RESEARCH CENTRE

Inria Paris Centre

IN PARTNERSHIP WITH:

Ecole normale supérieure de Paris, CNRS

2023
ACTIVITY REPORT

Project-Team
WILLOW

Embodied computer vision

IN COLLABORATION WITH: Département d'Informatique de l'Ecole Normale Supérieure

DOMAIN

Perception, Cognition and Interaction

THEME

Vision, perception and multimedia interpretation

Contents

Project-Team WILLOW

Creation of the Project-Team: 2007 June 01

Keywords

Computer sciences and digital sciences

A3.1.1. – Modeling, representation

A3.4. – Machine learning and statistics

A5.3. – Image processing and analysis

A5.4. – Computer vision

A5.10. – Robotics

A9. – Artificial intelligence

A9.1. – Knowledge

A9.2. – Machine learning

A9.3. – Signal analysis

A9.5. – Robotics

Other research topics and application domains

B9.5.1. – Computer science

B9.5.6. – Data science

1 Team members, visitors, external collaborators

Research Scientists

- Ivan Laptev [Team leader, INRIA, Senior Researcher, until Jul 2023, HDR]
- Justin Carpentier [Team leader, INRIA, Researcher, from Aug 2023]
- Stephane Caron [INRIA, Associate Professor Detachement]
- Justin Carpentier [INRIA, Researcher, until Jul 2023]
- Shizhe Chen [INRIA, Researcher, from Oct 2023]
- Cordelia Schmid [INRIA, Senior Researcher, HDR]

Faculty Member

- Jean Ponce [ENS Paris, Professor, HDR]

Post-Doctoral Fellows

- Shizhe Chen [INRIA, Post-Doctoral Fellow, until Sep 2023]
- Ewen Dantec [ENS Paris, Post-Doctoral Fellow, from Dec 2023]
- Etienne Moullet [INRIA, Post-Doctoral Fellow, from Oct 2023]
- Ajay Sathya [INRIA, Post-Doctoral Fellow, from Dec 2023]

PhD Students

- Alaaeldin Ali [FACEBOOK, until Sep 2023]
- Antoine Bambade [Corps des Ponts, Eaux et Forêts, until Aug 2023]
- Adrien Bardes [FACEBOOK]
- Theo Bodrito [INRIA]
- Oumayma Bounou [INRIA]
- Thomas Chabal [INRIA]
- Nicolas Chahine [DXO Mark]
- Elliot Chane-Sane [INRIA, until Aug 2023]
- Zerui Chen [INRIA]
- Hugo Cisneros [CTU Prague, until Feb 2023]
- Ludovic De Matteis [UNIV TOULOUSE III, from Sep 2023]
- Yann Dubois De Mont-Marin [INRIA]
- Gabriel Fiastre [INRIA, from Oct 2023]
- Matthieu Futeral-Peter [INRIA]
- Ricardo Garcia Pinel [INRIA]
- Pierre-Louis Guhur [ENS, until Feb 2023]

- Wilson Jallet [UNIV TOULOUSE III]
- Zeeshan Khan [INRIA, from Sep 2023]
- Yann Labbe [INRIA, until May 2023]
- Quentin Le Lidec [INRIA]
- Guillaume Le Moing [INRIA]
- Bruno Lecouat [Enhanced Lab, until Jun 2023]
- Zongmian Li [INRIA, until Feb 2023]
- Louis Montaut [INRIA]
- Fabian Schramm [INRIA]
- Robin Strudel [INRIA, until Feb 2023]
- Lucas Ventura [ENPC]
- Elliot Vincent [Ministère Transition]
- Antoine Yang [INRIA, until Oct 2023]

Technical Staff

- Roland Andrews [INRIA, Engineer, from Dec 2023]
- Etienne Arlaud [INRIA, Engineer]
- Umit Bora Gokbakan [INRIA, from Dec 2023]
- Megane Millan [INRIA, Engineer, from Feb 2023]
- Louis Montaut [INRIA, Engineer, from Oct 2023]
- Pierre-Guillaume Raverdy [INRIA, Engineer, from Sep 2023]
- Joris Vaillant [INRIA, Engineer, from Oct 2023]

Interns and Apprentices

- Hippolyte Bonabeau [INRIA, Intern, from May 2023 until Aug 2023]
- Mbaye Diongue [INRIA, Intern, from May 2023 until Sep 2023]
- Gabriel Fiastre [INRIA, Intern, from Apr 2023 until Sep 2023]
- Umit Bora Gokbakan [INRIA, Intern, from Apr 2023 until Oct 2023]
- Viviane Ledoux [INRIA, Intern, from Mar 2023 until Aug 2023]

Administrative Assistant

- Julien Guieu [INRIA]

Visiting Scientists

- Qikai Huang [UNIV GEORGIA, from Mar 2023]
- Armand Jordana [NYU, from Nov 2023 until Nov 2023]
- Ajay Sathya [UNIV LEUVEN, from Mar 2023 until Jun 2023]
- Kateryna Zorina [UNIV CHARLES - PRAGUE, until Sep 2023]

External Collaborator

- Josef Sivic [CTU Prague]

2 Overall objectives

2.1 Statement

Building machines that can automatically understand complex visual inputs is one of the central scientific challenges in artificial intelligence. Truly successful visual understanding technology will have a broad impact in application domains as varied as defense, entertainment, healthcare, human-computer interaction, image retrieval and data mining, industrial and personal robotics, manufacturing, scientific image analysis, surveillance and security, and transportation.

The problem is, however, very difficult due to the large variability of the visual world and the high complexity of the underling physical phenomena. For example, people easily learn how to perform complex tasks such as changing a car tire or performing resuscitation by observing other people. This involves advanced visual perception and interaction capabilities including interpreting sequences of human actions, learning new visuomotor skills from only a few example demonstrations, grounding instructions in appropriate scene elements and actions, and applying the learned skills in new environments and situations. Currently, however, there is no artificial system with a similar level of cognitive visual competence. Our goal for the next 10 years is to develop models, methods and algorithms providing sufficient level of visual intelligence to enable applications such as personal visual assistants or home robots that will, for example, prepare a meal in response to a chat request.

Despite the tremendous progress in visual recognition in the last decade, current visual recognition systems still require large amounts of carefully annotated training data, often use black-box architectures that do not model the 3D physical nature of the visual world, are typically limited to simple pattern recognition tasks such as detecting and recognizing objects from a predefined vocabulary, and do not capture real-world semantics. We plan to address these limitations with an ambitious research program that aims at developing models of the entire visual understanding process from image acquisition to the high-level embodied interpretation of visual scenes. We target learnable models that require minimal to no supervision, support complex reasoning about visual data, and are grounded in interactions with the physical world. More concretely, we will address fundamental scientific challenges along three research axes: (i) visual recognition in images and videos with an emphasis on weakly supervised learning and models grounded in the physical 3D world; (ii) learning embodied visual representations for robotic manipulation and locomotion; and (iii) image restoration and enhancement. These challenges will be tackled by a team of researchers with core expertise in computer vision and robotics, who will simultaneously advance both fields towards convergence. The complementary expertise in areas such as machine learning and natural language understanding will be gained through collaboration with relevant research teams.

We believe that foundational research should be grounded in applications and we plan to pursue applications with high scientific, societal, and/or economic impact in domains such as transportation; augmented reality; education; advanced manufacturing; and quantitative visual analysis in sciences, humanities and healthcare.

3 Research program

3.1 Visual recognition and reconstruction of images and videos

It is now possible to efficiently detect individual objects and people in cluttered images and videos. Current methods, however, rely on large-scale, manually-annotated image collections, often use black-box architectures that do not model the 3D physical nature of the visual world, and are typically limited to simple pattern recognition tasks. In this part of research program, we address these fundamental limitations. In particular, we address the following three key open challenges: (i) how to leverage available but weak annotations including text, audio and speech, (ii) how to enable automatic reasoning about visual data, and (iii) how to develop models grounded in the physical 3D world including learnable models for 3D object and scene reconstruction. We also continue theoretical work aimed at understanding the geometric underpinnings of computer vision.

Our current efforts in this area are outlined in detail in Section. 8.1.

3.2 Learning embodied representations

Computer vision has come a long way toward understanding images and videos in terms of scene geometry, object labels, locations and poses of people or classes of human actions. This “understanding”, however, remains largely disconnected from reasoning about the physical world. For example, what will happen when removing a tablecloth from a set table? What actions will be needed to resume an interrupted meal? We believe that a true *embodied* understanding of dynamic scenes from visual observations is the next major research challenge. We address this challenge by developing new models and algorithms with an emphasis on the synergy between vision, learning, robotics and natural language understanding. To this end, we study learning methods for motion planning and optimal control for known environments in state space. At the same time, we develop models and algorithms for learning visio-motor policies that do not rely on the known structure of environments and instead integrate visual perception directly into control algorithms. We also address natural language providing additional modality for more efficient learning and communication with emodied agents.

Our current efforts in this area are outlined in detail in Section 8.2.

3.3 Image restoration and enhancement

Although image processing is a mature field, it is more important than ever with the advent of high-quality camera phones, scientific applications in microscopy and astronomy and, recently, the emergence of multi-modal sensing for autonomous cars for example. In addition, it is an excellent proving ground for learning-based techniques since (a) it is in general (relatively) easy to generate realistic corrupted images from clean ones since reasonable models of the physical image corruption problem as often available (Abdelhamed et al., 2019; Nah et al., 2017), and (b) it is possible to incorporate natural image priors such as self-similarities (Buades et al., 2005) and sparsity (Mairal et al., 2009) in the modelling and optimization processes. We have conducted work on image restoration since the time of Julien Mairal’s PhD thesis, addressing problems such as demosaicking, denoising, inpainting, and inverse half-toning with a combination of sparse coding/dictionary learning methods and non-local means, then moving on to blind deblurring including motion segmentation and, more recently, deep-learning methods. In our on-going efforts we address several challenges for learning-based approaches to image restoration: (i) how to combine different modalities such as depth and RGB images to improve the quality of the joint observations; (ii) how to construct tunable, fully interpretable approaches to image restoration in a functional framework; and (iii) how to incorporate machine learning methods that go beyond the traditional fully supervised setting into the image restoration pipeline.

Our current work in this area is outlined in detail in Section 8.3.

4 Application domains

We believe that foundational modeling work should be grounded in applications. This includes (but is not restricted to) the following high-impact domains.

4.1 Automated visual assistants

The modern seamless video communication has enabled new applications in education, medicine and manufacturing, such as remote surgery or remotely-supervised product assembly. The abundance of online instructional videos further confirms the high demand of assistance including daily tasks such as cooking and gardening. Our work on embodied video understanding and on the joint modeling of vision and language will support automatic visual assistants. Similar to existing driving navigation assistants, such applications will guide people in daily living, inspection and manufacturing tasks. Some of these applications are studied within our MSR-Inria collaboration.

4.2 Robotics

In 2023, the Willow team has pursued the development of the Pinocchio library both from a scientific and software perspective. The recent versions of Pinocchio now accounts for closed-loop mechanisms (based on a proximal optimization), code source generation on GPUs, etc. All these new features make Pinocchio a unique tool to efficiently control complex robotic systems such as legged robots or industrial robots. We are now closely collaborating with Pal Robotics which plan to use Pinocchio to control its next generation of humanoid robots called Kangaroo. In the near future, the plan is to extend Pinocchio to become a generic-purposed and efficient robotic simulator simulating both rigid and compliant contact interactions between a robot and its environment, with the ambition of making Pinocchio the next golden framework for simulation in robotics, offering advanced features for optimal control, reinforcement learning, like differentiable simulation. Such features should position Pinocchio as the central simulator in Robotics.

4.3 Image restoration

We are pursuing applications of our image restoration work to personal photography, to enhance the images taken by consumer cameras and smartphones by deblurring and denoising them, and improving their spatial resolution and dynamic range. In this context, we are collaborating with DXOMark, the world leader in smartphone camera evaluation, through a CIFRE thesis. Two of the objectives are to develop a public database of portraits fully compliant with European GDRP regulations with informed consent from the models, and to automate the rating of image quality using this dataset. We also apply the mixture of physical image formation model and machine learning principles that has made our image restoration work successful to scientific fields: We collaborate with Anne-Marie Lagrange (Observatoire de Paris), Maud Langlois (SNRS/Observatoire de Lyon) and Julien Mairal (Inria) on direct exoplanet detection from ground-based telescope imagery. This work also involves a post-doc, Olivier Flasseur, and a PhD Student, Théo Bodrito. We will apply next year the same principles to molecular microscopy, in collaboration with Jean-Baptiste Masson (Institut Pasteur).

5 Social and environmental responsibility

Artificial intelligence holds great potential for improving our environment, for example, by reducing energy consumption and optimizing energy production. Computer vision, in particular, can be used to monitor emissions from coal plants and to track forest growth using satellite imagery. Autonomous drones can monitor and prevent failures of pipelines, power lines, power plants and other remote installations. However, as larger and more powerful AI models require increased compute power at training and deployment, AI itself stands for an increasingly high carbon footprint. One direction of our research aims to develop efficient and low-resource neural network models. To this end we have previously proposed Cross-Covariance Image Transformers (El-Nouby et al. NeurIPS'21) that avoid quadratic complexity in terms of image size. In 2023, we have been also working on the development of new optimization methods and associated software [36] to reduce the overall computationel burden and reduce their energetical impact when applied to industrial and practical scenarios. In the light of this devleopments, with the help of the Inria Soft infrastructure, we are considering creating a new software consortium, named **Mastro**, to accelerate the developement and the dissemination of efficient algorithmic solutions for the control of robotics systems. One objective of this consortium is to provide software solutions that reduce the computational burden and energetic consumption of modern robots currently deployed in industry or in societal sectors.

6 Highlights of the year

6.1 Awards

- C. Schmid received the Körber European Science Award;
- C. Schmid received the Helmholtz award for fundamental contributions in computer vision that have withstood the test of time;
- A. Young has been awarded a Google scholarship;
- J. Ponce has been elected member of Academia Europae.

6.2 Management plans

I. Laptev will be on a temporary leave from Inria strating from August 2023. In the coming year J. Carpentier will be replacing I. Laptev as an interim head of WILLOW benefiting from the help of J. Ponce and C. Schmid. J. Carpentier is planning to obtain HDR in 2024. The leadership of WILLOW will be reconsidered in one year depending on the circumstances.

7 New software, platforms, open data

7.1 New software

7.1.1 alignsdf

Keywords: Computer vision, 3D reconstruction

Functional Description: This is the PyTorch implementation of the AlignSDF research paper:

> AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction Zerui Chen, Yana Hasson, Ivan Laptev, Cordelia Schmid ECCV 2022

Publication: hal-03761124

Contact: Zerui Chen

Participants: Zerui Chen, Yana Hasson, Ivan Laptev, Cordelia Schmid

7.1.2 BLERC

Name: Benchmarking Learning Efficiency in Deep Reservoir Computing

Keywords: Machine learning, Continual Learning

Functional Description: Measuring learning efficiency of machine learning models.

URL: `https://github.com/hugcis/benchmark_learning_efficiency`

Publication: hal-03790477

Contact: Hugo Cisneros

7.1.3 BurstSR

Name: Super-resolution from image bursts

Functional Description: This is a research prototpye allowing to take as input a sequence of raw or rgb images produced by a smartphone or digital camera. This code produces a high quality color images with higher resolution.

Release Contributions: This new version, v0.2, introduces various improvements, as well as C++ code that accelerates the original Python code.

Publication: https://hal.inria.fr/hal-03323885

Contact: Julien Mairal

Participants: Bruno Lecouat, Julien Mairal, Jean Ponce

7.1.4 FrozenBiLM

Name: Zero-Shot Video Question Answering via Frozen Bidirectional Language Models

Keywords: Computer vision, Natural language processing, Deep learning

Functional Description: Code, datasets and models associated with the paper "Zero-Shot Video Question Answering via Frozen Bidirectional Language Models"

URL: `https://github.com/antoyang/FrozenBiLM`

Contact: Antoine Yang

7.1.5 hiveformer

Keywords: Robotics, NLP, Transformer

Functional Description: This is the PyTorch implementation of the Hiveformer research paper:

Instruction-driven history-aware policies for robotic manipulations Pierre-Louis Guhur, Shizhe Chen, Ricardo Garcia, Makarand Tapaswi, Ivan Laptev, Cordelia Schmid CoRL 2022 (oral)

Publication: guhur:hal-03775734

Contact: Pierre-Louis Guhur

Participants: Pierre-Louis Guhur, Shizhe Chen, Ricardo Garcia Pinel, Makarand Tapaswi, Ivan Laptev, Cordelia Schmid

7.1.6 HM3DAutoVLN

Name: Learning from Unlabeled 3D Environments for Vision-and-Language Navigation

Keyword: Computer vision

Functional Description: Open source release of the software package for the ECCV'22 paper by Chen et al. "Learning from Unlabeled 3D Environments for Vision-and-Language Navigation". This release provides a full implementation of the method, including code for training models, and testing on standard datasets, generated datasets as well as trained models.

