Karteek Alahari was a member of the CSI of Abid Ali (Univ. Côte d'Azur), Alaaeldin Ali (ENS), Adrien Bardes (ENS), Hugo Cisneros (ENS), Yann Labbé (ENS), Guillaume Le Moing (ENS).

# 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* (2022). URL: https://hal.archives-ouvertes.fr/hal-03508099.

[2] Z. Kang, M. Sadeghi, R. Horaud and X. Alameda-Pineda. 'Expression-preserving face frontalization improves visually assisted speech processing'. In: *International Journal of Computer Vision* (16th Dec. 2022). URL: https://hal.archives-ouvertes.fr/hal-03902610.

[3] B. Lecouat, T. Eboli, J. Ponce and J. Mairal. 'High Dynamic Range and Super-Resolution from Raw Image Bursts'. In: *ACM Transactions on Graphics* 41.4 (July 2022), pp. 1–21. DOI: 10.1145/3528223.3530180. URL: https://hal.inria.fr/hal-03740564.

**International peer-reviewed conferences**

[4] M. Arbel and J. Mairal. 'Amortized implicit differentiation for stochastic bilevel optimization'. In: The Tenth International Conference on Learning Representations. Online, France, 2022. URL: https://hal.archives-ouvertes.fr/hal-03455458.

[5] M. Arbel and J. Mairal. 'Non-Convex Bilevel Games with Critical Point Selection Maps'. In: NeurIPS 2022 - 36th Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems (NeurIPS) 2022. New Orleans, United States, 2022, pp. 1–34. DOI: 10.48550/arXiv.2207.04888. URL: https://hal.archives-ouvertes.fr/hal-03869097.

[6] F. Bartoccioni, É. Zablocki, A. Bursuc, P. Pérez, M. Cord and K. Alahari. 'LaRa: Latents and Rays for Multi-Camera Bird's-Eye-View Semantic Segmentation'. In: *Proceedings of CoRL*. CoRL 2022 - Conference on Robot Learning. Auckland, New Zealand, 27th June 2022, pp. 1–21. URL: https://hal.archives-ouvertes.fr/hal-03875582.

[7] G. Beugnot, A. Rudi and J. Mairal. 'On the Benefits of Large Learning Rates for Kernel Methods'. In: COLT 2022 - 35th Annual Conference on Learning Theory. Vol. 178. Proceedings of Thirty Fifth Conference on Learning Theory. London, United Kingdom, 5th July 2022, pp. 254–282. URL: https://hal.inria.fr/hal-03878527.

[8] J. Bourcier, G. Dashyan, J. Chanussot and K. Alahari. 'Evaluating the Label Efficiency of Contrastive Self-Supervised Learning for Multi-Resolution Satellite Imagery'. In: *Image and Signal Processing for Remote Sensing XXVIII*. Image and Signal Processing for Remote Sensing XXVIII. Vol. 12267. Berlin, Germany, 26th Oct. 2022, 122670K. DOI: 10.1117/12.2636350. URL: https://hal.archives-ouvertes.fr/hal-03812663.

[9] M. Choraria, L. Dadi, G. G. Chrysos, J. Mairal and V. Cevher. 'The Spectral Bias of Polynomial Neural Networks'. In: ICLR 2022 - International Conference on Learning Representations. Virtual, France, 25th Apr. 2022, pp. 1–30. URL: https://hal.inria.fr/hal-03878338.

[10] A. Dasgupta, C. V. Jawahar and K. Alahari. 'Overcoming Label Noise for Source-free Unsupervised Video Domain Adaptation'. In: ICVGIP 2022 - Indian Conference on Computer Vision, Graphics and Image Processing. Gandhinagar, India: ACM, 2022. URL: https://hal.science/hal-03929619.

[11] E. Fini, V. da Costa, X. Alameda-Pineda, E. Ricci, K. Alahari and J. Mairal. 'Self-Supervised Models are Continual Learners'. In: CVPR 2022 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, United States: IEEE, 19th June 2022, pp. 1–13. DOI: 10.1109/CVPR52688.2022.00940. URL: https://hal.inria.fr/hal-03626342.

[12] M. Martin, P. Mertikopoulos, T. Rahier and H. Zenati. 'Nested bandits'. In: ICML 2022 - 39th International Conference on Machine Learning. Baltimore, United States, 17th July 2022. URL: https://hal.archives-ouvertes.fr/hal-03874048.

[13] A. G. D. G. Matthews, M. Arbel, D. J. Rezende and A. Doucet. 'Continual Repeated Annealed Flow Transport Monte Carlo'. In: International Conference on Machine Learning 2022. Baltimore, United States, 2022. URL: https://hal.archives-ouvertes.fr/hal-03869105.

[14] L. Mezghani, S. Sukhbaatar, P. Bojanowski, A. Lazaric and K. Alahari. 'Learning Goal-Conditioned Policies Offline with Self-Supervised Reward Shaping'. In: CoRL 2022- Conference on Robot Learning. Auckland, New Zealand, 14th Dec. 2022, pp. 1–15. URL: https://hal.inria.fr/hal-03869706.

