RESEARCH CENTRE

**Inria Paris Center**

**IN PARTNERSHIP WITH:**

**Ecole normale supérieure de Paris, CNRS**

2022

# 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, HDR]
- Justin Carpentier [Inria, Researcher]
- Jean Ponce [Inria, until Aug 2022]
- Cordelia Schmid [Inria, Senior Researcher, HDR]

## Faculty Member

- Jean Ponce [ENS Paris, Professor, from Sep 2022]

## Post-Doctoral Fellows

- Alexandre Araujo [Inria]
- Shize Chen [Inria]
- Etienne Moullet [Inria, from Sep 2022]

## PhD Students

- Minttu Alakuijala [CIFRE Google, Inria]
- Alaaeldin Ali [CIFRE Facebook]
- Antoine Bambade [Corps des Ponts et Chaussées]
- Adrien Bardes [CIFRE Facebook]
- Theo Bodrito [Inria]
- Oumayma Bounou [Inria]
- Thomas Chabal [Inria]
- Nicolas Chahine [CIFRE, DXOMARK]
- Elliot Chane-Sane [Inria]
- Zerui Chen [Inria]
- Hugo Cisneros [CTU Prague, ENS]
- Yann Dubois De Mont-Marin [Inria]
- Matthieu Futeral-Peter [Inria]
- Ricardo Garcia Pinel [Inria]
- Pierre-Louis Guhur [Univ Paris Saclay, Inria, until Nov 2022]
- Wilson Jallet [Inria]
- Yann Labbe [ENS Paris, Inria]
- Quentin Le Lidec [Inria]
- Guillaume Le Moing [Inria]

- Bruno Lecouat [Inria]
- Zongmian Li [Inria]
- Louis Montaut [CTU Prague, ENS]
- Lucas Ventura [ENPC]
- Elliot Vincent [Ministère Transition]
- Antoine Yang [Inria]

### Technical Staff

- Etienne Arlaud [Inria]
- Rohan Budhiraja [Inria, Engineer, until Jun 2022]
- Fabian Schramm [Inria, Engineer, from Feb 2022]

### Administrative Assistants

- Julien Guieu [Inria, from Nov 2022]
- Scheherazade Rouag [Inria, until Nov 2022]

### Visiting Scientists

- Stephane Caron [CNRS, from Dec 2022]
- Armand Jordana [New York University, from Sep 2022 until Oct 2022]
- Sebastien Kleff [CNRS, New York University, from Aug 2022 until Aug 2022]
- Ludovic Righetti [New York University, from Jul 2022 until Jul 2022]

### External Collaborators

- Mathieu Aubry [ENPC, HDR]
- Josef Sivic [CTU Prague, HDR]
- Gül Varol [ENPC]

# 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 2022, 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 2022 we have introduced an efficient learning procedure for video question answering [23] that relies on large language models (LLMs) without expensive finetuning of such models. Moreover, in 2022 we have co-organized several climate-related events, namely "Mathématiques, numériques et climat" and "Semaine du climat" at Inria as well as a Workshop on facilitation for the Climate Fresk (x15) and MyCO2 (x4).

# 6 Highlights of the year

Jean Ponce co-founded in 2022 the “Enhance Lab” startup, which commercializes software for joint image demosaicing, denoising, super-resolution and high dynamic range imaging from raw image bursts. This softwaee is based on research done at Inria by Ponce in collaboration with Julien Mairal at Inria Grenoble and their joint PhD student Bruno Lecouat. It has been the subject of ICCV'21 and SIGGRAPH'22 papers, and a French patent, “Dispositif et procédé de formation d'une image à faible niveau de bruit à partir d'une rafale d'images”, B. Lecouat, J. Mairal and J. Ponce, French patent FR2207574, 92 INPI, pending.

In 2022 we have recruited two young roboticist researchers Majid Khadiv and Stéphane Caron. Cordelia Schmid has become a fellow of Asia-Pacific Artificial Intelligence Association. Shizhe Chen has won the 2nd place (Runner up) in the REVERIE Challenge @ CSIG 2022.

# 7 New software and platforms

## 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

**Keyword:** Image processing

**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, Shize 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:** Shize Chen

**Participants:** Shize 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:** Shize Chen

**Participants:** Shize 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:** Shize Chen

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

## 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 panipulator and a moving platform as well as the SOLO robot. These robotic patforms will enable our future research and experiments with locomotion navigation and manipulation.