URL: `https://github.com/cshizhe/HM3DAutoVLN`

Publication: hal-03890196

Contact: Shizhe Chen

Participants: Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, Ivan Laptev

7.1.7 Just Ask: Learning to Answer Questions from Millions of Narrated Videos

Keywords: Computer vision, Natural language processing, Deep learning

Functional Description: Code, datasets and models associated with the paper "Just Ask: Learning to Answer Questions from Millions of Narrated Videos"

URL: https://github.com/antoyang/just-ask

Contact: Antoine Yang

7.1.8 Pinocchio

Name: Pinocchio

Keywords: Robotics, Biomechanics, Mechanical multi-body systems

Functional Description: Pinocchio instantiates state-of-the-art Rigid Body Algorithms for poly-articulated systems based on revisited Roy Featherstone's algorithms. In addition, Pinocchio instantiates analytical derivatives of the main Rigid-Body Algorithms like the Recursive Newton-Euler Algorithms or the Articulated-Body Algorithm. Pinocchio is first tailored for legged robotics applications, but it can be used in extra contexts. It is built upon Eigen for linear algebra and FCL for collision detection. Pinocchio comes with a Python interface for fast code prototyping.

URL: https://github.com/stack-of-tasks/pinocchio

Contact: Justin Carpentier

Partner: CNRS

7.1.9 ProxSuite

Name: ProxSuite

Keywords: Conic optimization, Linear optimization, Robotics

Functional Description: ProxSuite is a collection of open-source, numerically robust, precise and efficient numerical solvers (e.g., LPs, QPs, etc.) rooted in revisited primal-dual proximal algorithms. Through ProxSuite, we aim to offer the community scalable optimizers that can deal with dense, sparse or matrix-free problems. While the first targeted application is Robotics, ProxSuite can be used in other contexts without limits.

ProxSuite is actively developed and supported by the Willow and Sierra research groups, joint research teams between Inria, École Normale Supérieure de Paris and Centre National de la Recherche Scientifique localized in France.

Contact: Justin Carpentier

7.1.10 SPE

Name: Semantics Preserving Encoder

Keywords: NLP, Adversarial attack, Word embeddings

Functional Description: Semantics Preserving Encoder is a simple, fully supervised sentence embedding technique for textual adversarial attacks.

URL: https://github.com/DavidHerel/semantics-preserving-encoder

Contact: Hugo Cisneros

Participants: Hugo Cisneros, David Herel, Daniela Hradilová

7.1.11 TubeDETR

Name: TubeDETR: Spatio-Temporal Video Grounding with Transformers

Keywords: Computer vision, Natural language processing, Deep learning

Functional Description: Code, datasets and models associated with the paper "TubeDETR: Spatio-Temporal Video Grounding with Transformers"

URL: `https://github.com/antoyang/TubeDETR`

Contact: Antoine Yang

7.1.12 vil3dref

Name: Language Conditioned Spatial Relation Reasoning for 3D Object Grounding

Keyword: Computer vision

Functional Description: Open source release of the software package for the NeurIPS'22 paper by Chen et al. "Language Conditioned Spatial Relation Reasoning for 3D Object Grounding". This release provides a full implementation of the method, including code for training models, and testing on standard datasets, as well as trained models.

URL: `https://github.com/cshizhe/vil3dref`

Publication: hal-03890174

Contact: Shizhe Chen

Participants: Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, Ivan Laptev

7.1.13 VLN-DUET

Name: Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation

Keyword: Computer vision

Functional Description: Open source release of the software package for the CVPR'22 paper by Chen et al. "Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation". This release provides a full implementation of the method, including codes for training models, and testing on standard datasets, as well as trained models.

URL: `https://github.com/cshizhe/VLN-DUET`

Publication: hal-03696868

Contact: Shizhe Chen

Participants: Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, Ivan Laptev

Participants: Jean Ponce, Justin Carpentier, Cordelia Schmid, Ivan Laptev, Etienne Arlaud, Pierre-Guillaume Raverdy, Stephane Caron, Shizhe Chen.

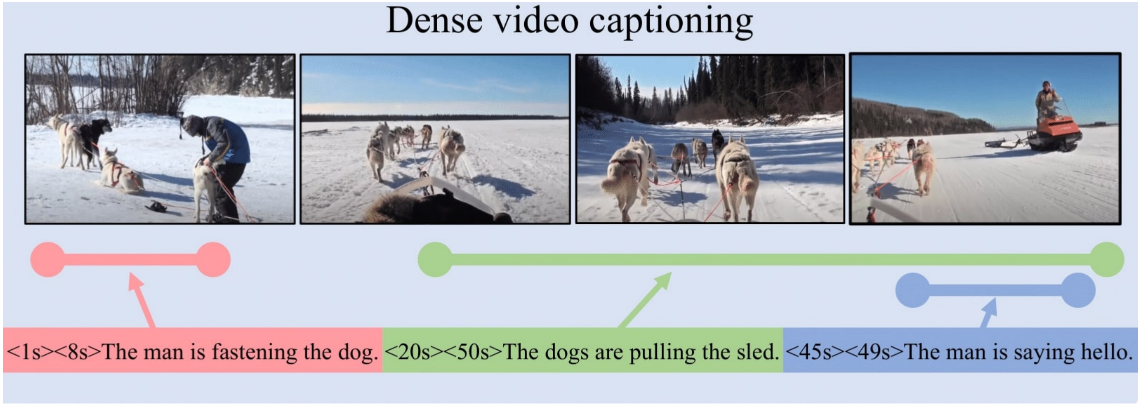

Dense video captioning
<1s><8s>The man is fastening the dog.
<20s><50s>The dogs are pulling the sled.
<45s><49s>The man is saying hello.

Figure 1: Vid2Seq is a visual language model that predicts dense event captions together with their temporal grounding in the video by generating a *single* sequence of tokens.

7.2 New platforms

Together with SED we are bulding the new robotics laboratory at Inria Paris located on the 5th floor of the C building. This laboratory is currently composed of two robotic anthropomorphic arms for manipulation experiments mounted on a fixed frame basement, the Tigao++ robot equipped with a manipulator and a moving platform as well as the SOLO quadruped robot. The robotics laboratory is also equipped with a dedicated Motion Capture system for precise object localization and robot calibration. These robotic patforms will enable our future research and experiments with locomotion navigation and manipulation.

In 2023, we have made the acquisition of one ALEGRO hand for dexterous manipulation to achieve fine manipulation of everydaylife objects.

8 New results

8.1 Visual recognition and reconstruction of images and videos

8.1.1 Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

Participants: Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, Cordelia Schmid.

In this work [24], we introduce Vid2Seq, a multi-modal single-stage dense event captioning model pretrained on narrated videos which are readily-available at scale. The Vid2Seq architecture augments a language model with special time tokens, allowing it to seamlessly predict event boundaries and textual descriptions in the same output sequence (see Figure 1). Such a unified model requires large-scale training data, which is not available in current annotated datasets. We show that it is possible to leverage unlabeled narrated videos for dense video captioning, by reformulating sentence boundaries of transcribed speech as pseudo event boundaries, and using the transcribed speech sentences as pseudo event captions. The resulting Vid2Seq model pretrained on the YT-Temporal-1B dataset improves the state of the art on a variety of dense video captioning benchmarks including YouCook2, ViTT and ActivityNet Captions. Vid2Seq also generalizes well to the tasks of video paragraph captioning and video clip captioning, and to few-shot settings.

8.1.2 VidChapters-7M: Video Chapters at Scale

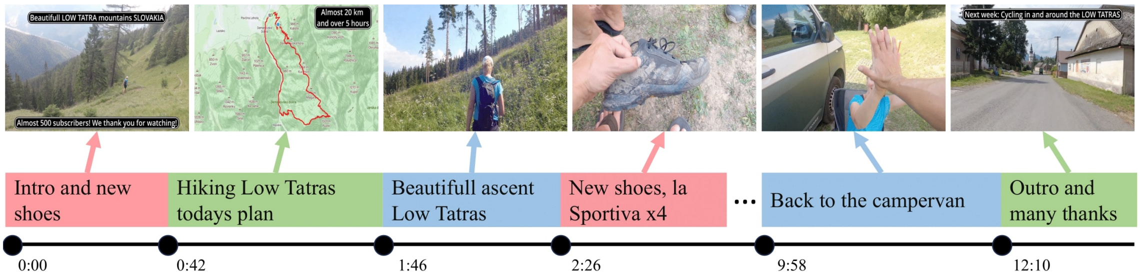

Beautifull LOW TATRA mountains SLOVAKIA
Almost 500 subscribers! We thank you for watching!
Almost 20 km and over 5 hours
Next week: Cycling in and around the LOW TATRAS
Intro and new shoes
Hiking Low Tatras todays plan
Beautifull ascent Low Tatras
New shoes, la Sportiva x4
...
Back to the campervan
Outro and many thanks
0:00
0:42
1:46
2:26
9:58
12:10

Figure 2: A video with user-annotated chapters in VidChapters-7M: the video is temporally segmented into chapters, which are annotated with a chapter title in free-form natural language.

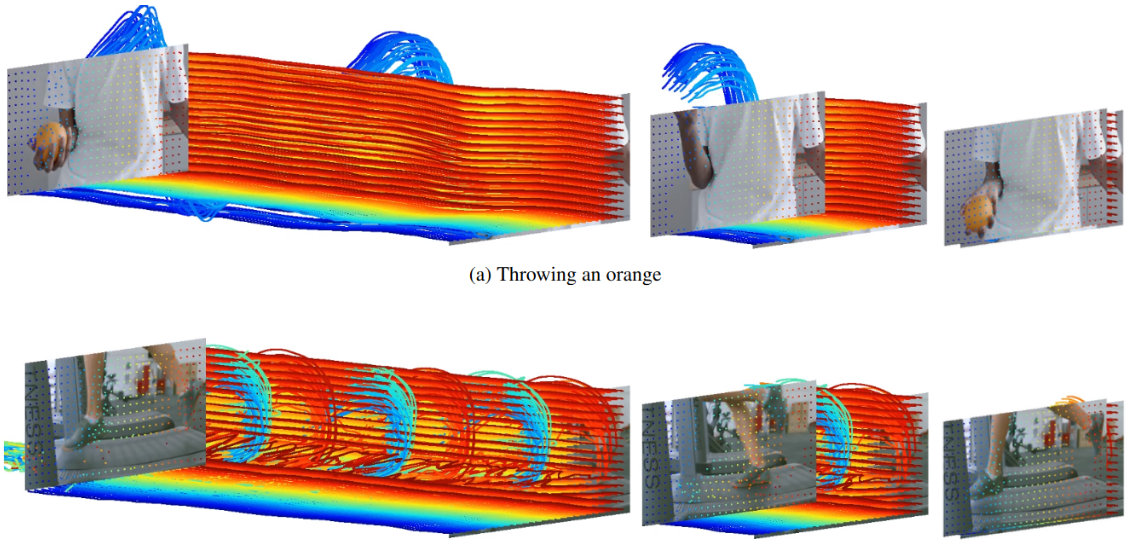
Participants: Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, Cordelia Schmid.

Segmenting long videos into chapters enables users to quickly navigate to the information of their interest. This important topic has been understudied due to the lack of publicly released datasets. To address this issue, in [23], we present VidChapters-7M, a dataset of 817K user-chaptered videos including 7M chapters in total (see Figure 2). VidChapters-7M is automatically created from videos online in a scalable manner by scraping user-annotated chapters and hence without any additional manual annotation. We introduce the following three tasks based on this data. First, the video chapter generation task consists of temporally segmenting the video and generating a chapter title for each segment. To further dissect the problem, we also define two variants of this task: video chapter generation given ground-truth boundaries, which requires generating a chapter title given an annotated video segment, and video chapter grounding, which requires temporally localizing a chapter given its annotated title. We benchmark both simple baselines and state-of-the-art video-language models for these three tasks. We also show that pretraining on VidChapters-7M transfers well to dense video captioning tasks in both zero-shot and finetuning settings, largely improving the state of the art on the YouCook2 and ViTT benchmarks. Finally, our experiments reveal that downstream performance scales well with the size of the pretraining dataset.

8.1.3 Dense Optical Tracking: Connecting the Dots

Participants: Guillaume Le Moing, Jean Ponce, Cordelia Schmid.

Recent approaches to point tracking are able to recover the trajectory of any scene point through a large portion of a video despite the presence of occlusions. They are, however, too slow in practice to track every point observed in a single frame in a reasonable amount of time. In this work [20], we introduce DOT, a novel, simple and efficient method for solving this problem. It first extracts a small set of tracks from key regions at motion boundaries using an off-the-shelf point tracking algorithm. Given source and target frames, DOT then computes rough initial estimates of a dense flow field and visibility mask through nearest-neighbor interpolation, before refining them using a learnable optical flow estimator that explicitly handles occlusions and can be trained on synthetic data with ground-truth correspondences. DOT is significantly more accurate than current optical flow techniques, outperforms sophisticated "universal" trackers like OmniMotion, and is on par with, or better than, the best point tracking algorithms like CoTracker while being at least two orders of magnitude faster. Some qualitative results are shown in Figure 3. Code, data, and videos showcasing the capabilities of our approach are available in the project webpage.



(a) Throwing an orange

(b) Running at the gym

Figure 3: Space-time visualizations. We track all pixels from the first frame of different videos, and show the trajectory for a subset of points, laid on a regular grid in the first frame. These trajectories are displayed in 3D, by using time as an additional dimension (progressing from right to left in the figure) alongside the two spatial dimensions. Each point on the grid is uniquely represented by a distinct color. Our method is able to track objects accurately, even when they go out of the frame multiple times like in video (a). DOT is able to cope with intra-object occlusions, as in video (b) where the left leg occludes the right one repeatedly.

8.1.4 WALDO: Future Video Synthesis using Object Layer Decomposition and Parametric Flow Prediction

Participants: Guillaume Le Moing, Jean Ponce, Cordelia Schmid.

Predicting the future from a video stream is an important tool for improving the versatility and safety of autonomous agents. In [21] we propose WALDO (WArping Layer-Decomposed Objects), a novel approach to the prediction of future video frames from past ones. Individual images are decomposed into multiple layers combining object masks and a small set of control points. The layer structure is shared across all frames in each video to build dense inter-frame connections. Complex scene motions are modeled by combining parametric geometric transformations associated with individual layers, and video synthesis is broken down into discovering the layers associated with past frames, predicting the corresponding transformations for upcoming ones and warping the associated object regions accordingly, and filling in the remaining image parts. Extensive experiments on the Cityscapes and KITTI datasets show that WALDO significantly outperforms prior works. Video samples synthesized by our approach are illustrated in Figure 4. More videos can be found from the project webpage.

8.1.5 Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation

Participants: Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, Rachel Bawden.

One of the major challenges of machine translation (MT) is ambiguity, which can in some cases be

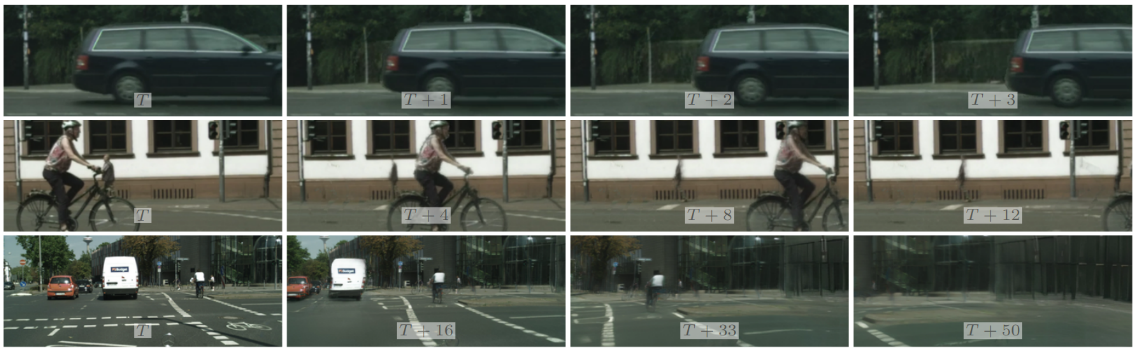


Figure 4: Video frame T and future frames $T+K$ predicted by WALDO from frames 1 to T. In our experiments, we use in general T=4 (1/4s) and K up to 50 (3s).

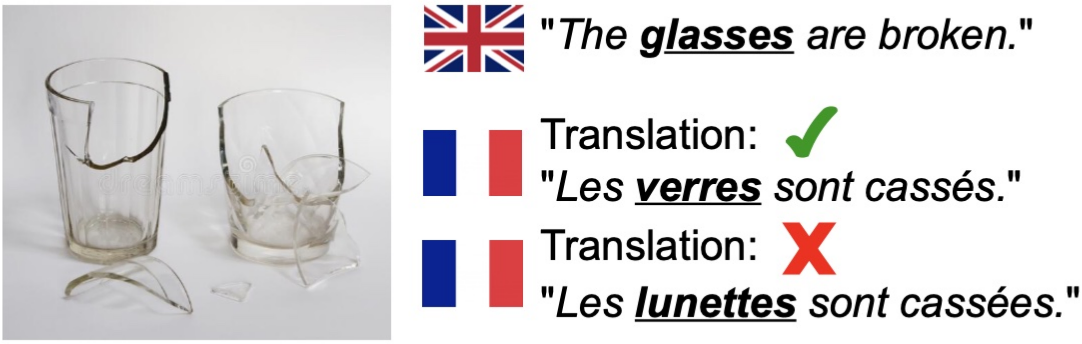

"The glasses are broken."
Translation: ✓
"Les verres sont cassés."
Translation: ✗
"Les lunettes sont cassées."

Figure 5: Visual context resolving the ambiguity of English word glasses for English-to-French translation.

resolved by accompanying context such as images, as exemplified by Figure 5. In this work [16], we present a new MMT approach based on a strong text-only MT model, which uses neural adapters, a novel guided self-attention mechanism and which is jointly trained on both visually-conditioned masking and MMT. We also introduce CoMMuTE, a Contrastive Multilingual Multimodal Translation Evaluation set of ambiguous sentences and their possible translations, accompanied by disambiguating images corresponding to each translation.