[15] L. Mezghani, S. Sukhbaatar, T. Lavril, O. Maksymets, D. Batra, P. Bojanowski and K. Alahari. 'Memory-Augmented Reinforcement Learning for Image-Goal Navigation'. In: IROS - IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, Japan, 23rd Oct. 2022. URL: https://hal.inria.fr/hal-03110875.

[16] T. Moskovitz, M. Arbel, J. Parker-Holder and A. Pacchiano. 'Towards an Understanding of Default Policies in Multitask Policy Optimization'. In: 25th International Conference on Artificial Intelligence and Statistics. Volume 130: International Conference on Artificial Intelligence and Statistics. Online, France, 2022. URL: https://hal.archives-ouvertes.fr/hal-03455465.

[17] A. Saha and P. Gaillard. 'Versatile Dueling Bandits: Best-of-both-World Analyses for Online Learning from Preferences'. In: ICML 2022 - 39th International Conference on Machine Learning. Baltimore (MA), United States, 14th Feb. 2022. URL: https://hal.inria.fr/hal-03922380.

[18] H. Zenati, A. Bietti, E. Diemert, J. Mairal, M. Martin and P. Gaillard. 'Efficient Kernel UCB for Contextual Bandits'. In: *Proceedings of Machine Learning Research*. International Conference on Artificial Intelligence and Statistics. Vol. 151. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022. Valencia, Spain: PMLR, 3rd May 2022, pp. 5689–5720. URL: https://hal.archives-ouvertes.fr/hal-03575953.

**Conferences without proceedings**

[19] V. Gabeur, P. Hongsuck Seo, A. Nagrani, C. Sun, K. Alahari and C. Schmid. 'AVATAR: Unconstrained Audiovisual Speech Recognition'. In: INTERSPEECH 2022 - Conference of the International Speech Communication Association. Incheon, South Korea, 18th Sept. 2022, pp. 1–6. URL: https://hal.archives-ouvertes.fr/hal-03717330.

[20] V. Gabeur, A. Nagrani, C. Sun, K. Alahari and C. Schmid. 'Masking Modalities for Cross-modal Video Retrieval'. In: WACV 2022 - Winter Conference on Applications of Computer Vision. Waikoloa, United States, 4th Jan. 2022, pp. 1–10. URL: https://hal.archives-ouvertes.fr/hal-03420133.

**Reports & preprints**

[21] P. Bideau, E. Learned-Miller, C. Schmid and K. Alahari. *The Right Spin: Learning Object Motion from Rotation-Compensated Flow Fields*. 2nd Mar. 2022. URL: https://hal.inria.fr/hal-03593853.

[22] J. Bourcier, T. Floquet, G. Dashyan, T. Ceillier, K. Alahari and J. Chanussot. *Self-Supervised Pre-training on Satellite Imagery: a Case Study on Label-Efficient Vehicle Detection*. May 2022. URL: https://hal.archives-ouvertes.fr/hal-03823164.

[23] P. Gaillard, A. Saha and S. Dan. *One Arrow, Two Kills: An Unified Framework for Achieving Optimal Regret Guarantees in Sleeping Bandits*. 26th Oct. 2022. URL: https://hal.inria.fr/hal-03922350.

[24] P. Glaser, M. Arbel, A. Doucet and A. Gretton. *Maximum Likelihood Learning of Energy-Based Models for Simulation-Based Inference*. 24th Nov. 2022. DOI: 10.48550/arXiv.2210.14756. URL: https://hal.archives-ouvertes.fr/hal-03869080.

[25] R. Jézéquel, D. M. Ostrovskii and P. Gaillard. *Efficient and Near-Optimal Online Portfolio Selection*. 27th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03787674.

[26] Z. Kang, E. Fini, M. Nabi, E. Ricci and K. Alahari. *A soft nearest-neighbor framework for continual semi-supervised learning*. 10th Dec. 2022. URL: https://hal.archives-ouvertes.fr/hal-03893056.

[27] H. Kwon, F. M. Castro, M. J. Marin-Jimenez, N. Guil and K. Alahari. *Lightweight Structure-Aware Attention for Visual Understanding*. 30th Dec. 2022. URL: https://hal.archives-ouvertes.fr/hal-03916268.

[28] H. Leterme, K. Polisano, V. Perrier and K. Alahari. *From CNNs to Shift-Invariant Twin Wavelet Models*. 1st Dec. 2022. URL: https://hal.archives-ouvertes.fr/hal-03880520.

[29] H. Leterme, K. Polisano, V. Perrier and K. Alahari. *On the Shift Invariance of Max Pooling Feature Maps in Convolutional Neural Networks*. 16th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03779434.

[30] M. B. Sariyildiz, K. Alahari, D. Larlus and Y. Kalantidis. *Fake it till you make it: Learning(s) from a synthetic ImageNet clone*. 19th Dec. 2022. URL: https://hal.inria.fr/hal-03916262.

[31] M. B. Sariyildiz, Y. Kalantidis, K. Alahari and D. Larlus. *Improving the Generalization of Supervised Models*. 30th June 2022. URL: https://hal.inria.fr/hal-03929621.

[32] A. Zouaoui, G. Muhawenayo, B. Rasti, J. Chanussot and J. Mairal. *Entropic Descent Archetypal Analysis for Blind Hyperspectral Unmixing*. 26th Sept. 2022. URL: https://hal.inria.fr/hal-03788427.