# 8 New results

## 8.1 Visual recognition and reconstruction of images and videos

### 8.1.1 VICRegL: Self-Supervised Learning of Local Visual Features

**Participants:** Adrien Bardes, Jean Ponce, Yann LeCun.

Most recent self-supervised methods for learning image representations focus on either producing a global feature with invariance properties, or producing a set of local features. The former works best for classification tasks while the latter is best for detection and segmentation tasks. In our work [9], we explore the fundamental trade-off between learning local and global features. A new method called VICRegL is proposed that learns good global and local features simultaneously, yielding excellent performance on detection and segmentation tasks while maintaining good performance on classification tasks. Concretely, two identical branches of a standard convolutional net architecture are fed two differently distorted versions of the same image. The VICReg criterion is applied to pairs of global feature vectors. Simultaneously, the VICReg criterion [8] is applied to pairs of local feature vectors occurring before the last pooling layer. Two local feature vectors are attracted to each other if their $l^2$-distance is below a threshold or if their relative locations are consistent with a known geometric transformation between the two input images. We demonstrate strong performance on linear classification and segmentation transfer tasks. The architecture of VICRegL is presented in Figure 1. This work was presented at NeurIPS'22 [9].

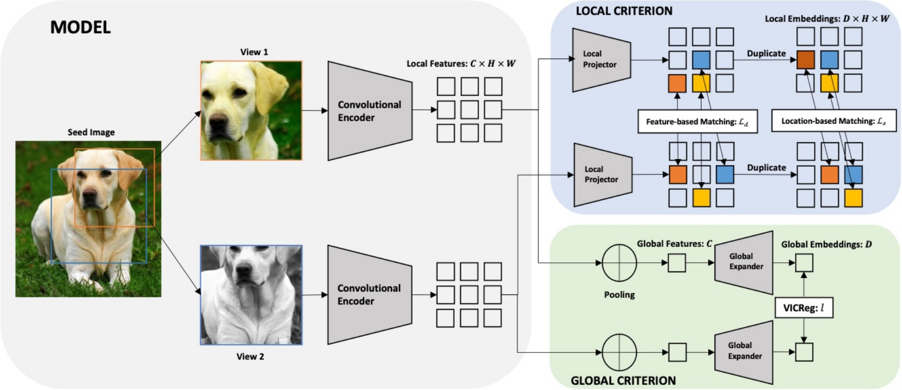


Figure 1: **Overview of VICRegL: Learning local and global features with VICReg.** Given a seed image, two views are produced and fed to an encoder that produces local features. The features are further processed by a local projector that embed them into a smaller space, without destroying the localization information. Two matchings, one based on the spatial information provided by the transformation between the views, the other one based on the $l^2$-distance in the embedding space are computed, and the VICReg criterion is then applied between matched spatial embeddings. Additionally, the local features from the encoder are pooled together, and the pooled features are fed to a global expander. The VICReg criterion is finally applied between the two resulting embeddings.

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

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

Following language instructions to navigate in unseen environments is a challenging problem for autonomous embodied agents. The agent not only needs to ground languages in visual scenes, but also should explore the environment to reach its target. In our work [13], we propose a **du**al-scal**e** graph **t**ransformer (DUET) for joint long-term action planning and fine-grained cross-modal understanding. We build a topological map on-the-fly to enable efficient exploration in global action space. To balance the complexity of large action space reasoning and fine-grained language grounding, we dynamically combine a fine-scale encoding over local observations and a coarse-scale encoding on a global map via graph transformers. Figure 2 presents an overview of the DUET model. The proposed approach, DUET, significantly outperforms state-of-the-art methods on goal-oriented vision-and-language navigation (VLN) benchmarks REVERIE and SOON. It also improves the success rate on the fine-grained VLN benchmark R2R. This work was presented at CVPR'22 [13].