8.1.6 gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction

Participants: Zerui Chen, Shizhe Chen, Cordelia Schmid, Ivan Laptev.

Signed distance functions (SDFs) is an attractive framework that has recently shown promising results for 3D shape reconstruction from images. SDFs seamlessly generalize to different shape resolutions and topologies but lack explicit modelling of the underlying 3D geometry. In this work [14], we exploit the hand structure and use it as guidance for SDF-based shape reconstruction. In particular, we address reconstruction of hands and manipulated objects from monocular RGB images. To this end, we estimate poses of hands and objects and use them to guide 3D reconstruction. More specifically, we predict kinematic chains of pose transformations and align SDFs with highly-articulated hand poses. We improve the visual features of 3D points with geometry alignment and further leverage temporal information to enhance the robustness to occlusion and motion blurs. We conduct extensive experiments on the

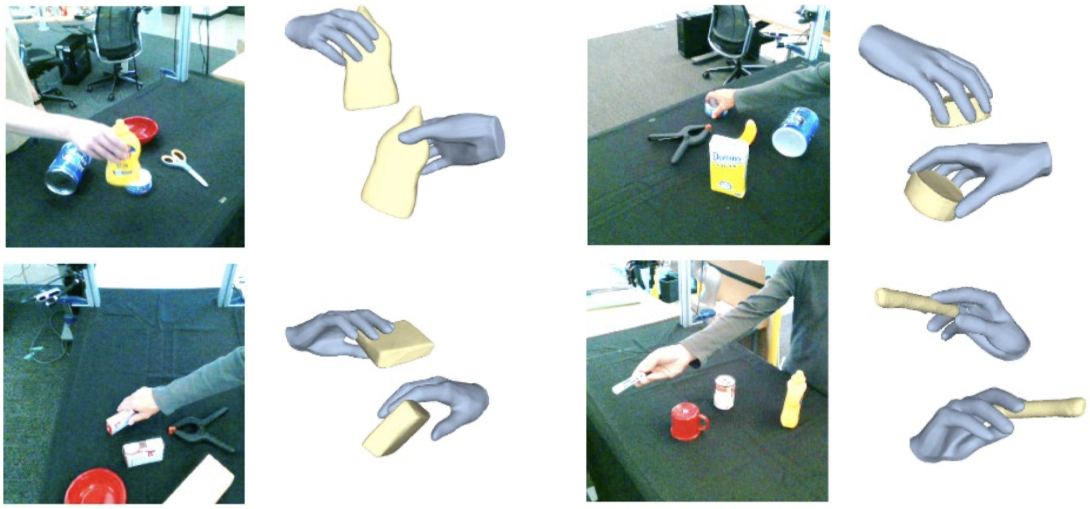


Figure 6: Qualitative results of our model on test images from the DexYCB benchmarks. Our model produces convincing results for different grasping poses and diverse objects.

challenging ObMan and DexYCB benchmarks and demonstrate significant improvements of the proposed method over the state of the art. Figure 6 presents some example results.

8.1.7 CoVR: Learning Composed Video Retrieval from Web Video Captions

Participants: Lucas Ventura, Antoine Yang, Cordelia Schmid, Gül Varol.

Composed Image Retrieval (CoIR) has recently gained popularity as a task that considers *both* text and image queries together, to search for relevant images in a database. Most CoIR approaches require manually annotated datasets, comprising image-text-image triplets, where the text describes a modification from the query image to the target image. However, manual curation of CoIR *triplets* is expensive and prevents scalability. In this work [47], we instead propose a scalable automatic dataset creation methodology that generates triplets given video-caption *pairs*, while also expanding the scope of the task to include composed *video* retrieval (CoVR) as seen in Figure 7. To this end, we mine paired videos with a similar caption from a large database, and leverage a large language model to generate the corresponding modification text. Applying this methodology to the extensive WebVid2M collection, we automatically construct our dataset, resulting in 1.6 million triplets. Moreover, we introduce a new benchmark for CoVR with a manually annotated evaluation set, along with baseline results. Our experiments further demonstrate that training a CoVR model on our dataset effectively transfers to CoIR, leading to improved state-of-the-art performance in the zero-shot setup on both the CIRR and FashionIQ benchmarks. Our code, datasets, and models are publicly available at the project webpage.

8.1.8 On the duality between contrastive and non-contrastive self-supervised learning

Participants: Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, Yann Lecun.

Recent approaches in self-supervised learning of image representations can be categorized into different families of methods and, in particular, can be divided into contrastive and non-contrastive approaches. While differences between the two families have been thoroughly discussed to motivate new

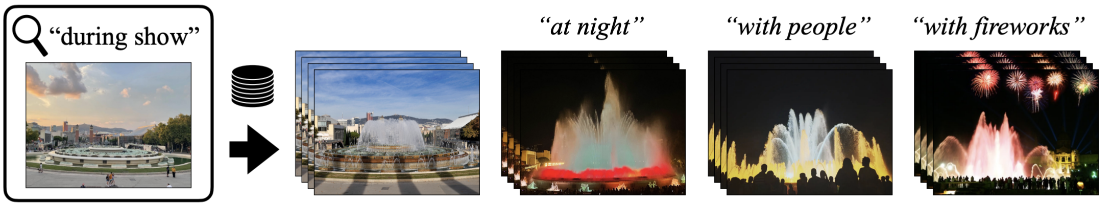


Figure 7: Task: Composed Video Retrieval (CoVR) seeks to retrieve *videos* from a database by searching with both a query image and a query text. The text typically specifies the desired modification to the query image. In this example, a traveller might wonder how the photographed place looks like during a fountain show, by describing several modifications, such as "during show at night, with fireworks".

approaches, in [18], we focus more on the theoretical similarities between them. By designing contrastive and covariance based non-contrastive criteria that can be related algebraically and shown to be equivalent under limited assumptions, we show how close those families can be. We further study popular methods and introduce variations of them, allowing us to relate this theoretical result to current practices and show the influence (or lack thereof) of design choices on downstream performance. Motivated by our equivalence result, we investigate the low performance of SimCLR and show how it can match VICReg's with careful hyperparameter tuning, improving significantly over known baselines. We also challenge the popular assumption that non-contrastive methods need large output dimensions. Our theoretical and quantitative results suggest that the numerical gaps between contrastive and non-contrastive methods in certain regimes can be closed given better network design choices and hyperparameter tuning. The evidence shows that unifying different SOTA methods is an important direction to build a better understanding of self-supervised learning.

8.1.9 Pixel-wise Agricultural Image Time Series Classification: Comparisons and a Deformable Prototype-based Approach

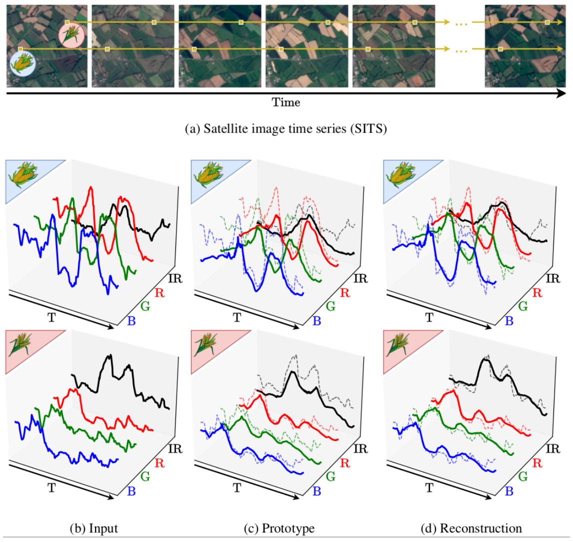
Participants: Elliot Vincent, Jean Ponce, Mathieu Aubry.

Improvements in Earth observation by satellites allow for imagery of ever higher temporal and spatial resolution. Leveraging this data for agricultural monitoring is key for addressing environmental and economic challenges. Current methods for crop segmentation using temporal data either rely on annotated data or are heavily engineered to compensate the lack of supervision. In this work [48], we present and compare datasets and methods for both supervised and unsupervised pixel-wise segmentation of satellite image time series (SITS). We also introduce an approach to add invariance to spectral deformations and temporal shifts to classical prototype-based methods such as K-means and Nearest Centroid Classifier (NCC). Figure 8 summarizes this framework. We show this simple and highly interpretable method leads to meaningful results in both the supervised and unsupervised settings and significantly improves the state of the art for unsupervised classification of agricultural time series on four recent SITS datasets.

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

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

Direct imaging is an active research topic in astronomy for the detection and the characterization of young sub-stellar objects. The very high contrast between the host star and its companions makes



(a) Satellite image time series (SITS)

(b) Input

(c) Prototype

(d) Reconstruction

Figure 8: Reconstructing pixel sequences from satellite image time series (SITS) through learned prototypes and transformations. Given a SITS (a), we reconstruct pixel-wise multi-spectral sequences using learned prototypes and transformations. Here, we show the RGB and IR spectral intensities over time for a corn and a wheat pixel sequence (b), along with their corresponding prototype before (c) and after (d) transformation.

the observations particularly challenging. In this context, post-processing methods combining several images recorded with the pupil tracking mode of telescope are needed. In previous works, we have presented a data-driven algorithm, PACO, capturing locally the spatial correlations of the data with a multivariate Gaussian model. PACO delivers improved detection sensitivity and confidence, as well as more accurate astro-photometric estimates than the standard post-processing methods of the field. However, there is room for improvement due to the approximate fidelity of the PACO statistical model to the time evolving observations. In this paper [39], we propose to combine the statistical model of PACO with supervised deep learning. The data are first pre-processed with the PACO framework to improve the stationarity and the contrast. A convolutional neural network (CNN) is then trained in a supervised fashion to detect the residual signature of synthetic sources. Finally, the trained network delivers a detection map. The photometry of detected sources is estimated by a second CNN. Both models are trained from scratch with custom data augmentation strategies allowing the creation of large training sets from a single spatio-temporal dataset. We apply the proposed approach to several datasets from the infrared imager of the VLT/SPHERE instrument. Our results show that its detection stage performs significantly better than baseline methods of the fields (cADI, PCA), and leads to a contrast improvement up to half a magnitude compared to PACO. The characterization stage of the proposed method performs on average on par with or better than the comparative algorithms (PCA, PACO) for angular separation above 0.5 arcsec. A typical reduction of the absolute error of photometric estimation by a factor two is obtained for sources of contrast up to 10^{-6}. Closer to the star, PCA and PACO remain slightly better. For both stages, controlling the uncertainty and the data-dependence of the current models remains an important avenue for improvement.

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

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

This paper [44] introduces a novel approach to the fine alignment of images in a burst captured by a handheld camera. In contrast to traditional techniques that estimate two-dimensional transformations between frame pairs or rely on discrete correspondences, the proposed algorithm establishes dense correspondences by optimizing both the camera motion and surface depth and orientation at every pixel. This approach improves alignment, particularly in scenarios with parallax challenges. Extensive experiments with synthetic bursts featuring small and even tiny baselines demonstrate that it outperforms the best optical flow methods available today in this setting, without requiring any training. Beyond enhanced alignment, our method opens avenues for tasks beyond simple image restoration, such as depth estimation and 3D reconstruction, as supported by promising preliminary results. This positions our approach as a versatile tool for various burst image processing applications.

8.2 Learning embodied representations

8.2.1 Learning Video-Conditioned Policies for Unseen Manipulation Tasks

Participants: Elliot Chane-Sane, Cordelia Schmid, Ivan Laptev.

The ability to specify robot commands by a non-expert user is critical for building generalist agents capable of solving a large variety of tasks. One convenient way to specify the intended robot goal is by a video of a person demonstrating the target task. While prior work typically aims to imitate human demonstrations performed in robot environments, here we focus on a more realistic and challenging setup with demonstrations recorded in natural and diverse human environments. In this work [11], we propose *Video-conditioned Policy learning (ViP)*, a data-driven approach that maps human demonstrations of previously unseen tasks to robot manipulation skills. To this end, we learn our policy to generate appropriate actions given current scene observations and a video of the target task. To encourage generalization to new tasks, we avoid particular tasks during training and learn our policy from unlabelled robot trajectories and corresponding robot videos. Both robot and human videos in our framework are represented by video embeddings pre-trained for human action recognition. At test time we first translate human videos to robot videos in the common video embedding space, and then use resulting embeddings to condition our policies. Figure 9 illustrates our method. Our approach enables robot control by human demonstrations in a *zero-shot manner*, i.e., without using robot trajectories paired with human instructions during training. We validate our approach on a set of challenging multi-task robot manipulation environments and outperform state of the art. Our method also demonstrates excellent performance in a new challenging zero-shot setup where no paired data is used during training.

8.2.2 Contact models in robotics: a comparative analysis

Participants: Quentin Le Lidec, Wilson Jallet, Louis Montaut, Ivan Laptev, Cordelia Schmid, Justin Carpentier.

Physics simulation is ubiquitous in robotics. Whether in model-based approaches (*e.g.*, trajectory optimization), or model-free algorithms (*e.g.*, reinforcement learning), physics simulators are a central component of modern control pipelines in robotics. Over the past decades, several robotic simulators have been developed, each with dedicated contact modeling assumptions and algorithmic solutions. In this article [43], we survey the main contact models and the associated numerical methods commonly used in robotics for simulating advanced robot motions involving contact interactions. In particular, we recall the physical laws underlying contacts and friction (*i.e.*, Signorini condition, Coulomb's law, and the maximum dissipation principle), and how they are transcribed in current simulators, as illustrated in Figure 10. For each physics engine, we expose their inherent physical relaxations along with their limitations due to the numerical techniques employed. Based on our study, we propose theoretically grounded quantitative criteria on which we build benchmarks assessing both the physical and computational aspects of simulation. We support our work with an open-source and efficient C++ implementation of the existing algorithmic variations. Our results demonstrate that some approximations or algorithms commonly used in robotics can severely widen the reality gap and impact target applications. We hope this work will help motivate the development of new contact models, contact solvers, and robotic simulators in general, at the root of recent progress in motion generation in robotics.

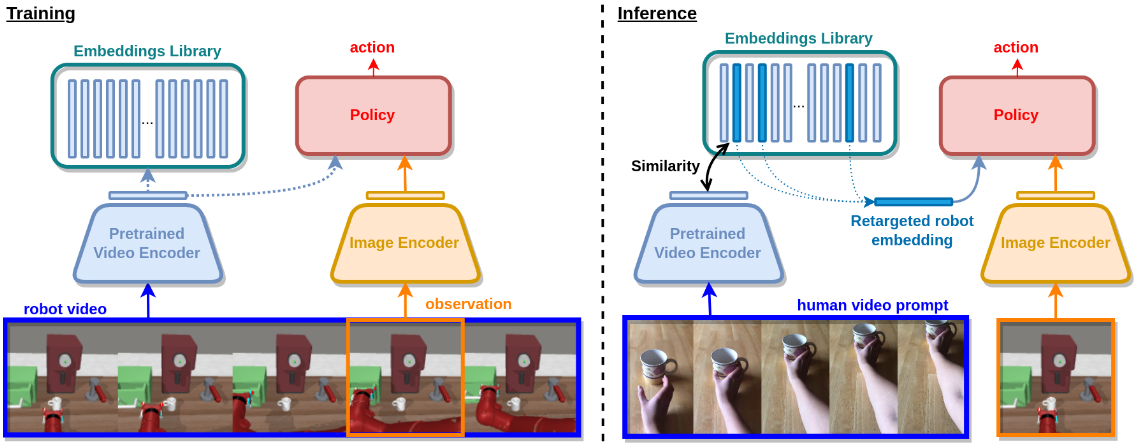


Figure 9: (Left) During training, we learn a manipulation policy conditioned on robot video embeddings of the full robot trajectories from the robot dataset. At the same time, the robot video embedding of each trajectory in the robot dataset is added to an embeddings library. (Right) At inference, we encode the human video instruction into a human video embedding. We then average the robot embeddings from the library that have highest cosine similarity to the human embedding into a selected robot embedding. Finally, we execute the policy conditioned on this selected embedding.

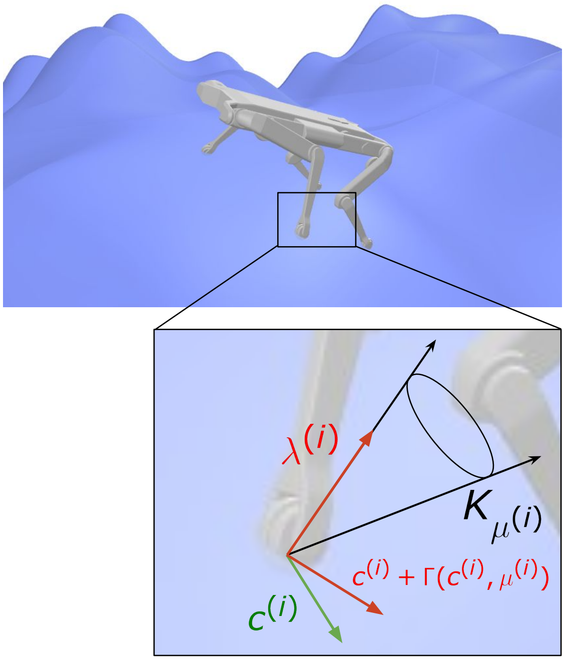


Figure 10: **Illustration of the dynamics of frictional contacts** between rigid bodies, which are governed by the Signorini condition, Coulomb's law, and the maximum dissipation principle. Combining these three principles leads to the Non-linear Complementarity Problem.

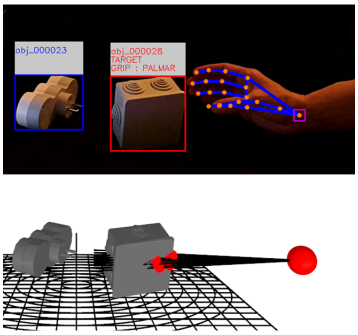


Figure 11: Visualization of the adaptive grasping assistance algorithm.

8.2.3 Vision-based interface for grasping intention detection and grip selection: towards intuitive upper-limb assistive devices

Participants: Etienne Moullet, François Bailly, Christine Azevedo Coste, Justin Carpentier.

Grasping is crucial for many daily activities, and its impairment considerably impacts quality of life and autonomy. Attempts to restore this function may rely on various approaches and devices (functional electrical stimulation, exoskeletons, prosthesis...) with command modalities often exerting considerable cognitive loads on users and lacking controllability and intuitiveness in daily life. In this work [22], we propose a novel user interface for grasping movement control in which the user delegates the grasping task decisions to the device, only requiring the user to move their (potentially prosthetic) hand toward the targeted object. Applying hand and object pose estimation to a live video feed of hands and objects, we analyse the kinematics of a movement to predict its target and estimate its impact zone to select the most appropriate grip.

8.2.4 Leveraging Proximal Optimization for Differentiating Optimal Control Solvers

Participants: Oumayma Bounou, Jean Ponce, Justin Carpentier.

Over the past few years, differentiable optimization has gained in maturity and attractivity within both machine learning and robotics communities. It consists in computing the derivatives of a given optimization problem which can then be used by learning algorithms, and enables to generically plug computational blocks reflecting the solving of generic mathematical programming problems into a learning pipeline. Until now, dedicated approaches have been proposed to compute the derivatives of various types of optimization problems (LPs, QPs, SOCPs, etc.). However, these approaches assume the problems are well-posed (e.g., satisfaction of the linearly independent constraint qualifications), limiting de facto their application to ill-posed problems. In this work [9], we focus on the differentiation of optimal control solvers widely used in robotics. We notably introduce a differentiable proximal formulation for solving equality-constrained LQR problems that is effective in solving ill-posed and rank-deficient problems accurately. Importantly, we show that this proximal formulation allows us to compute accurate gradients even in the case of ill-posed problems which do not satisfy the classical constraints qualification. Because any optimal control problem can be casted as an equalityconstrained LQR problem in the vicinity of the optimal solution, ours robust LQR derivatives computation can then be exploited to obtain the derivatives of general optimal control problems. We demonstrate the effectiveness of our approach in dynamics learning and system parameters identification experiments in linear optimal control problems (see Fig. 12).

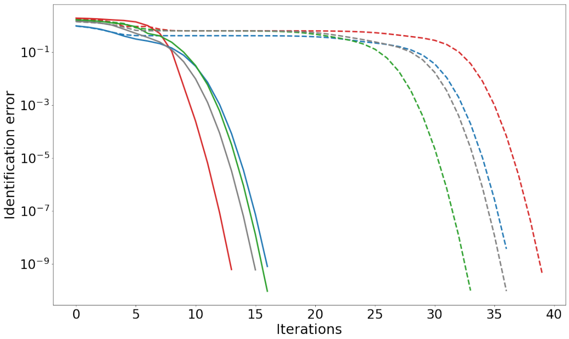


Figure 12: **Identification error** on identifying the dynamics matrices of a linear system. Pairs of curves with the same colors are identification experiments on the same problem parameters solved using different solvers: diff-mpc in dashed lines and ours in solid lines.

8.2.5 PROXDDP: Proximal Constrained Trajectory Optimization

Participants: Wilson Jallet, Antoine Bambade, Etienne Arlaud, Sarah El-Kazdadi, Nicolas Mansard, Justin Carpentier.

Trajectory optimization (TO) has proven, over the last decade, to be a versatile and effective framework for robot control. Several numerical solvers have been demonstrated to be fast enough to allow recomputing full-dynamics trajectories for various systems at control time, enabling model predictive control (MPC) of complex robots. These first implementations of MPC in robotics predominantly utilize some differential dynamic programming (DDP) variant for its computational speed and ease of use in constraint-free settings. Nevertheless, many scenarios in robotics call for adding hard constraints in TO problems (e.g., torque limits, obstacle avoidance), which existing solvers, based on DDP, often struggle to handle. Effectively addressing path constraints still poses optimization challenges (e.g., numerical stability, efficiency, accuracy of constraint satisfaction) that we propose to solve by combining advances in numerical optimization with the foundational efficiency of DDP. In this article [40], we leverage proximal methods for constrained optimization and introduce a DDP-like method to achieve fast, constrained trajectory optimization with an efficient warm-starting strategy particularly suited for MPC applications. Compared to earlier solvers, our approach effectively manages hard constraints without warm-start limitations and exhibits commendable convergence accuracy. Additionally, we leverage the computational efficiency of DDP, enabling real-time resolution of complex problems such as whole-body quadruped locomotion. We provide a complete implementation as part of an open-source and flexible C++ trajectory optimization library called ALIGATOR. These algorithmic contributions are validated through several trajectory planning scenarios from the robotics literature and the real-time whole-body MPC of a quadruped robot.

8.2.6 PROXQP: an Efficient and Versatile Quadratic Programming Solver for Real-Time Robotics Applications and Beyond

Participants: Antoine Bambade, Fabian Schramm, Sarah El Kazdadi, Stéphane Caron, Adrien Taylor, Justin Carpentier.

Convex Quadratic programming (QP) has become a core component in the modern engineering toolkit, particularly in robotics, where QP problems are legions, ranging from real-time whole-body controllers to planning and estimation algorithms. Many of those QPs need to be solved at high frequency. Meeting timing requirements requires taking advantage of as many structural properties as possible for the problem at hand. For instance, it is generally crucial to resort to warm-starting to exploit the resemblance of consecutive control iterations. While a large range of off-the-shelf QP solvers is available, only a few are suited to exploit problem structure and warm-starting capacities adequately. In this work [36], we propose the PROXQP algorithm, a new and efficient QP solver that exploits QP structures by leveraging primal-dual augmented Lagrangian techniques. For convex QPs, PROXQP features a global convergence guarantee to the closest feasible QP, an essential property for safe closedloop control. We illustrate its practical performance on various standard robotic and control experiments, including a real-world closed-loop model predictive control application. While originally tailored for robotics applications, we show that PROXQP also performs at the level of state of the art on generic QP problems, making PROXQP suitable for use as an off-the-shelf solver for regular applications beyond robotics.

8.2.7 QPLayer: efficient differentiation of convex quadratic optimization

Participants: Antoine Bambade, Fabian Schramm, Adrien Taylor, Justin Carpentier.

Optimization layers within neural network architectures have become increasingly popular for their ability to solve a wide range of machine learning tasks and to model domain-specific knowledge. However, designing optimization layers requires careful consideration as the underlying optimization problems might be infeasible during training. Motivated by applications in learning, control, and robotics, this work [37] focuses on convex quadratic programming (QP) layers. The specific structure of this type of optimization layer can be efficiently exploited for faster computations while still allowing rich modeling capabilities. We leverage primal-dual augmented Lagrangian techniques for computing derivatives of both feasible and infeasible QPs. Not requiring feasibility allows, as a byproduct, for more flexibility in the QP to be learned. The effectiveness of our approach is demonstrated in a few standard learning experiments, obtaining three to ten times faster computations than alternative state-of-the-art methods while being more accurate and numerically robust. Along with these contributions, we provide an open-source C++ software package called QPLayer for efficiently differentiating convex QPs and which can be interfaced with modern learning frameworks.

8.2.8 Stagewise Implementations of Sequential Quadratic Programming for Model-Predictive Control

Participants: Armand Jordana, Sébastien Kleff, Avadesh Meduri, Justin Carpentier, Nicolas Mansard, Ludovic Righetti.

The promise of model-predictive control in robotics has led to extensive development of efficient numerical optimal control solvers in line with differential dynamic programming because it exploits the sparsity induced by time. In this work [41], we argue that this effervescence has hidden the fact that sparsity can be equally exploited by standard nonlinear optimization. In particular, we show how a tailored implementation of sequential quadratic programming achieves state-of-the-art model-predictive control. Then, we clarify the connections between popular algorithms from the robotics community and well-established optimization techniques. Further, the sequential quadratic program formulation naturally encompasses the constrained case, a notoriously difficult problem in the robotics community. Specifically, we show that it only requires a sparsity-exploiting implementation of a state-of-theart quadratic programming solver. We illustrate the validity of this approach in a comparative study and experiments on a torque-controlled manipulator. To the best of our knowledge, this is the first demonstration of nonlinear model-predictive control with arbitrary constraints on real hardware.

8.2.9 Risk-Sensitive Extended Kalman Filter

Participants: Armand Jordana, Avadesh Meduri, Etienne Arlaud, Justin Carpentier, Ludovic Righetti.

In robotics, designing robust algorithms in the face of estimation uncertainty is a challenging task. Indeed, controllers often do not consider the estimation uncertainty and only rely on the most likely estimated state. Consequently, sudden changes in the environment or the robot's dynamics can lead to catastrophic behaviors. In this work [42], we present a risk-sensitive Extended Kalman Filter that allows doing outputfeedback Model Predictive Control (MPC) safely. This filter adapts its estimation to the control objective. By taking a pessimistic estimate concerning the value function resulting from the MPC controller, the filter provides increased robustness to the controller in phases of uncertainty as compared to a standard Extended Kalman Filter (EKF). Moreover, the filter has the same complexity as an EKF, so that it can be used for real-time model-predictive control. The paper evaluates the risk-sensitive behavior of the proposed filter when used in a nonlinear model-predictive control loop on a planar drone and industrial manipulator in simulation, as well as on an external force estimation task on a real quadruped robot. These experiments demonstrate the abilities of the approach to improve performance in the face of uncertainties significantly.

8.2.10 PolarNet: 3D Point Clouds for Language-Guided Robotic Manipulation

Participants: Shizhe Chen, Ricardo Garcia, Cordelia Schmid, Ivan Laptev.

The ability for robots to comprehend and execute manipulation tasks based on natural language instructions is a long-term goal in robotics. The dominant approaches for language-guided manipulation use 2D image representations, which face difficulties in combining multi-view cameras and inferring precise 3D positions and relationships as shown in Figure 13b. To address these limitations, in this work [13], we propose a 3D point cloud based policy called PolarNet for language-guided manipulation. It leverages carefully designed point cloud inputs (see Figure 13c), efficient point cloud encoders, and multimodal transformers to learn 3D point cloud representations and integrate them with language instructions for action prediction. PolarNet is shown to be effective and data efficient in a variety of experiments conducted on the RLBench benchmark. It outperforms state-of-the-art 2D and 3D approaches in both single-task and multi-task learning. It also achieves promising results on a real robot. Code, trained models and videos are available at the project website.

8.2.11 Object Goal Navigation with Recursive Implicit Maps

Participants: Shizhe Chen, Thomas Chabal, Ivan Laptev, Cordelia Schmid.

Object goal navigation aims to navigate an agent to locations of a given object category in unseen environments. Classical methods explicitly build maps of environments and require extensive engineering while lacking semantic information for object-oriented exploration. On the other hand, end-to-end learning methods alleviate manual map design and predict actions using implicit representations. Such methods, however, lack an explicit notion of geometry and may have limited ability to encode navigation history. In this work [12], we propose an implicit spatial map for object goal navigation. Our implicit map is recursively updated with new observations at each step using a transformer. To encourage spatial reasoning, we introduce auxiliary tasks and train our model to reconstruct explicit maps as well as to predict visual features, semantic labels and actions. Figure 14 illustrates the proposed framework and auxiliary training tasks. Our method significantly outperforms the state of the art on the challenging MP3D dataset and generalizes well to the HM3D dataset. We successfully deploy our model on a real robot and achieve encouraging object goal navigation results in real scenes using only a few real-world demonstrations. Code, trained models and videos are available at the project website.

8.2.12 GJK++: Leveraging Acceleration Methods for Faster Collision Detection

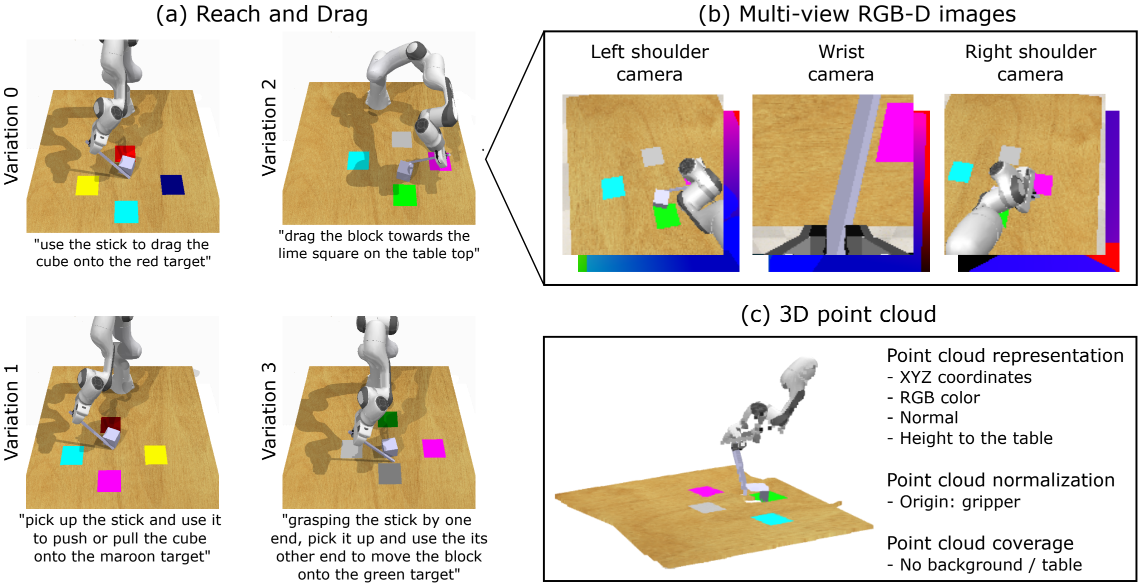

(a) Reach and Drag
Variation 0
"use the stick to drag the cube onto the red target"
Variation 2
"drag the block towards the lime square on the table top"
Variation 1
"pick up the stick and use it to push or pull the cube onto the maroon target"
Variation 3
"grasping the stick by one end, pick it up and use the its other end to move the block onto the green target"
(b) Multi-view RGB-D images
Left shoulder camera
Wrist camera
Right shoulder camera
(c) 3D point cloud
Point cloud representation
- XYZ coordinates
- RGB color
- Normal
- Height to the table
Point cloud normalization
- Origin: gripper
Point cloud coverage
- No background / table

Figure 13: (a): Variations of the "Reach and Drap" task with difference target colors per variation. (b): Although different views are complementary to present the scene, they are not explicitly aligned with each other. (c) Merging multi-view cameras to construct a unified point cloud in 3D space with optimized design choices.

Participants: Louis Montaut, Quentin Le Lidec, Vladimir Petrik, Josef Sivic, Justin Carpentier.

Collision detection is a fundamental problem in various domains, such as robotics, computational physics, and computer graphics. In general, collision detection is tackled as a computational geometry problem, with the so-called Gilbert, Johnson, and Keerthi (GJK) algorithm being the most adopted solution nowadays. While introduced in 1988, GJK remains the most effective solution to compute the distance or the collision between two 3D convex geometries. Over the years, it was shown to be efficient, scalable, and generic, operating on a broad class of convex shapes, ranging from simple primitives (sphere, ellipsoid, box, cone, capsule, etc.) to complex meshes involving thousands of vertices. In this article [46], we introduce several contributions to accelerate collision detection and distance computation between convex geometries by leveraging the fact that these two problems are fundamentally optimization problems. Notably, we establish that the GJK algorithm is a specific sub-case of the well-established Frank-Wolfe (FW) algorithm in convex optimization. By adapting recent works linking Polyak and Nesterov accelerations to Frank-Wolfe methods, we also propose two accelerated extensions of the classic GJK algorithm: the Polyak and Nesterov accelerated GJK algorithms. Through an extensive benchmark over millions of collision pairs involving objects of daily life, we show that these two accelerated GJK extensions significantly reduce the overall computational burden of collision detection, leading to computation times that are up to two times faster. Finally, we hope this work will significantly reduce the computational cost of modern robotic simulators, allowing the speed-up of modern robotic applications that heavily rely on simulation, such as reinforcement learning or trajectory optimization.

8.2.13 Robust Visual Sim-to-Real Transfer for Robotic Manipulation

Participants: Ricardo Garcia, Robin Strudel, Shizhe Chen, Etienne Arlaud, Ivan Laptev, Cordelia Schmid.