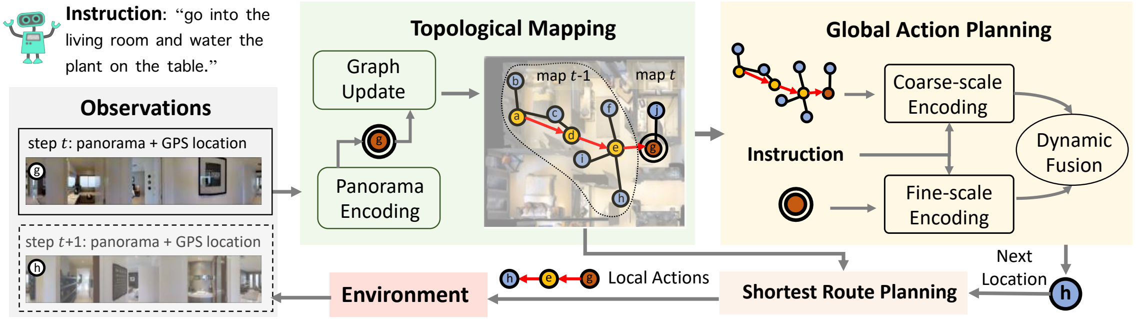


Figure 2: An overview of the DUET approach for vision-and-language navigation. An agent is required to navigate in unseen environments to reach target locations according to language instructions. It only obtains local observations of the environment and is allowed to make local actions, i.e. moving to neighboring locations. In this work, we propose to build topological maps on-the-fly to enable long-term action planning.

### 8.1.3 Learning from Unlabeled 3D Environments for Vision-and-Language Navigation

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

In vision-and-language navigation (VLN), an embodied agent is required to navigate in realistic 3D environments following natural language instructions. One major bottleneck for existing VLN approaches is the lack of sufficient training data, resulting in unsatisfactory generalization to unseen environments. While VLN data is typically collected manually, such an approach is expensive and prevents scalability. In our work [12], we address the data scarcity issue by proposing to automatically create a large-scale VLN dataset from 900 unlabeled 3D buildings from HM3D. As illustrated in Figure 3, we first generate a navigation graph for each building and transfer object predictions from 2D to generate pseudo 3D object labels by cross-view consistency. We then fine-tune a pretrained language model using pseudo object labels as prompts to alleviate the cross-modal gap in instruction generation. Our resulting HM3D-AutoVLN dataset is an order of magnitude larger than existing VLN datasets in terms of navigation environments and instructions. We experimentally demonstrate that HM3D-AutoVLN significantly increases the generalization ability of resulting VLN models. On the SPL metric, our approach improves

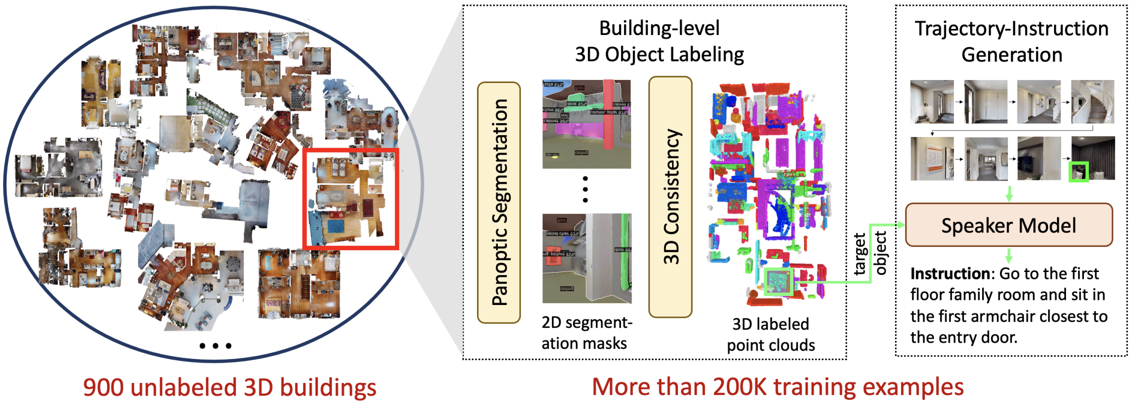


Figure 3: **HM3D-AutoVLN dataset**. We use 900 unlabeled 3D buildings from the HM3D dataset. We improve labels obtained with a 2D segmentation model based on 3D cross-view consistency. Then we rely on these pseudo labels to generate instructions via a speaker model. We automatically create over 200K realistic training samples for VLN.

over state of the art by 7.1% and 8.1% on the unseen validation splits of REVERIE and SOON datasets respectively. This work was presented at ECCV'22 [12].