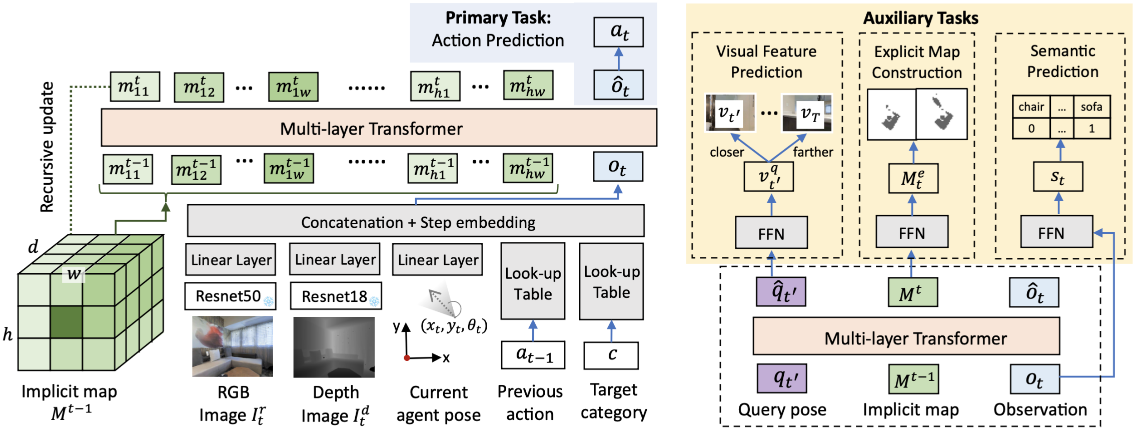


Figure 14: An overview of our ObjectNav model with a recursive implicit map (RIM). It encodes observations at step t into o_t, and uses a multi-layer transformer to recursively update the implicit spatial map M^t. Besides the main task of action prediction, three auxiliary tasks are proposed to improve spatial reasoning of M^t and semantic understanding. The Multi-layer Transformer on the right is the same as on the left, while we add three feed-forward networks (FFN) for auxiliary tasks.

In this work [17], we systematically explore visual domain randomization methods and benchmark them on a rich set of challenging robotic manipulation tasks. In particular, we propose an off-line proxy task of cube localization to select DR parameters for texture randomization, lighting randomization,variations of object colors and camera parameters. Notably, we demonstrate that DR parameters have similar impact on our off-line proxy task and on-line policies. As illustrated in Figure 16, we use off-line optimized DR parameters to train visuomotor policies in simulation and directly apply such policies to a real robot. Our approach achieves 93% success rate on average when tested on a diverse set of challenging manipulation tasks. Moreover, we evaluate the robustness of policies to visual variations in real scenes and show that our simulator-trained policies outperform policies learned using real but limited data. Code, simulation environment, real robot datasets and trained models are available at the project website.

8.2.14 Learning Reward Functions for Robotic Manipulation by Observing Humans

Participants: Minttu Alakuijala, Gabriel Dulac-Arnold, Julien Mairal, Jean Ponce, Cordelia Schmid.

Observing a human demonstrator manipulate objects provides a rich, scalable and inexpensive source of data for learning robotic policies. However, transferring skills from human videos to a robotic manipulator poses several challenges, not least a difference in action and observation spaces. In this work [8], we use unlabeled videos of humans solving a wide range of manipulation tasks to learn a task-agnostic reward function for robotic manipulation policies. Thanks to the diversity of this training data (see Figure 17), the learned reward function sufficiently generalizes to image observations from a previously unseen robot embodiment and environment to provide a meaningful prior for directed exploration in reinforcement learning. We propose two methods for scoring states relative to a goal image: through direct temporal regression, and through distances in an embedding space obtained with time-contrastive learning. By conditioning the function on a goal image, we are able to reuse one model across a variety of tasks. Unlike prior work on leveraging human videos to teach robots, our method, Human Offline Learned Distances (HOLD) requires neither a priori data from the robot environment, nor a set of task-specific human demonstrations, nor a predefined notion of correspondence across morphologies, yet it is able to accelerate training of several manipulation tasks on a simulated robot arm compared to using only a sparse reward obtained from task completion.

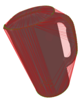
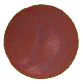
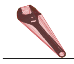
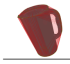
		$N = 240$			$N = 1811$			$N = 3585$		
										
		GJK	**Polyak**	**Nesterov**	**GJK**	**Polyak**	**Nesterov**	**GJK**	**Polyak**	**Nesterov**
	T_D^μ	1.1 ± 0.3	$\mathbf{0.9 \pm 0.3}$	$\mathbf{0.9 \pm 0.3}$	1.9 ± 0.5	1.5 ± 0.5	$\mathbf{1.4 \pm 0.5}$	3.2 ± 0.8	$\mathbf{2.3 \pm 0.7}$	2.4 ± 0.7
	T_C^μ	0.9 ± 0.4	$\mathbf{0.7 \pm 0.4}$	0.8 ± 0.4	1.5 ± 0.6	1.2 ± 0.6	$\mathbf{1.1 \pm 0.6}$	2.5 ± 0.9	$\mathbf{1.7 \pm 0.8}$	1.9 ± 1.0
	T_D^μ				2.7 ± 0.8	2.1 ± 0.7	$\mathbf{1.9 \pm 0.6}$	4.0 ± 1.0	$\mathbf{2.9 \pm 0.9}$	2.9 ± 1.0
	T_C^μ				2.3 ± 0.8	1.6 ± 0.8	$\mathbf{1.5 \pm 0.8}$	3.1 ± 1.2	$\mathbf{2.1 \pm 1.0}$	2.2 ± 1.3
	T_D^μ							4.4 ± 1.3	$\mathbf{2.9 \pm 1.0}$	3.0 ± 0.9
	T_C^μ							3.0 ± 1.9	$\mathbf{2.1 \pm 1.4}$	$\mathbf{2.1 \pm 1.4}$

Figure 15: Computation times on collision detection problems of our methods, Polyak and Nesterov accelerated GJK algorithms, and the state-of-the-art GJK algorithm. We report the average runtime (in μs) for distance computation (T_D^μ) and boolean collision checking (T_C^μ) on the ycb benchmark dataset for close-proximity or shallowly intersecting shapes. N denotes the number of vertices for each mesh.

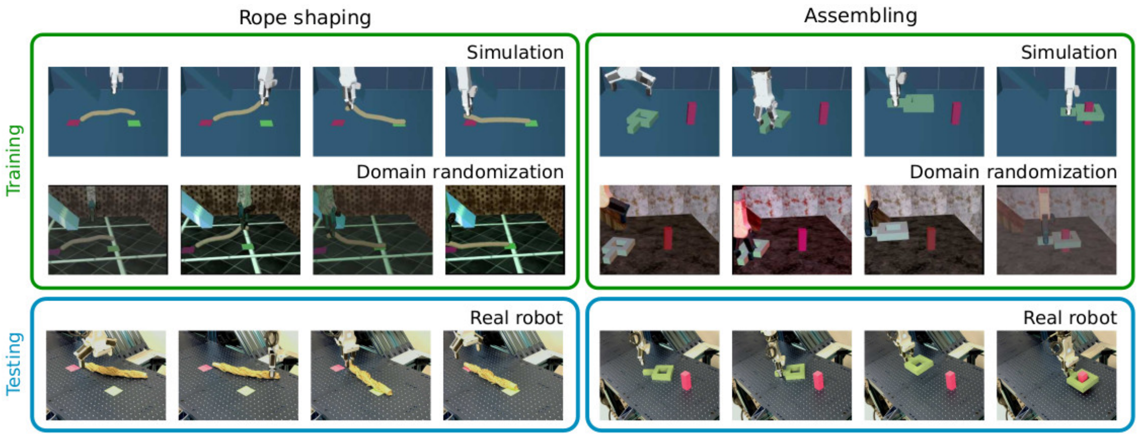
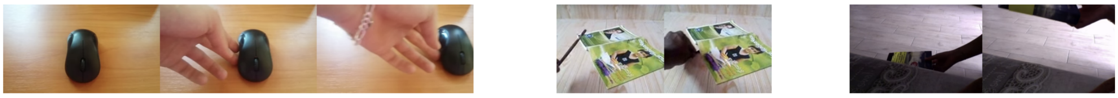


Figure 16: **An overview of our approach.** We learn visuomotor manipulation policies in simulation (row 1) with domain randomization by sampling high-quality textures, lighting, object colors and camera parameters (row 2). We analyze different sampling options and demonstrate that simulator-trained policies can be directly deployed on a real robot for diverse and challenging manipulation tasks (row 3), such as rope-shaping (left) and assembling (right).



(a) Pushing mouse from left to right

(b) Putting paint brush underneath magazine

(c) Moving book up

Figure 17: Example human videos from Something-Something v2 used to train the distance models.

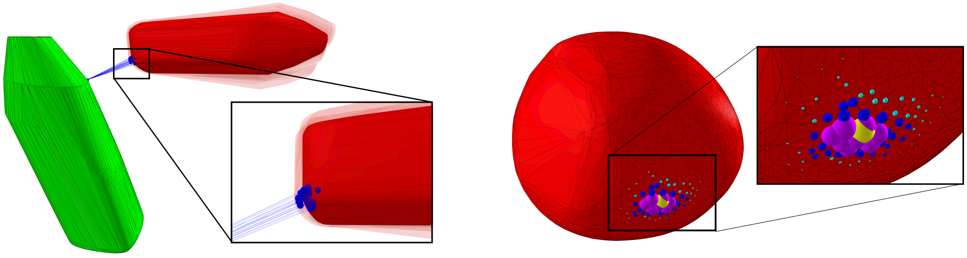


Figure 18: Illustration of randomized smoothing approximation on meshes from the YCB dataset. Left: 0^{th} order estimator, using a Gaussian distribution with 25 samples. Right: 1^{st} order estimator using a Gumbel distribution.

8.2.15 Differentiable Collision Detection: A Randomized Smoothing Approach

Participants: Louis Montaut, Quentin Le Lidec, Antoine Bambade, Vladimir Petrik, Josef Sivic, Justin Carpentier.

Collision detection appears as a canonical operation in a large range of robotics applications from robot control to simulation, including motion planning and estimation. While the seminal works on the topic date back to the 80's, it is only recently that the question of properly differentiating collision detection has emerged as a central issue, thanks notably to the ongoing and various efforts made by the scientific community around the topic of differentiable physics. Yet, very few solutions have been suggested so far, and only with a strong assumption on the nature of the shapes involved. In our work [45], we introduce a generic and efficient approach to compute the derivatives of collision detection for *any* pair of convex shapes, by notably leveraging randomized smoothing techniques which have shown to be particularly adapted to capture the derivatives of non-smooth problems. This approach is implemented in the HPP-FCL and Pinocchio ecosystems, and evaluated on classic datasets and problems of the robotics literature, demonstrating few micro-second timings to compute informative derivatives directly exploitable by many real robotic applications including differentiable simulation.

8.2.16 Multi-Contact Task and Motion Planning Guided by Video Demonstration

Participants: Kateryna Zorina, David Kovar, Florent Lamiraux, Nicolas Mansard, Justin Carpentier, Josef Sivic, Vladimír Petrík.

In our work [25], we aim to leverage instructional video to guide the solving of complex multi-contact task-and-motion planning tasks in robotics. Towards this goal, we propose an extension of the well-established Rapidly-Exploring Random Tree (RRT) planner, which simultaneously grows multiple trees around grasp and release states extracted from the guiding video. Our key novelty lies in combining contact states, and 3D object poses extracted from the guiding video with a traditional planning algorithm that allows us to solve tasks with sequential dependencies, for example, if an object needs to be placed at a specific location to be grasped later. To demonstrate the benefits of the proposed video-guided planning approach, we design a new benchmark with three challenging tasks: (i) 3D rearrangement of multiple objects between a table and a shelf, (ii) multi-contact transfer of an object through a tunnel, and (iii) transferring objects using a tray in a similar way a waiter transfers dishes. We demonstrate the effectiveness of our planning algorithm on several robots, including the Franka Emika Panda and the KUKA KMR iiwa.

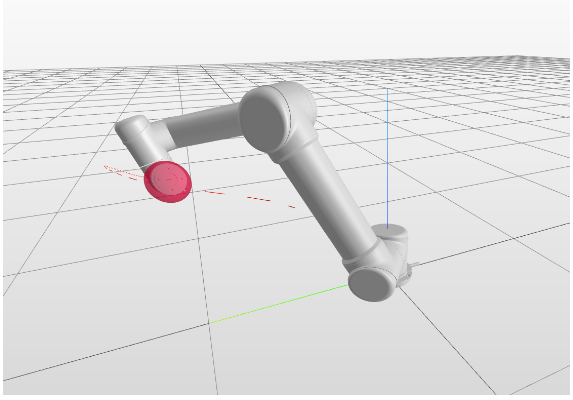


Figure 19: After training, a rollout of the policy leads to precise control on a test problem.

8.2.17 Enforcing the consensus between Trajectory Optimization and Policy Learning for precise robot control

Participants: Quentin Le Lidec, Wilson Jallet, Ivan Laptev, Cordelia Schmid, Justin Carpentier.

Reinforcement learning (RL) and trajectory optimization (TO) present strong complementary advantages. On one hand, RL approaches are able to learn global control policies directly from data, but generally require large sample sizes to properly converge towards feasible policies. On the other hand, TO methods are able to exploit gradient-based information extracted from simulators to quickly converge towards a locally optimal control trajectory which is only valid within the vicinity of the solution. Over the past decade, several approaches have aimed to adequately combine the two classes of methods in order to obtain the best of both worlds. Following on from this line of research, our work [19] proposes several improvements on top of these approaches to learn global control policies quicker, notably by leveraging sensitivity information stemming from TO methods via Sobolev learning, and augmented Lagrangian techniques to enforce the consensus between TO and policy learning. We evaluate the benefits of these improvements on various classical tasks in robotics through comparison with existing approaches in the literature.

8.2.18 Talking about moving machines

Participants: Céline Pieters, Emmanuelle Danblon, Philippe Souères, Jean-Paul Laumond.

Globally, robots can be described as some sets of moving parts that are dedicated to a task while using their own energy. Yet, humans commonly qualify those machines as being intelligent, autonomous or being able to learn, know, feel, make decisions, etc. Is it merely a way of talking or does it mean that robots could eventually be more than a complex set of moving parts? On the one hand, the language of robotics allows multiple interpretations (leading sometimes to misreading or confusion in various contexts). On the other hand, the status of robots is challenged more and more by technical achievements and humans' own empirical beliefs. In this paper [7], we follow a linguistic approach in order to explore the relevance of these words when talking about robots. Since we note that the words impose themselves (even if opposed), we discuss the efficiency of a rhetorical strategy in order to work with such a lexicon in robotics. More precisely, we explore the argumentative technique of the dissociation of notions through the study of a practical case: the case of robot lawn mowers versus hedgehogs.

8.2.19 Co-designing versatile quadruped robots for dynamic and energy-efficient motions

Participants: Gabriele Fadini, Shivesh Kumar, Rohit Kumar, Thomas Flayols, Andrea Del Prete, Justin Carpentier, Philippe Souères.

This paper [38] presents a bi-level optimization framework to concurrently optimize a quadruped hardware and control policies for achieving dynamic cyclic behaviors. The longterm vision to drive the design of dynamic and efficient robots by means of computational techniques is applied to improve the development of a new quadruped prototype. The scale of the robot and its actuators are optimized for energy efficiency considering a complete model of the motor, that includes friction, torque, and bandwidth limitations. This model is used to optimize the power consumption during bounding and backflip tasks and is validated by tracking the output trajectories on the first prototype iteration. The co-design results show an improvement of up to 87% for a single task optimization. It appears that, for jumping forward, robots with longer thighs perform better, while for backflips, longer shanks are better suited. To understand the trade-off between these different choices, a Pareto set is constructed to guide the design of the next prototype.

8.3 Image restoration and enhancement

8.3.1 An Image Quality Assessment Dataset for Portraits

Participants: Nicolas Chahine, Stefania Calarasanu, Davide Garcia-Civiero, Théo Cayla, Sira Ferradans, Jean Ponce.

In the field of smartphone photography, particularly portrait photography, the demand for superior image quality assessment (IQA) is paramount. In this work [10], we introduce the "PIQ23" dataset, specifically curated for advancing IQA in portrait photography. This dataset comprises 5116 images across 50 distinct scenarios, captured with 100 varied smartphone models. It includes portraits of individuals from diverse genders and ethnicities, with their explicit consent for public research use. Examples are presented in Figure 20. Over 30 image quality experts annotated the dataset, focusing on face detail preservation, face target exposure, and overall image quality. This annotation process was rigorously analyzed to ensure consistency, a crucial aspect often lacking in mean opinion score (MOS) methods. Our findings also highlight the significant role of semantic information in enhancing IQA predictions. The dataset, along with our detailed statistical analysis and BIQA algorithms, is available here.

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

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

Exoplanet detection by direct imaging is a difficult task: the faint signals from the objects of interest are buried under a spatially structured nuisance component induced by the host star. The exoplanet signals can only be identified when combining several observations with dedicated detection algorithms. In contrast to most of existing methods, in [15], we propose to learn a model of the spatial, temporal and spectral characteristics of the nuisance, directly from the observations. In a pre-processing step, a statistical model of their correlations is built locally, and the data are centered and whitened to improve both their stationarity and signal-to-noise ratio (SNR). A convolutional neural network (CNN) is then trained in a supervised fashion to detect the residual signature of synthetic sources in the preprocessed images. Our method leads to a better trade-off between precision and recall than standard approaches in the field. It also outperforms a state-of-the-art algorithm based solely on a statistical framework. Besides, the exploitation of the spectral diversity improves the performance compared to a similar model built solely from spatio-temporal data.

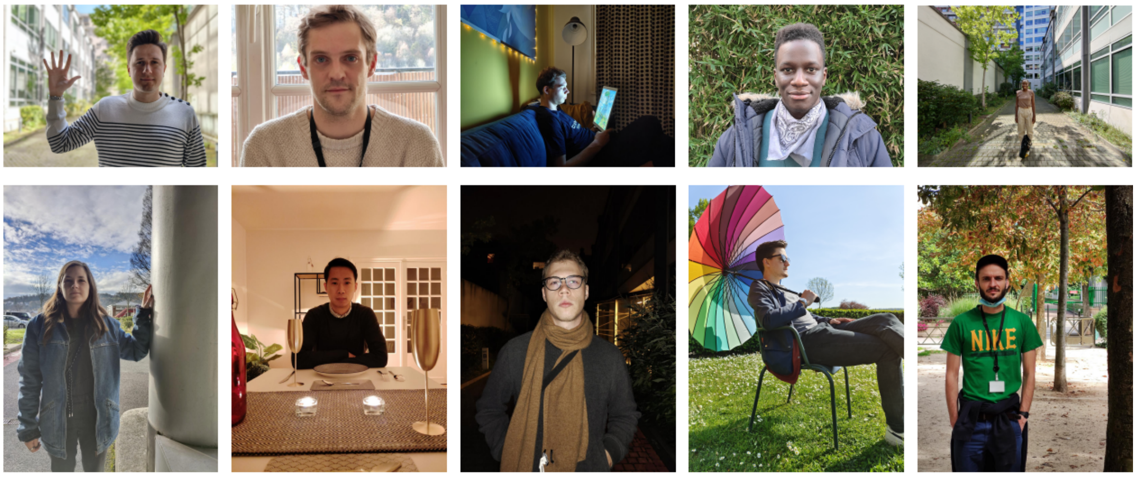


Figure 20: A selection of images from the PIQ23 dataset, demonstrating the range of portrait scenarios and smartphone models used in our study.

8.4 Doctoral dissertations and habilitation theses

8.4.1 Unsupervised Learning in Complex Systems

Participants: Hugo Cisneros.

In this thesis [28], we explore the use of complex systems to study learning and adaptation in natural and artificial systems. The goal is to develop autonomous systems that can learn without supervision, develop on their own, and become increasingly complex over time. Complex systems are identified as a suitable framework for understanding these phenomena due to their ability to exhibit growth of complexity. Being able to build learning algorithms that require limited to no supervision would enable greater flexibility and adaptability in various applications. By understanding the fundamental principles of learning in complex systems, we hope to advance our ability to design and implement practical learning algorithms in the future. This thesis makes the following key contributions: the development of a general complexity metric that we apply to search for complex systems that exhibit growth of complexity, the introduction of a coarse-graining method to study computations in large-scale complex systems, and the development of a metric for learning efficiency as well as a benchmark dataset for evaluating the speed of learning algorithms. Our findings add substantially to our understanding of learning and adaptation in natural and artificial systems. Moreover, our approach contributes to a promising new direction for research in this area. We hope these findings will inspire the development of more effective and efficient learning algorithms in the future.

8.4.2 Language-guided navigation and manipulation in robotics using transformers

Participants: Pierre-Louis Guhur.

Recent progress in machine learning has enabled groundbreaking improvements notably in computer vision, natural language processing, and robotics. Can we go one step further and combine these research fields? This would allow new applications, such as language-guided robotics, where a robot must follow instructions provided by an operator. While people learn to follow natural language instructions from their childhood, the same task is difficult for robots. Current challenges include (i) the limited amount of training data, (ii) the multiple levels of reasoning, and (iii) the multi-dimensional continuous action space. The goal of this thesis [29] is to improve language-guided robotics by addressing these challenges. We break down the difficulty of language-guided robotics by considering two types of tasks: (i) vision-and-language navigation, where a mobile robot must go to a target location, and (ii) vision-and-language manipulation, where a robotic arm should manipulate objects on a tabletop. Our contributions are the following: (i) we address the scarcity of training data and develop an efficient pre-training procedure based on the new BnB dataset, (ii) we propose a hierarchical approach based on the Transformer architecture to encode several layers of abstractions, and (iii) we propose a new method predicting continuous and multi-dimensional actions for solving a large number of robotics tasks on a tabletop. Methods developed in this thesis have been tested in photo-realistic simulators and on a real-world robot. They have outperformed the state-of-the-art performance on a dozen of benchmarks.

8.4.3 Pose estimation of rigid objects and robots

Participants: Yann Labbé.

The goal of this thesis [30] is to develop methods for recovering the 3D configuration of scenes containing rigid objects and articulated robots with known 3D models using one or multiple RGB images as inputs. We consider the following challenging scenes and visual conditions: (i) textureless and/or symmetric objects (ii) robot arms with several degrees of freedom, (iii) scenes imaged under challenging conditions (e.g. viewpoint or illumination) and (iv) objects or robots partially occluded. The key contributions of this thesis are as follows. First, we introduce a method for identifying a variable number of objects in a robot's workspace and estimate the 2D coordinate of the object's centroids in the robot coordinate frame. Our approach does not require extrinsic camera-to-robot calibration. Second, we propose a method for efficiently solving the planar rearrangement planning problem. We propose a discrete action parametrization of this problem, and efficiently apply Monte-Carlo Tree Search (MCTS) to solve it. Third, we introduce a novel learning-based method for 6D pose estimation of rigid objects with known 3D models. Our approach relies on the render-and-compare strategy. We introduce innovations of the training loss and rotation parametrization to explicitly handle object symmetries and achieve stable training. We train our approach on synthetic data using heavy image augmentations and show the crucial importance of data augmentation for the trans- fer to real scenes. Fourth, we introduce an approach for multi-view multi-object 6D pose estimation. We introduce a novel object-level RANSAC strategy to jointly estimate relative camera poses and find correspondences between single-view pose hypotheses. Poses of all objects and cameras are jointly refined by solving an object-level bundle adjustment problem. Fifth, we develop an approach to estimate the pose of novel objects, i.e. objects unseen during training, but for which the 3D model is available at test time. We introduce a scoring network for finding the best initial estimate among a set of coarse hypotheses, and design a network for iterative refinement where the object shape and coordinate system are implicitly provided as inputs. The model is trained on a novel large-scale synthetic dataset displaying thousands of different objects in challenging visual conditions. Finally, we introduce a method for estimating the 6D pose and joint angles of an articulated robot. We extend the render-and-compare strategy to handle robots with several degrees of freedom. We show the crucial importance of robot parametrization in this problem, and propose an effective strategy that is independent of the robot. The methods presented in this thesis advance the state-of-the-art on existing datasets and benchmarks for object and robot pose estimation. For known rigid objects, our single-view approach CosyPose is the winning entry in the BOP Challenge 2020. Our approach for unseen objects, MegaPose, achieves similar performance while not requiring the objects to be known in advance for training, paving the way for real applications where rapid deployment is key.

8.4.4 Estimating 3D Motion and Forces from Monocular Videos

Participants: Zongmian Li.

In this thesis [32], we investigate the problem of automatically reconstructing the 3D dynamic scene depicting a person interacting with a tool in a single RGB video. The objective is to obtain a 3D interpretation of the scene represented by the 3D poses of the person and the manipulated object over time, the contact positions and the contact forces exerted on the human body. This problem is challenging because of occlusions, depth ambiguities and the thin, texture-less nature of the manipulated tools such as the spade or the hammer. The main contributions of this thesis are as follows. First, we introduce an approach to jointly estimate the motion and the actuation forces of the person on the manipulated object by modeling the contacts and the dynamics of the interactions. This is cast as a large-scale trajectory optimization problem by minimizing a set of loss functions integrated over time and summed over person joints and object keypoints. The problem is subject to several constraints based on the laws of physics, which include contact and friction models and the Lagrangian dynamics equation. Second, we develop a method to automatically recognize from the input video the 2D position and timing of contacts between the person and the object or the ground. Instead of modeling contact states as binary variables during optimization, we automatically recognize contacts in the input video using a convolutional neural network (CNN) trained from manually annotated contact data that combine both still images and videos harvested from the Internet, thereby significantly reducing the complexity of the optimization. Third, we validate our approach on a recent video-MoCap dataset capturing typical parkour actions and equipped with ground truth forces and trajectories. We also demonstrate the benefits of our approach on a new dataset of Internet videos showing people manipulating a variety of tools in unconstrained environments. The experiments show that our method improves results on both 3D human pose estimation and 2D object localization, and achieves reasonable force estimates on this data.

8.4.5 Learning representations for visually-guided robotics

Participants: Robin Strudel.

The goal of this thesis [34] is to develop models, representations and learning algorithms for the automatic acquisition of visually-guided robotic skills from demonstrations and for object localization. We first introduce a method to acquire robotic skills from demonstrations by learning a vocabulary of basic skills with behavioral cloning. Skills are then combined with a planning policy learned with reinforcement learning in order to perform more complex tasks. We show successful transfer of multiple tasks from simulation to a real robot by using a method developed in this thesis optimizing a sequence of data augmentations on synthetic data to solve a proxy object localization task on real data. We then focus on sensor-based motion planning and propose an approach leveraging the knowledge of surrounding obstacles observed with a camera to accelerate the finding of collision-free paths. The learned representation generalizes across a large variety of objects, and the planning policy can handle new environments with dynamically moving obstacles. While visually-guided policies learn task-centric image representations from control supervision, another line of work consists in learning object-centric representations that can be plugged into classical robotics methods. Object-centric approaches rely on a segmentation backbone for which we propose the following contributions. Towards this goal we propose a transformer-based semantic segmentation model that leverages global context of the image at every stage of the model and show state-of-the-art results when compared to convolution-based approaches on classical benchmarks. Our segmentation model presents two limitations, it localizes a pre-defined set objects and requires dense annotations to be trained, which limits its scalability to large datasets. To address these limitations, we propose a method that segments an open set of visual concepts defined by natural language and does not require pixel-level supervision. Our method learns to localize objects by using image-level labels such as the presence of an object in the image.

8.4.6 Learning Visual Language Models for Video Understanding

Participants: Antoine Yang.

The goal of this thesis [35] is to build and train machine learning models that combine the power of natural language processing with visual understanding, enabling a comprehensive and detailed comprehension of the content within videos. First, we propose two scalable approaches to develop video question answering models without the need for costly manual annotation. We automatically generate video question answering data from narrated videos using text-only question-generation models. We then show that a multi-modal transformer trained contrastively on the generated data can answer visual questions in a zero-shot manner. In order to bypass the data generation procedure, we present an alternative approach, dubbed FrozenBiLM, that directly leverages bidirectional masked language models. Second, we develop TubeDETR, a transformer model that can spatially and temporally localize a natural language query in an untrimmed video. Unlike prior spatio-temporal grounding approaches, TubeDETR can be effectively trained end-to-end on untrimmed videos. Third, we present a new model and a new dataset for multi-event understanding in untrimmed videos. We introduce the Vid2Seq model which generates dense natural language descriptions and corresponding temporal boundaries for all events in an untrimmed video by predicting a single sequence of tokens. Moreover, Vid2Seq can be effectively pretrained on narrated videos at scale using transcribed speech as pseudo-supervision. Finally, we introduce VidChapters-7M, a large-scale dataset of user-chaptered videos. Based on this dataset, we evaluate state-of-the-art models on three tasks including video chapter generation. We also show that video chapter generation models transfer well to dense video captioning in both zero-shot and finetuning settings.

8.4.7 Contributions to the Design and Training of Transformers in Computer Vision

Participants: Alaaeldin Ali.

The goal of this thesis [33] is to explore the potential of Transformer models for computer vision. We propose architectural innovations to overcome their limitations, developing sample-efficient self-supervised pre-training methods, and advancing multimodal learning with Transformers. First, we propose Cross-Covariance Attention to reduce the quadratic complexity of self-attention achieving similar performance as vision transformers with lower memory footprint and computational cost, enabling the application of vision transformers to higher-resolution images. We then investigate self-supervised pre-training for vision transformers. We propose SplitMask, a denoising autoencoosing method based on masked image modeling. Unlike joint embedding methods, SplitMask does not require large-scale pre-training datasets and can be applied to diverse visual data. SplitMask matches the performance of joint embedding methods when pre-trained on datasets two orders of magnitude smaller, highlighting its improved sample efficiency. Moreover, we apply masked image modeling to neural image compression in the form of an improved entropy model yielding a strong rate-distortion performance and enabling the compression of images to the size of a short SMS or tweet. Finally, we propose ImageBind, a method for learning a shared embedding space across six modalities. ImageBind leverages the abundance of images and text on the web to enable transfer to modalities with scarce annotations like depth, thermal, audio, and IMU. In summary, this thesis demonstrates the potential of Transformers for computer vision through architectural innovations, new self-supervised objectives, and multimodal knowledge transfer. The methods proposed in this thesis push the boundaries of transformers in vision by enhancing their scalability and generality, enabling more sample-efficient representation learning, and facilitating transfer across modalities.

8.4.8 Image Reconstruction from Multiple Shots with Trainable Algorithms

Participants: Bruno Lecouat.

The goal of this thesis [31] is dedicated to an in-depth exploration of hybrid methods for solving inverse problems, with a specific focus on their pragmatic implementation in burst photography for real-world applications. The first part of this thesis studies hybrid methods for single-image restoration, providing some methodological tools and some tricks for unrolled optimization. We propose a trainable non-local sparse model for image restoration, leveraging a differentiable relaxation of the unrolled group lasso solver. Taking it a step further, we propose a framework providing differentiable relaxations of convex non-smooth optimization solvers for classic image priors and some. These models demonstrate comparable performance to large neural networks but with significantly fewer parameters, increased interpretability, and faster training times, requiring less training data. The second part of the thesis delves into combining hybrid methods with multi-frame image restoration for super-resolution and HDR reconstruction applications. In this section, our primary focus is reconstructing scenes using rial-world images rather than relying on experiments conducted with synthetic data. The design of plug-and-play (PnP) algorithms for burst photography is explored, with efforts directed toward practical implementation and optimization for mobile devices. Throughout our investigation, we have consistently identified registration quality as a prominent bottleneck. Finally, we propose a novel, dense multi-frame registration algorithm to tackle this challenge effectively, enabling 3D scene reconstruction from image bursts with tiny baselines.

8.4.9 Learning Multi-Task Policies for Robotics

Participants: Elliot Chane-Sane.

The goal of this thesis [27] is to introduce novel methods for learning multi-task policies for robotics. In our first contribution, we present a novel reinforcement learning algorithm that learns goal-reaching policies by interacting with the environment. Our approach incorporates imagined subgoals to guide policy learning during training, resulting in higher sample efficiency and the ability to solve more complex temporally extended tasks. In our second contribution, we propose a method for learning policies in multi-task vision-based manipulation environments that can follow human video instructions. By utilizing an existing large dataset of labeled human videos, we achieve this without requiring annotated robot demonstrations or task-specific reward shaping.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

9.1.1 MSR-Inria joint lab: Image and video mining for science and humanities (Inria)

Participants: Ivan Laptev, Jean Ponce, Josef Sivic.

This collaborative project brings together the WILLOW and THOTH project-teams with MSR researchers in Cambridge and elsewhere. The concept builds on several ideas articulated in the 2020 Sciencea report, including the importance of data mining and machine learning in computational science. Rather than focusing only on natural sciences, however, we propose here to expand the breadth of e-science to include humanities and social sciences. The project we propose will focus on fundamental computer science research in computer vision and machine learning, and its application to archaeology, cultural heritage preservation, environmental science, and sociology, and it will be validated by collaborations with researchers and practitioners in these fields.

In 2018 a new agreement has been signed with a new focus on video understanding for personal assistants. The scientific objectives are to develop models, representations and learning algorithms for (i) automatic understanding of task-driven complex human activities from videos narrated with natural language in order to (ii) give people instructions in a new environment via an augmented reality device such as the Microsoft HoloLens. Besides the clear scientific interest of automatically understanding human activities in video streams, the main high-impact motivation of this project it to develop virtual assistants that may guide a child through simple games to improve his/her manipulation and language skills; help an elderly person to achieve everyday tasks; or facilitate the training of a new worker for highly-specialized machinery maintenance.

9.1.2 Louis Vuitton/ENS chair on artificial intelligence

Participants: Ivan Laptev, Jean Ponce, Josef Sivic.

The scientific chair Louis Vuitton - École normale supérieure in Artificial Intelligence has been created in 2017 and inaugurated on April 12, 2018 by the ENS Director Marc Mézard and the LV CEO Michael Burke. The goal of the chair is to establish a close collaboration between LV and ENS in the area of Artificial Intelligence. The chair enjoys the generous annual contribution of 200K Euros provided by LV in support of research activities in statistical learning and computer vision. In particular, the chair supports the costs of researchers, students, missions, computational resources as well as seminars and meetings, including the two days of meeting annually organized by LV and ENS. During 2020 ENS and LV have organized several joint meetings with the participation of researchers from SIERRA and WILLOW teams. The chair has also supported the hiring of one PhD student at the WILLOW team, missions to conferences and international research labs as well as data collection for research projects. In 2020 the chair has been extended to the next three-year period until 2023. We are planning to start a CIFRE PhD of François Gardères together with Louis Vuitton in 2023.

9.1.3 Casino/ENS chair on algorithmic and machine learning

Participants: Justin Carpentier.

The scientific chair Casino/ENS - École normale supérieure on algorithmic and machine learning has been created in 2021. J. Carpentier is in charge of the robotics axis of this chair.

9.2 Bilateral grants with industry

9.2.1 Facebook AI Research Paris: Weakly-supervised interpretation of image and video data (Inria)

Participants: Jean Ponce.

We will develop in this project (Facebook gift) new models of image and video content, as well as new recognition architectures and algorithms, to address the problem of understanding the visual content of images and videos using weak forms of supervision, such as the fact that multiple images contain instances of the same objects, or the textual information available in television or film scripts.

9.2.2 Google: Multimodal video representation with cross-modal learning (Inria)

Participants: Ivan Laptev.