### 8.1.4 Language Conditioned Spatial Relation Reasoning for 3D Object Grounding

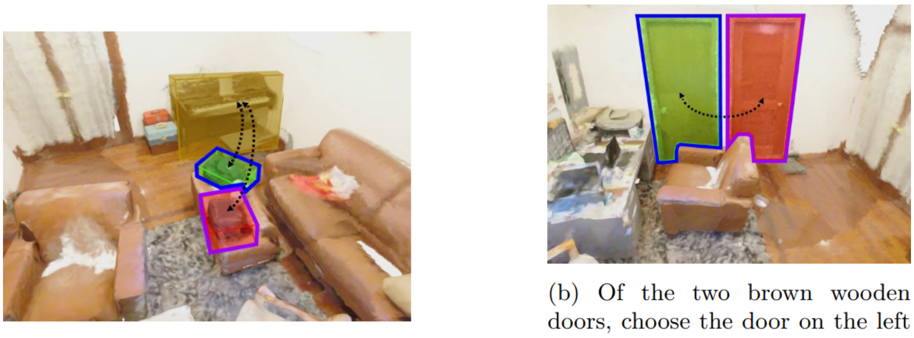
**Participants:** Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, Ivan Laptev.

Localizing objects in 3D scenes based on natural language requires understanding and reasoning about spatial relations. In particular, it is often crucial to distinguish similar objects referred by the text, such as "the left most chair" and "a chair next to the window". In our work [11] we propose a language-conditioned transformer model for grounding 3D objects and their spatial relations. To this end, we design a spatial self-attention layer that accounts for relative distances and orientations between objects in input 3D point clouds. Training such a layer with visual and language inputs enables to disambiguate spatial relations and to localize objects referred by the text. To facilitate the cross-modal learning of relations, we further propose a teacher-student approach where the teacher model is first trained using ground-truth object labels, and then helps to train a student model using point cloud inputs. We perform ablation studies showing advantages of our approach. We also demonstrate our model to significantly outperform the state of the art on the challenging Nr3D, Sr3D and ScanRefer 3D object grounding datasets. Figure 4 shows qualitative examples predicted by our model. This work was presented at NeurIPS'22 [11].

### 8.1.5 AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction

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

Recent work achieved impressive progress towards joint reconstruction of hands and manipulated objects from monocular color images. Existing methods focus on two alternative representations in terms of either parametric meshes or signed distance fields (SDFs). On one side, parametric models can benefit from prior knowledge at the cost of limited shape deformations and mesh resolutions. Mesh models, hence, may fail to precisely reconstruct details such as contact surfaces of hands and objects. SDF-based methods, on the other side, can represent arbitrary details but are lacking explicit priors. In this work we aim to improve SDF models using priors provided by parametric representations. In [14], as shown in Figure 5, we propose a joint learning framework that disentangles the pose and the shape. We obtain



(a) The backpack closest to the piano.

(b) Of the two brown wooden doors, choose the door on the left when facing them.

Figure 4: Example sentences that refer to objects in 3D scenes. The green box denotes the ground-truth object, the blue box is the prediction from our model, and the purple one is from a baseline model without explicit spatial reasoning and knowledge distillation.

hand and object poses from parametric models and use them to align SDFs in 3D space. We show that such aligned SDFs better focus on reconstructing shape details and improve reconstruction accuracy both for hands and objects. We evaluate our method and demonstrate significant improvements over the state of the art on the challenging ObMan and DexYCB benchmarks. This work was presented at ECCV'22 [14].

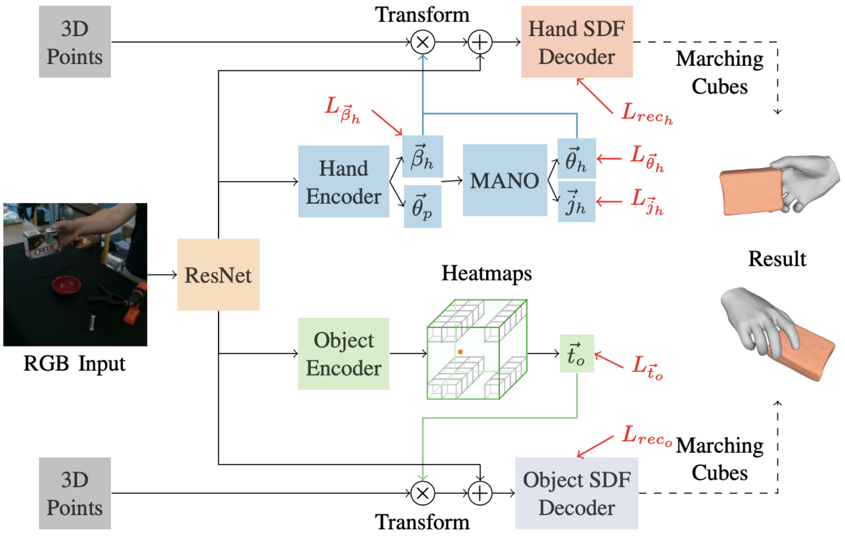
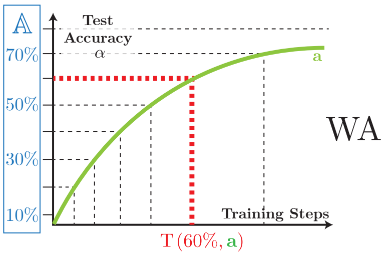


Figure 5: Our proposed method extends SDFs with prior knowledge on hand and object poses obtained via parametric models and can produce detailed meshes for hands and manipulated objects from monocular RGB images.

### 8.1.6 Benchmarking Learning Efficiency in Deep Reservoir Computing

**Participants:** Hugo Cisneros, Josef Sivic, Tomas Mikolov.

It is common to evaluate the performance of a machine learning model by measuring its predictive power on a test dataset. This approach favors complicated models that can smoothly fit complex functions and generalize well from training data points. Although essential components of intelligence, speed and data efficiency of this learning process are rarely reported or compared between different candidate models. In [15], we introduce a benchmark of increasingly difficult tasks together with a data efficiency metric to measure how quickly machine learning models learn from training data. We compare the learning speed of some established sequential supervised models, such as RNNs, LSTMs, or Transformers, with relatively less known alternative models based on reservoir computing. The proposed tasks require a wide range of computational primitives, such as memory or the ability to compute Boolean functions, to be effectively solved. Surprisingly, we observe that reservoir computing systems that rely on dynamically evolving feature maps learn faster than fully supervised methods trained with stochastic gradient optimization while achieving comparable accuracy scores. The code, benchmark, trained models, and results to reproduce our experiments are available at this link. Figure 6 illustrates the construction of the efficiency metric. This work presented at CoLLAs'22 [15].



$$\mathrm{WADE}(\mathbf{a}) = \frac{1}{\sum \alpha} \sum_{\alpha\in\mathbb{A}} \frac{\alpha}{\mathrm{T}(\alpha,\mathbf{a})}$$

Figure 6: A learning curve alongside the equation defining the WADE metric. The efficiency metric is based on measuring the time needed to reach various accuracy checkpoints ($T(\alpha,\mathbf{a})$).

### 8.1.7 Preserving Semantics in Textual Adversarial Attacks

**Participants:** David Herel, Hugo Cisneros, Tomas Mikolov.

Adversarial attacks in NLP challenge the way we look at language models. The goal of this kind of adversarial attack is to modify the input text to fool a classifier while maintaining the original meaning of the text. Although most existing adversarial attacks claim to fulfill the constraint of semantics preservation, careful scrutiny shows otherwise. We show that the problem lies in the text encoders used to determine the similarity of adversarial examples, specifically in the way they are trained. Unsupervised training methods make these encoders more susceptible to problems with antonym recognition. To overcome this, we introduce a simple, fully supervised sentence embedding technique called Semantics-Preserving-Encoder (SPE), illustrated in Figure 7. The results show that our solution minimizes the variation in the meaning of the adversarial examples generated. It also significantly improves the overall quality of adversarial examples, as confirmed by human evaluators. Furthermore, it can be used as a component in any existing attack to speed up its execution while maintaining similar attack success.

### 8.1.8 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 [42] 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