The proposed project (Google gift) aims to learn a detailed correspondence between the text and the visual content of the video from large-scale unlabeled video collections. It will significantly extend current representations which rely on frame/clip based features and at best learn correlation based on transformers, but fail to provide the in-depth understanding of spatial and temporal structure of the visual content in the video. This will enable advanced multimodal video representations and hence will improve downstream tasks such as video captioning, search and summarization. The main challenge of the project is to build new state-of-the-art models and methods for self-supervised learning based on large-scale but imprecise textual information obtained from video transcripts and other video metadata. The project includes the collection of a dataset allowing a detailed analysis of the visual representation by extending the HowTo100Million dataset with manual annotations.

9.2.3 Google: Structured learning from video and natural language (Inria)

Participants: Ivan Laptev, Josef Sivic.

People can easily learn how to change a flat tire of a car or assemble an IKEA shelve by observing other people doing the same task, for example, by watching a narrated instruction video. In addition, they can easily perform the same task in a different context, for example, at their home. This involves advanced visual intelligence abilities such as recognition of objects and their function as well as interpreting sequences of human actions that achieve a specific task. However, currently there is no artificial system with a similar cognitive visual competence. The goal of this proposal is to develop models, representations and learning algorithms for automatic understanding of complex human activities from videos narrated with natural language.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria associate team not involved in an IIL or an international program

WILLOW has had a long-term (2007–), on-going collaboration with MSR researchers from Cambridge and Redmond, funded by the MSR-Inria joint centre. The new phase of the project started in January 2018 aims at understanding instructional videos and hand-object manipulations in relation to AR devices such as HoloLens. WILLOW has also had a long-term and still ongoing collaboration with Carnegie-Mellon University, funded in part by the GAYA Inria associate team.

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

Ajay Sathya

Status PhD

Institution of origin: KU Leuven

Country: Belgium

Dates: March 2023 – June 2023

Context of the visit: Collaboration on the development of new rigid-body dynamics algorithms for efficient simulation of constrained dynamical systems

Mobility program/type of mobility: research stay

Bruce Wingo

Status PhD

Institution of origin: Georgia Tech

Country: USA

Dates: March 2023 – June 2024

Context of the visit: Collaboration on the development of efficient and generic control schemes for the control of constrained dynamical systems

Mobility program/type of mobility: research stay

Kateryna Zorina

Status PhD

Institution of origin: CTU Prague

Country: Czech Republic

Dates: Jan 2023 – Sept 2023

Context of the visit: Collaboration on the development of new compuser vision and optimal estimation frameworks for the reconstruction of human motion from differentiable simulation

Mobility program/type of mobility: research stay

10.3 European initiatives

10.3.1 Horizon Europe

AGIMUS

Participants: Justin Carpentier, Etienne Arlaud, Pierre-Guillaume Raverdy, Quentin Le Lidec, Louis Montaut, Wilson Jallet.

AGIMUS project on cordis.europa.eu

Title: Next generation of AI-powered robotics for agile production

Duration: From October 1, 2022 to September 30, 2026

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- AIRBUS, France
- KLEEMANN HELLAS SA (KLEEMANN HELLAS SA), Greece
- PAL ROBOTICS SL (PAL ROBOTICS), Spain
- Q-PLAN INTERNATIONAL ADVISORS PC (Q-PLAN INTERNATIONAL), Greece
- TOWARD SAS, France
- THIMM OBALY, K.S., Czechia
- CESKE VYSOKE UCENI TECHNICKE V PRAZE (CVUT), Czechia
- CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS (CNRS), France

Inria contact: Justin Carpentier

Coordinator:

Summary: AGIMUS aims to deliver an open-source breakthrough innovation in AI-powered agile production, introducing solutions that push the limits of perception, planning, and control in robotics, enabling general-purpose robots to be quick to set-up, autonomous and to easily adapt to changes in the manufacturing process. To achieve such agile production, AGIMUS leverages on cutting-edge technologies and goes beyond the state-of-the-art to equip current mobile manipulators with a combination of (i) an advanced task and motion planner that can learn from online available video demonstrations; (ii) optimal control policies obtained from advances in reinforcement learning based on efficient differentiable physics simulations of the manufacturing process; as well as (iii) advanced perception algorithms able to handle objects and situations unseen during initial training. Along the way, optimization of energy efficiency and the use of 5G technology will support further pushing the limits of autonomy. The AGIMUS solutions and their impact will be demonstrated and thoroughly stress-tested in 3 testing zones, as well as 3 industrial pilots in Europe, under numerous diverse real-world case studies and scenarios (different tools, environments, processes, etc.). In every step, and from the very beginning, AGIMUS will go beyond current norms and involve a wide range of stakeholders, starting from the production line itself, to identify the essential ethical-by-design principles and guidelines that can maximise acceptance and impact.

10.4 National initiatives

10.4.1 PRAIRIE

Participants: Justin Carpentier, Ivan Laptev, Jean Ponce, Cordelia Schmid.

The Prairie Institute (PaRis AI Research InstitutE) is one of the four French Institutes for Interdisciplinary Artificial Intelligence Research (3IA), which were created as part of the national French initiative on AI announced by President Emmanuel Macron on May 29, 2018. It brings together five academic partners (CNRS, Inria, Institut Pasteur, PSL University, and University of Paris) as well as 17 industrial partners, large corporations which are major players in AI at the French, European and international levels, as well as 45 Chair holders, including four of the members of WILLOW (Carpentier, Laptev, Ponce, Schmid). Ponce is the scientific director of PRAIRIE.

10.4.2 VideoPredict: Predicting future video content

Participants: Cordelia Schmid, Jean Ponce.

Predicting future video content is a challenging problem with high potential impact in downstream tasks such as self-driving cars and robotics, but also much promise for the learning process itself, from self-supervised learning to data augmentation. Existing approaches range from predicting future actions with semantic labels to creating realistic renderings of future frames. Most of them use straight predictions from convolutional features of previous frames. We propose instead to model the causality effects involved in the video formation process, and disentangle motion and appearance factors. This will result in better prediction, but also and maybe more importantly in a better, more structured understanding of the video content, leading to explicable and interpretable results, and eventually to more trustworthy learning systems. The German and French partners are, respectively, experts in machine learning and computer vision, with complementary research threads in causality and disentangled data models on the one hand, and video understanding and action recognition on the other hand, that are ideally suited for this collaborative project

10.4.3 PEPR Organic Robotics

Participants: Justin Carpentier, Stephane Caron.

The PEPR O2R "Organic Robotics" aims to initiate a change in robotics to create a new generation of robots capable of fluid and natural interactions with users, of social adaptation in their interactions, and which accompanies the technological transitions of societies by producing adapted, responsive and reliable services to citizens. In the frame of this national program, WILLOW is involved in Structuring Action 2 (Robot motion with physical interactions and social adaptation) led by Philippe Souères at LAAS-CNRS, and Structuring Action 4 (Modelling, Simulation, Multi-scale, and Biomechanics) led by Jérémie Dequidt at Inria DEFROST. J. Carpentier is also a member of the executive committee of the PEPR.

10.4.4 NIMBLE (ANR JCJC): Inexact optimization for robot control

Participants: Justin Carpentier, Stephane Caron, Etienne Arlaud, Jean Ponce, Oumayma Bounou, Joris Vaillant, Wilson Jallet.

The limited agility and dexterity of modern robots prevent them from being deployed outside of laboratories, not even mentioning outside of factories. With NIMBLE, we want to point the classical sense-plan-act design pattern, widely adopted in robotics, as one of the main limiting factor. We propose to replace this three-part control paradigm by learning, from real robot experiments, a predictive model of the robot sensorimotor capabilities. This sensorimotor model will be notably exploited to take complex decisions generalizing to unforeseen situations directly from sensor measurements. While NIMBLE's innovation takes its roots in the observation of the human motor control organization, it is grounded by advanced and principled mathematical methodologies, in particular, the Koopman operator sitting on top of (deep) learning, and exploits our recognized expertise in robot modeling, optimal control and machine learning for real robots. It will notably enable complex tasks to be defined and executed directly in the sensor space. The success of NIMBLE will be asserted by clear benchmarks in quadrupedal locomotion able to optimally adapt to unstructured terrains and in mobile manipulation for opening unknown doors using the sound combination of force and visual feedback.

10.4.5 INEXACTE (ANR PRCE): Inexact optimization for robot control

Participants: Justin Carpentier, Stephane Caron, Etienne Arlaud, Antoine Bambade, Joris Vaillant, Wilson Jallet.

Robotic systems are expected to take a large place in tomorrow's society, far beyond current industrial robots in tightly controlled factory environments, with large impacts in terms of safety, health at work, comfort and productivity. The motion of robots is typically designed and controlled by specifying numerical objectives and constraints on what they must do, and within which limits. These specifications often conflict, and the actual control must be computed to satisfy all of them in the best possible way. This is naturally achieved by solving a numerical optimization problem. Such problems are often small enough in robotics that they can be solved exactly in theory, but they are always based on models, and by definition, models reflect reality imperfectly, even more so as we get away from tightly controlled (factory) environments.

We propose a complete change of paradigm, to acknowledge that we actually solve inaccurate optimization problems that provide inaccurate solutions by construction, and explore the following two hypotheses: (H1) We can obtain the exact same performance with imprecise numerical solutions, (H2) we can obtain these imprecise numerical solutions using less costly numerical methods, which can be computed faster, using less demanding hardware. To the best of our knowledge, these questions have barely been explored and INEXACT will provide the first comprehensive exploration of this topic.

Our short-term ambition is to significantly lower the computational requirements for solving control problems, taking advantage of the imprecisions inherent to robotics control to compute appropriate solutions faster. But ultimately, our long-term ambition is to design less fragile, less expensive and less polluting robots, since being less dependent on precise models can make us less dependent on precise and therefore complex, fragile, expensive and resource-demanding mechatronics.

10.5 Regional initiatives

10.5.1 (

AI4IDF)

Participants: Justin Carpentier, Jean Ponce, Etienne Arlaud, Pierre-Guillaume Raverdy.

The Ile-de-France region is home to the world's largest mathematics community, several of France's largest computer science laboratories, and a dense industrial fabric in Artificial Intelligence. In this extremely rich context, the four main AI institutes – DATAIA, Hi! PARIS, PRAIRIE and SCAI – have joined forces within the "AI4IDF" program to structure and animate the community, and offer to industrial and international partners a unified vision of the exceptional forces at work. The scientific program of the AI4IDF project aims to deepen knowledge in AI while keeping the human in-the-loop. It is divided into four axes: (1) Learning and optimization, (2) NLP and dialogue with humans, (3) Robotics, movement and interaction with humans, (4) AI in human life: health, education and creation.

11 Dissemination

Participants: Ivan Laptev, Jean Ponce, Josef Sivic, Justin Carpentier, Stephane Caron, Cordelia Schmid, Gabriel Fiastre, Ricardo Garcia, Quentin Le Lidec, Louis Montaut, Yann Dubois De Mont-Marin.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- J. Ponce and C. Schmid have been co-general chair for ICCV 2023, one of the biggest conference in Computer Vision.

Member of the organizing committees

- Ellis workshop for Computer Vision and Machine Learning, Metzingen, May 2023

11.1.2 Scientific events: selection

Chair of conference program committees

- I. Laptev has been program chair for ICCV 2023, one of the biggest conference in Computer Vision.

Member of the conference program committees

- C. Schmid was Area Chair of CVPR 2023, ICLR 2023, ECCV 2024.
- S. Chen was Area Chair of CVPR 2023, ICCV 2023, ACM MM 2023, NeurIPS 2023, ICLR 2024.

Reviewer

- IEEE-RAS International Conference on Robotics and Automation, 2023 (J. Carpentier, L. Montaut)
- IEEE-RAS International Conference on Humanoid Robots, 2023 (W. Jallet)
- IEEE/RSJ International Conference on Intelligent Robots, 2023 (W. Jallet, S. Chen)
- Robotics: Science and Systems, 2023 (S. Caron)

11.1.3 Journal

Member of the editorial boards

- Associate Editor, IEEE Transactions on Robotics (J. Carpentier, S. Caron).
- Associate Editor, IEEE Robotics and Automation Letters (J. Carpentier).
- Associate Editor, International Journal of Computer Vision (I. Laptev).

Reviewer - reviewing activities

- IEEE Transactions on Robotics (W. Jallet, S. Chen, Y. Labbé, J. Carpentier, S. Caron, Y. de Mont-Marin, Q. Le Lidec, E. Moullet, S. Chen).
- IEEE Robotics and Automation Letters (W. Jallet, J. Carpentier, S. Caron, T. Chabal, R. Garcia, S. Chen)
- IEEE Transactions on Multimedia (S. Chen)
- International Journal of Computer Vision (S. Chen)
- Transactions on Pattern Analysis and Machine Intelligence (S. Chen)

11.1.4 Invited talks

- R. Garcia and T. Chabal, NYU Tandon School of Engineering, 2023.
- C. Schmid, 3rd NAVER LABS Europe International Workshop on AI for Robotics, Grenoble, November 2023.
- C. Schmid, Workshop on New Ideas in Vision Transformers, in conjunction with ICCV'23, Paris, October 2023.
- C. Schmid, Fifth Large-scale Video Object Segmentation Workshop, in conjunction with ICCV'23, Paris, October 2023.
- C. Schmid, keynote at Workshop "Fondements Mathématiques de l'IA", Paris, October 2023.
- C. Schmid, Video AI Symposium, London, September 2023
- C. Schmid, Nature Webcast, September 2023.
- C. Schmid, MPI Summer Colloquium, Tubingen, July 2023.
- C. Schmid, "T4V: Transformers for Vision" workshop, in conjunction with CVPR'23, June 2023.
- C. Schmid, "Scholars & Big Models: How Can Academics Adapt?", in conjunction with CVPR'23, June 2023.
- C. Schmid, Fourth International Workshop on Large Scale Holistic Video Understanding, in conjunction with CVPR'23, June 2023.

- C. Schmid, New Frontiers for Zero-Shot Image Captioning Evaluation Workshop, in conjunction with CVPR'23, June 2023.
- C. Schmid, seminar at Ellis workshop, May 2023.
- C. Schmid, Bavarian International Conference on AI, Munich, February 2023.
- C. Schmid, Soirée networking Prairie, January 2023.
- J. Carpentier, Speacker at Journées Nationales de la Recherche en Robotique, Oct 2023.
- J. Carpentier, Invited Speacker at Journées Nationales de la Recherche en Robotique Humanoide, Bordeaux, July 2023.
- J. Carpentier, KU Leuven Robotics Seminar, May 2023.
- J. Carpentier, Soirée networking Prairie, March 2023.
- J. Carpentier, IEEE-RAS TC on Model-Based Optimization for Robotics, Feb 2023.
- S. Caron, Invited Speacker at Journées Nationales de la Recherche en Robotique Humanoide, Bordeaux, July 2023.
- Q. Le Lidec, Invited Speacker at Journées Nationales de la Recherche en Robotique Humanoide, Bordeaux, July 2023.
- L. Montaut, Invited Speacker at Journées Nationales de la Recherche en Robotique Humanoide, Bordeaux, July 2023.
- S. Chen, BAAI Young Researcher Seminar, August 2023.
- S. Chen, CAAI Tutorial on Embodied AI, July 2023.
- J. Ponce, Séminaire du Directeur de l'ENS, Nov. 2023.
- J. Ponce, Table ronde "Implications de l'IA dans le domaine de la recherche scientifiqueENS", Paris, Nov. 2023
- J. Ponce, Flatiron Institute, New York, April 2023
- J. Ponce, Mad Seminar, NYU, New York, April 2023
- J. Ponce, PRAIRIE/Riken AIP Workshop, Tokyo, March 2023
- J. Ponce, Tokyo Tech, Tokyo, March 2023
- J. Ponce, INDAM Workshop on Learning for Inverse Problems, Rome,June 2023
- J. Ponce, International Summer School on Computer Vision, Sicily, July 2023
- J. Ponce, Keynote, Junior Conference on Data Science and Engineering, Saclay, Sep. 2023
- J. Ponce, Seoul National University, Seoul, Nov. 2023
- J. Ponce, AI Summit, Seoul, Nov. 2023
- J. Ponce, Présentation "IA et vision artificielle", Club We are, Paris, Nov. 2023

11.1.5 Leadership within the scientific community

- Board Member, European Laboratory for Learning and Intelligent Systems (J. Sivic).
- Executive committee member, PEPR O2R (J. Carpentier).
- Executive committee member, Computer Science department, ENS (J. Carpentier).
- Director of Ellis program on Computer Vision and Machine Learning, (C. Schmid).
- Global Member of the Bavarian AI Council (J. Sivic).
- Member of the PAMI-TC executive committee (C. Schmid).
- Member of the PAMI-TC awards committee (C. Schmid).
- Member of IEEE Computer Society and the Computer Vision Foundation CVPR/ICCV Steering committee (C. Schmid).
- Member of Fellows Appointment Committee of the T ubingen ELLIS Institute (C. Schmid).
- Member of the Munich Center for Machine Learning (MCML) Advisory Board (C. Schmid).
- Director of Ellis program on Computer Vision and Machine Learning (C. Schmid).

11.1.6 Scientific expertise

- Head of scientific board at VisionLabs (I. Laptev).

11.1.7 Research administration

- Scientific director, PRAIRIE 3IA Institute (J. Ponce).