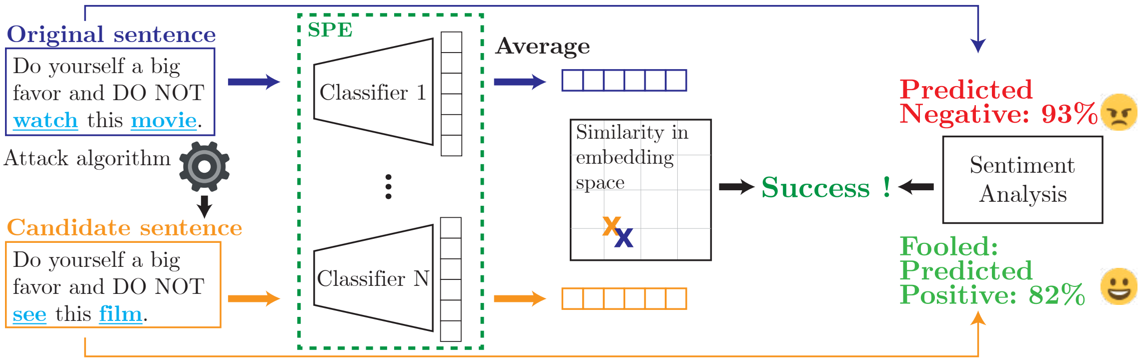

Original sentence
Do yourself a big favor and DO NOT watch this movie.
Attack algorithm
Candidate sentence
Do yourself a big favor and DO NOT see this film.
SPE
Classifier 1
Classifier N
Average
Similarity in embedding space
Success !
Sentiment Analysis
Predicted Negative: 93%
Fooled: Predicted Positive: 82%

Figure 7: The Semantics Preserving Encoder pipeline illustrated.

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 8. More videos can be found from the project webpage.

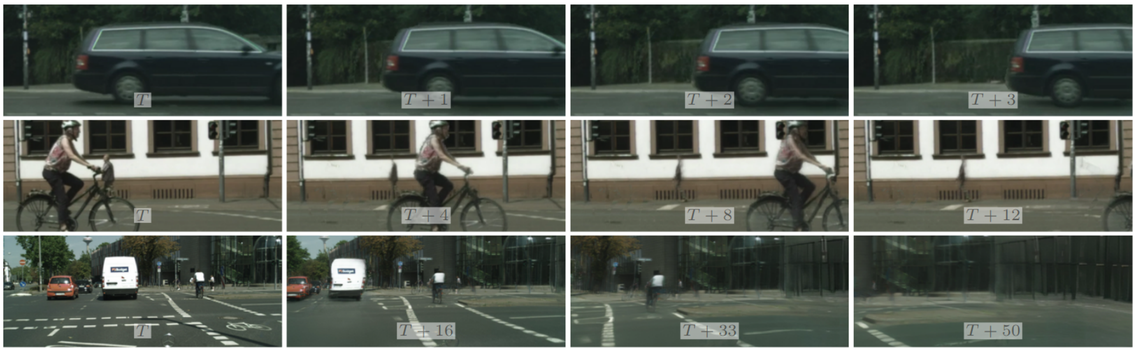


Figure 8: Video frame $T$ and future frames $T+K$ predicted by WALDO from frames 1 to $T$. In our experiments, we use in general $T$=4 (1/4s) and $K$ up to 50 (3s).

### 8.1.9 Spatially-consistent Feature Matching and Learning for Heritage Image Analysis

**Participants:** Xi Shen, Robin Champenois, Shiry Ginosar, Ilaria Pastrolin, Morgane Rousselot, Oumayma Bounou, Tom Monnier, Spyros Gidaris, François Bougard, Pierre-Guillaume Raverdy, Marie-Françoise Limon, Christine Bénévent, Marc Smith, Olivier Poncet, K Bender, Béatrice Joyeux-Prunel, Elizabeth Honig, Alexei Efros, Mathieu Aubry.

Progress in the digitization of cultural assets leads to online databases that become too large for a human to analyze. In our work [4], we explore two applications of computer vision to analyze historical data: watermark recognition and one-shot repeated pattern detection in artwork collections. Both problems present computer vision challenges representative of the ones encountered in cultural heritage applications: limited supervision, fine-grained recognition tasks, and multi-modal data. Both applications are practical, as recognizing watermarks makes it possible to date and locate documents, while detecting repeated patterns allows exploring visual links between artworks. We demonstrate on both the benefits of relying on deep mid-level features. More precisely, we define an image similarity score based on geometric verification of mid-level features and show how spatial consistency can be used to fine-tune out-of-the-box features to the target dataset with weak or no supervision. Our code and data are available at this link.

### 8.1.10 TubeDETR: Spatio-Temporal Video Grounding with Transformers

**Participants:** Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, Cordelia Schmid.

We consider the problem of localizing a spatio-temporal tube in a video corresponding to a given text query, see Figure 9. It is a challenging task that requires the joint modeling of temporal, spatial and multi-modal interactions. To address this task, in [22] we propose TubeDETR, a transformer-based architecture inspired by the recent success of such models for text-conditioned object detection. Our model notably includes: (i) an efficient video and text encoder that models spatial multi-modal interactions over sparsely sampled frames and (ii) a space-time decoder that jointly performs spatio-temporal localization. We demonstrate the advantage of our proposed components through an extensive ablation study. We also evaluate our full approach on the spatio-temporal video grounding task and demonstrate improvements over the state of the art on the challenging VidSTG and HC-STVG benchmarks. Code, datasets and trained models are available at this address. This work was presented at CVPR'22 [22].

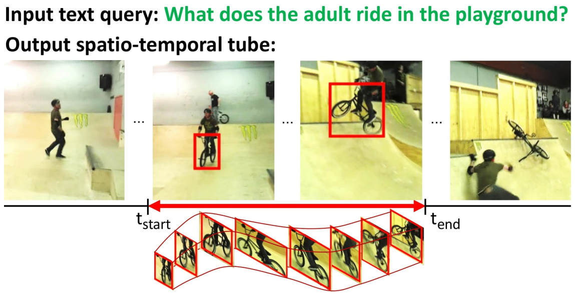
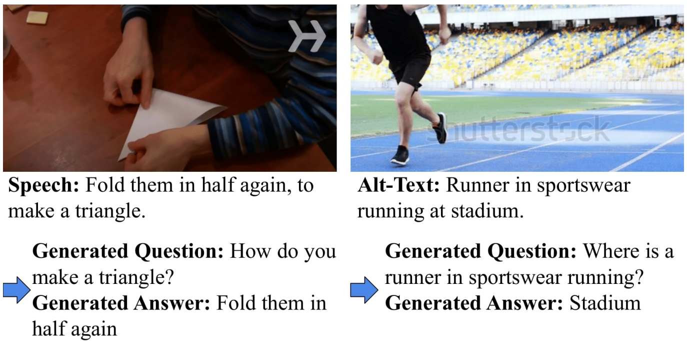

Input text query: What does the adult ride in the playground?
Output spatio-temporal tube:
...
...
...
$t_{start}$
$t_{end}$

Figure 9: Spatio-temporal video grounding requires reasoning about space, time, and language.

### 8.1.11 Learning to Answer Visual Questions from Web Videos

**Participants:** Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, Cordelia Schmid.

Recent methods for visual question answering rely on large-scale annotated datasets. Manual annotation of questions and answers for videos, however, is expensive and prevents scalability. In this work, we propose to avoid it and generate a large-scale training dataset for video question answering making use of automatic cross-modal supervision. We leverage a question generation transformer trained on text data and use it to generate question-answer pairs from transcribed video narrations. Given narrated videos, we then automatically generate the HowToVQA69M dataset with 69M video-question-answer triplets, see Figure 10. To handle the open vocabulary of diverse answers in this dataset, we propose a training procedure based on a contrastive loss between a video-question multi-modal transformer and an answer transformer. We introduce the zero-shot VideoQA task and the VideoQA feature probe evaluation setting and show excellent results, in particular for rare answers. Furthermore, our method achieves competitive results on MSRVTT-QA, ActivityNet-QA, MSVD-QA and How2QA datasets. We also show that our VideoQA dataset generation approach generalizes to another source of web video and text data. We use our method to generate the WebVidVQA3M dataset from the WebVid dataset (videos with alt-text



**Speech:** Fold them in half again, to make a triangle.

**Generated Question:** How do you make a triangle?
**Generated Answer:** Fold them in half again

**Alt-Text:** Runner in sportswear running at stadium.

**Generated Question:** Where is a runner in sportswear running?
**Generated Answer:** Stadium

Figure 10: Given videos with transcribed narration (left) or videos with "alt-text" annotations (right), we leverage language models and cross-modal supervision to obtain large-scale VideoQA data. Top: frame with the corresponding text annotation. Bottom: automatically generated question and answer pair.

annotations) and show its benefits for training VideoQA models. Finally, for a detailed evaluation we introduce iVQA, a new VideoQA dataset with reduced language bias and high-quality manual annotations. Code, datasets and trained models are available at this address. This work was published at TPAMI [5].