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Master: Convex Optimization, M2, École normale supérieure, and MVA, École normale supérieure Paris-Saclay (Q. Le Lidec, Teaching Assistant).
- Master: Introduction à la vision artificielle, M1, École normale supérieure Paris (J. Ponce, O. Bounou).
- Master: Introduction to computer vision, NYU (J. Ponce).
- Master: Object recognition and computer vision, M2, École normale supérieure, and MVA, École normale supérieure Paris-Saclay, 36h (I. Laptev, J. Ponce, J. Sivic and C. Schmid, with G. Le Moing as Teaching Assistant).
- Master: Robotics courses at ENS-DI, 30h (J. Carpentier, S. Caron and Y. DM as Teaching Assistant).
- Master: Robotics courses at ENS-MVA, 30h (J. Carpentier, S. Caron).
- Master: Robotics courses at Formation des Inge‘nieurs de l'Automobile, PSL, 15h (J. Carpentier).
- Master: Computer Vision courses at Formation des Inge‘nieurs de l'Automobile, PSL, 18h (J. Ponce).
- Master: Three lectures (3 x 1.5h) in the 3D computer vision class of V. Hlavac at Charles University in Prague (J. Sivic).
- Introduction to computer vision course at the PSL Intensive week, 3H, (A. Yang as Teaching Assistant).

11.2.2 Supervision

- PhD in progress: Antoine Bambade, started in Oct 2020, J. Carpentier, A. Taylor (Sierra) and J. Ponce.
- PhD in progress: Adrien Bardes, started in Oct 2020, J. Ponce.
- PhD in progress: Theo Bodrito, started in Sep 2021, J. Ponce and J. Mairal (Inria Grenoble)
- PhD in progress: Oumayma Bounou, started in Oct 2020, J. Ponce and J. Carpentier.
- PhD in progress: Nicolas Chahine, started in Aug 2021, J. Ponce.
- PhD in progress: Thomas Chabal, started in Sept 2021, J. Ponce and C. Schmid.
- PhD in progress: Zerui Chen, started in March 2022, I. Laptev and C. Schmid.
- PhD finished in 2023: Elliot Chane-Sane, started in Oct. 2020, graduated in Sept. 2023, I. Laptev and C. Schmid.
- PhD in progress: Yann Dubois de Mont-Marin, started in Sept. 2020, J. Ponce and J. Carpentier.
- PhD in progress: Matthieu Futeral-Peter, started in Nov 2021, I. Laptev and C. Schmid.
- PhD in progress: Ricardo Garcia Pinel, started in Sep 2021, I. Laptev and C. Schmid.
- PhD finished in 2023: Pierre-Louis Guhur, started in Oct 2019, graduated in Feb. 2023, I. Laptev and C. Schmid.
- PhD in progress: Wilson Jallet, started in Oct 2021, J. Carpentier and N. Mansard.
- PhD finished in 2023: Yann Labbe, started in Oct 2018, graduated in June 2023, J. Sivic and I. Laptev.
- PhD in progress: Quentin Le Lidec, started in Oct 2021, J. Carpentier, I. Laptev and C. Schmid.
- PhD in progress: Guillaume Le Moing, started in Nov 2020, J. Ponce and C. Schmid.
- PhD finished in 2023: Bruno Lecouat, started in Sept 2019, graduated in Dec 2023, J. Ponce and J. Mairal (Inria Grenoble).
- PhD finished in 2023: Zongmian Li, started in Oct 2017, graduated in Feb 2023, I. Laptev, J. Sivic and N. Mansard (LAAS/CNRS, Toulouse).
- PhD in progress: Louis Montaut, started in Sept 2020, J. Carpentier, I. Laptev and J. Sivic.
- PhD finished in 2023: Robin Strudel, started in Oct 2018, graduated in Feb 2023, I. Laptev, C. Schmid and J. Sivic.
- PhD in progress: Lucas Ventura, started in Oct 2022, C. Schmid and G. Varol (ENPC).
- PhD in progress: Elliot Vincent, started in Sep 2021, J. Ponce and M. Aubry (ENPC).
- PhD finished in 2023: Antoine Yang, started in Oct. 2020, graduated in Oct. 2023, I. Laptev, C. Schmid and J. Sivic.
- PhD in progress: Fabian Schramm, started in Feb 2023, J. Carpentier and N. Perrin-Gilbert (ISIR).
- PhD in progress: Zeeshan Khan, started in Sept 2023, S. Chen and C. Schmid.
- PhD in progress: Gabriel Fiastre , started in Sept 2023, C. Schmid.
- PhD in progress: Ludovic de Matteis, started in Oct 2023, J. Carpentier and N. Mansard (LAAS).

11.2.3 Juries

- Alexandre Ramé, October 2023, Sorbonne University, Paris (C. Schmid).
- Medhini Narasimhan, August 2023, University of California, Berkeley (C. Schmid).
- Florent Bartoccioni, May 2023, UGA (C. Schmid).
- Mert Bulent Sariyildiz, June 2023, UGA (C. Schmid).
- Florent Bartoccioni, May 2023, UGA (C. Schmid).
- Tonmoy Saikia, April 2023, Universität Freiburg (C. Schmid).
- Etienne Ménager, Dec 2023, Univeristé de Lille (J. Carpentier).
- Alejandro Astudillo Vigoya, May 2023, KU Leuven (J. Carpentier).
- Antun Skuric, Nov 2023, Université de Bordeaux (S. Caron).

11.3 Popularization

11.3.1 Articles and contents

- Carte blanche au Monde, article spécilisé mensuel, Le Monde (J. Ponce)
- Le Dessous des Images, ARTE (J. Carpentier)

12 Scientific production

12.1 Major publications

[1] P.-L. Guhur, S. Chen, R. Garcia, M. Tapaswi, I. Laptev and C. Schmid. 'Instruction-driven history-aware policies for robotic manipulations'. In: CoRL 2022 - Conference on Robot Learning. Aukland, New Zealand, 14th Dec. 2022. URL: https://hal.science/hal-03775734.

[2] Y. de Mont-Marin, J. Ponce and J.-P. Laumond. *A minimum swept-volume metric structure for configuration space*. 21st Nov. 2022. DOI: 10.48550/arXiv.2211.11811. URL: https://inria.hal.science/hal-03856704.

[3] L. Montaut, Q. Le Lidec, A. Bambade, V. Petrík, J. Sivic and J. Carpentier. *Differentiable Collision Detection: a Randomized Smoothing Approach*. 14th Apr. 2023. URL: https://hal.science/hal-03780482.

[4] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic and C. Schmid. 'Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning'. In: CVPR 2023 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023. URL: https://inria.hal.science/hal-04039246.

12.2 Publications of the year

International journals

[5] Q. Le Lidec, F. Schramm, L. Montaut, C. Schmid, I. Laptev and J. Carpentier. 'Leveraging Randomized Smoothing for Optimal Control of Nonsmooth Dynamical Systems'. In: *Nonlinear Analysis: Hybrid Systems* 52 (22nd Jan. 2024), p. 101468. DOI: 10.1016/j.nahs.2024.101468. URL: https://hal.science/hal-03480419.

[6] Y. de Mont-Marin, J. Ponce and J.-P. Laumond. 'A minimum swept-volume metric structure for configuration space'. In: *IEEE International Conference on Robotics and Automation (ICRA)* (1st July 2023), pp. 3686–3692. DOI: 10.48550/arXiv.2211.11811. URL: https://inria.hal.science/hal-03856704.

[7] C. Pieters, E. Danblon, P. Souères and J.-P. Laumond. 'Talking about moving machines'. In: *Interaction Studies*. Interaction Studies 23.2 (15th Mar. 2023), pp. 322–340. DOI: 10.1075/is.22005.pie. URL: https://laas.hal.science/hal-04269587.

International peer-reviewed conferences

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

[9] O. Bounou, J. Ponce and J. Carpentier. 'Leveraging Proximal Optimization for Differentiating Optimal Control Solvers'. In: IEEE 62nd Conference on Decision and Control (CDC). Singapore, Singapore, 12th Dec. 2023. URL: https://hal.science/hal-03786820.

[10] N. Chahine, A.-S. Calarasanu, D. Garcia-Civiero, T. Cayla, S. Ferradans and J. Ponce. 'An Image Quality Assessment Dataset for Portraits'. In: CVPR 2023 - Conference on Computer Vision and Pattern Recognition 2023. Vancouver, Canada: IEEE, 18th June 2023. URL: https://hal.science/hal-04062434.

[11] E. Chane-Sane, C. Schmid and I. Laptev. 'Learning Video-Conditioned Policies for Unseen Manipulation Tasks'. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE, 29th May 2023, pp. 909–916. DOI: 10.1109/ICRA48891.2023.10161336. URL: https://inria.hal.science/hal-04226856.

[12] S. Chen, T. Chabal, I. Laptev and C. Schmid. 'Object Goal Navigation with Recursive Implicit Maps'. In: The 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023). Detroit (Michigan), United States, 1st Oct. 2023. URL: https://inria.hal.science/hal-04215744.

[13] S. Chen, R. Garcia, C. Schmid and I. Laptev. 'PolarNet: 3D Point Clouds for Language-Guided Robotic Manipulation'. In: 7th Conference on Robot Learning (CoRL 2023). Atlanta, GA, United States, 6th Nov. 2023. URL: https://hal.science/hal-04221153.

[14] Z. Chen, S. Chen, C. Schmid and I. Laptev. 'gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction'. In: CVPR 2023 - IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023. Vancouver (Canada), Canada, 2023. URL: https://hal.science/hal-04095352.

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

[16] M. Futeral, C. Schmid, I. Laptev, B. Sagot and R. Bawden. 'Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 3rd July 2023, pp. 5394–5413. DOI: 10.18653/v1/2023.acl-long.295. URL: https://inria.hal.science/hal-03977982.

[17] R. Garcia, R. Strudel, S. Chen, E. Arlaud, I. Laptev and C. Schmid. 'Robust Visual Sim-to-Real Transfer for Robotic Manipulation'. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit, United States: ieee, 28th July 2023. URL: https://hal.science/hal-04192660.

[18] Q. Garrido, Y. Chen, A. Bardes, L. Najman and Y. Lecun. 'On the duality between contrastive and non-contrastive self-supervised learning'. In: ICLR 2023 - Eleventh International Conference on Learning Representations. Kigali, Rwanda, 1st May 2023. DOI: 10.48550/arXiv.2206.02574. URL: https://hal.science/hal-03685169.

[19] Q. Le Lidec, W. Jallet, I. Laptev, C. Schmid and J. Carpentier. ‘Enforcing the consensus between Trajectory Optimization and Policy Learning for precise robot control’. In: *2023 International Conference on Robotics and Automation (ICRA)*. ICRA 2023 - IEEE International Conference on Robotics and Automation. London, United Kingdom, June 2023. URL: https://hal.science/hal-03780392.

[20] G. L. Moing, J. Ponce and C. Schmid. ‘Dense Optical Tracking: Connecting the Dots’. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, United States, 1st Dec. 2023. URL: https://inria.hal.science/hal-04320282.

[21] G. L. Moing, J. Ponce and C. Schmid. ‘WALDO: Future Video Synthesis using Object Layer Decomposition and Parametric Flow Prediction’. In: IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, 4th Oct. 2023. URL: https://inria.hal.science/hal-03889664.

[22] E. Moullet, F. Bailly, J. Carpentier and C. Azevedo Coste. ‘Vision-based interface for grasping intention detection and grip selection : towards intuitive upper-limb assistive devices’. In: Congrès annuel de la Société de Biomécanique. Grenoble, France, 25th Oct. 2023. URL: https://hal.science/hal-04141469.

[23] A. Yang, A. Nagrani, I. Laptev, J. Sivic and C. Schmid. ‘VidChapters-7M: Video Chapters at Scale’. In: NeurIPS 2023 - Conference on Neural Information Processing Systems - Track on Datasets and Benchmarks. Vol. 36. New Orleans (LA), United States, 2023. URL: https://hal.science/hal-04217697.

[24] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic and C. Schmid. ‘Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning’. In: CVPR 2023 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023. URL: https://inria.hal.science/hal-04039246.

[25] K. Zorina, D. Kovar, F. Lamiraux, N. Mansard, J. Carpentier, J. Sivic and V. Petrik. ‘Multi-Contact Task and Motion Planning Guided by Video Demonstration’. In: ICRA 2023 - International Conference on Robotics and Automation. Londres, United Kingdom, 29th May 2023. URL: https://laas.hal.science/hal-03945110.

Conferences without proceedings

[26] R. Loiseau, E. Vincent, M. Aubry and L. Landrieu. ‘Learnable Earth Parser: Discovering 3D Prototypes in Aerial Scans’. In: CVPR. Seattle (USA), United States, 19th Apr. 2023. URL: https://hal.science/hal-04135416.

Doctoral dissertations and habilitation theses

[27] E. Chane-Sane. ‘Learning Multi-Task Policies for Robotics’. Inria Paris; Ecole Normale Superieure, 6th Sept. 2023. URL: https://inria.hal.science/tel-04499924.

[28] H. Cisneros. ‘Unsupervised Learning in Complex Systems’. École normale supérieure, 15th May 2023. URL: https://inria.hal.science/tel-04159511.

[29] P.-L. Guhur. ‘Language-guided navigation and manipulation in robotics using transformers.’ Inria; Ecole Normale Supérieure, 4th Apr. 2023. URL: https://inria.hal.science/tel-04125019.

[30] Y. Labbé. ‘Pose estimation of rigid objects and robots’. Ecole Normale Supérieure, 5th June 2023. URL: https://hal.science/tel-04124865.

[31] B. Lecouat. ‘Image Reconstruction from Multiple Shots with Trainable Algorithms’. INRIA Paris; Ecole Normale Supérieure (ENS), 15th Nov. 2023. URL: https://theses.hal.science/tel-04489120.

[32] Z. Li. ‘Estimating 3D Motion and Forces from Monocular Videos’. INRIA Paris; École Normale Supérieure, 17th Feb. 2023. URL: https://hal.science/tel-04141548.

[33] A. Mohamed Elnouby Abdallah Ali. ‘Contributions to the Design and Training of Transformers in Computer Vision’. INRIA Paris; ENS Paris - Ecole Normale Supérieure de Paris; PSL University, 11th July 2023. URL: https://inria.hal.science/tel-04477587.

[34] R. Strudel. ‘Learning representations for visually-guided robotics’. INRIA Paris; Ecole Normale Superieure, 1st Feb. 2023. URL: https://inria.hal.science/tel-04071261.

[35] A. Yang. ‘Learning Visual Language Models for Video Understanding’. Ecole Normale Superieure de Paris - ENS Paris, 23rd Nov. 2023. URL: https://hal.science/tel-04307117.

Reports & preprints

[36] A. Bambade, F. Schramm, S. E. Kazdadi, S. Caron, A. Taylor and J. Carpentier. *PROXQP: an Efficient and Versatile Quadratic Programming Solver for Real-Time Robotics Applications and Beyond*. 1st Sept. 2023. URL: https://inria.hal.science/hal-04198663.

[37] A. Bambade, F. Schramm, A. Taylor and J. Carpentier. *QPLayer: efficient differentiation of convex quadratic optimization*. 19th June 2023. URL: https://inria.hal.science/hal-04133055.

[38] G. Fadini, S. Kumar, R. Kumar, T. Flayols, A. D. Prete, J. Carpentier and P. Souères. *Co-designing versatile quadruped robots for dynamic and energy-efficient motions*. 16th July 2023. URL: https://laas.hal.science/hal-04162737.

[39] O. Flasseur, T. Bodrito, J. Mairal, J. Ponce, M. Langlois and A.-M. Lagrange. *deep PACO: Combining statistical models with deep learning for exoplanet detection and characterization in direct imaging at high contrast*. 25th May 2023. DOI: 10.48550/arXiv.2303.02461. URL: https://hal.science/hal-04106501.

[40] W. Jallet, A. Bambade, E. Arlaud, S. El-Kazdadi, N. Mansard and J. Carpentier. *PROXDDP: Proximal Constrained Trajectory Optimization*. 8th Dec. 2023. URL: https://inria.hal.science/hal-04332348.

[41] A. Jordana, S. Kleff, A. Meduri, J. Carpentier, N. Mansard and L. Righetti. *Stagewise Implementations of Sequential Quadratic Programming for Model-Predictive Control*. 7th Dec. 2023. URL: https://laas.hal.science/hal-04330251.

[42] A. Jordana, A. Meduri, E. Arlaud, J. Carpentier and L. Righetti. *Risk-Sensitive Extended Kalman Filter*. 1st May 2023. URL: https://inria.hal.science/hal-04107207.

[43] Q. Le Lidec, W. Jallet, L. Montaut, I. Laptev, C. Schmid and J. Carpentier. *Contact Models in Robotics: a Comparative Analysis*. 13th Apr. 2023. URL: https://hal.science/hal-04067291.

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

[45] L. Montaut, Q. Le Lidec, A. Bambade, V. Petrík, J. Sivic and J. Carpentier. *Differentiable Collision Detection: a Randomized Smoothing Approach*. 14th Apr. 2023. URL: https://hal.science/hal-03780482.

[46] L. Montaut, Q. Le Lidec, V. Petrík, J. Sivic and J. Carpentier. *GJK++: Leveraging Acceleration Methods for Faster Collision Detection*. 1st Nov. 2023. URL: https://hal.science/hal-04070039.

[47] L. Ventura, A. Yang, C. Schmid and G. Varol. *CoVR: Learning Composed Video Retrieval from Web Video Captions*. 28th Aug. 2023. URL: https://hal.science/hal-04327307.

[48] E. Vincent, J. Ponce and M. Aubry. *Pixel-wise Agricultural Image Time Series Classification: Comparisons and a Deformable Prototype-based Approach*. 22nd Mar. 2023. URL: https://hal.science/hal-04135119.