### 8.1.12 Zero-Shot Video Question Answering via Frozen Bidirectional Language Models

**Participants:** Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, Cordelia Schmid.

Video question answering (VideoQA) is a complex task that requires diverse multi-modal data for training. Manual annotation of questions and answers for videos, however, is tedious and prohibits scalability. To tackle this problem, recent methods consider zero-shot settings with no manual annotation of visual question-answer. In particular, a promising approach adapts *frozen autoregressive* language models pretrained on Web-scale text-only data to multi-modal inputs. In contrast, we here build on *frozen bidirectional* language models (BiLM) and show that such an approach provides a stronger and cheaper alternative for zero-shot VideoQA. In particular, (i) we combine visual inputs with the frozen BiLM using light trainable modules, (ii) we train such modules using Web-scraped multi-modal data, and finally (iii) we perform zero-shot VideoQA inference through masked language modeling, where the masked text is the answer to a given question, see Figure 11. Our proposed approach, FrozenBiLM, outperforms the state of the art in zero-shot VideoQA by a significant margin on a variety of datasets, including LSMDC-FiB, iVQA, MSRVTT-QA, MSVD-QA, ActivityNet-QA, TGIF-FrameQA, How2QA and TVQA. It also demonstrates competitive performance in the few-shot and fully-supervised setting. Our code and models are publicly available at this address. This work was presented at NeurIPS'22 [23].

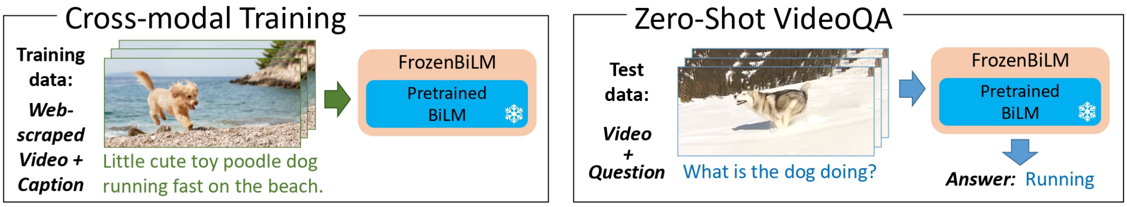


Figure 11: Our model FrozenBiLM builds on a pretrained and frozen bidirectional language model (BiLM), and is trained from Web-scraped video-caption pairs. FrozenBiLM excels in the zero-shot video question answering task without using any explicit visual question-answer supervision.

## 8.2 Learning embodied representations

### 8.2.1 ProxQP: Yet another Quadratic Programming Solver for Robotics and beyond

**Participants:** Antoine Bambade, Sarah El-Kazdadi, Adrien Taylor, Justin Carpentier.

In this paper [25], we introduce ProxQP, an open-source and state-of-the-art quadratic programming (QP) software with modular API that implements a new optimization algorithm based on primal dual augmented Lagrangian method. ProxQP is designed for robotic applications but not only. Indeed, quadratic programming has become a core modelling component in the modern engineering toolkit. This is particularly true for simulation, planning and control in robotics. Yet, modern numerical solvers have not reached the level of efficiency and reliability required in practical applications where speed, robustness, and accuracy are all necessary. Hence, along with our method, we present a benchmark studying the practical performances of modern optimization solvers for convex QPs on generic and complex problems of the literature as well as on common robotic scenarios. This benchmark notably highlights that this approach outperforms modern solvers in terms of efficiency, accuracy and robustness for small to medium-sized problems, while remaining competitive for higher dimensions. For example, Figure 12 shows performance profiles of different state-of-the-art solvers on a set of very hard QP problems from the literature (proposed by both industry and academia). We can see that ProxQP demonstrates the best performance profile and is thus both more accurate, robust and quick.

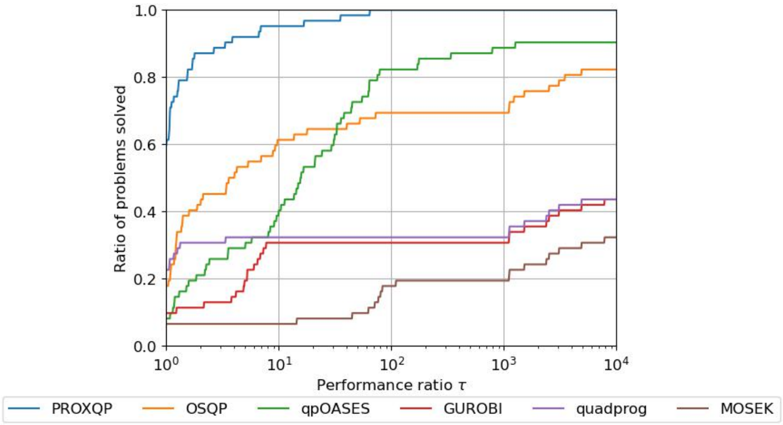


Figure 12: Performance profiles on small to medium-sized Maros-Mészàros problems, using a Core i5 - 5300U - 5th Generation @ 2,3 GHz processor. The higher the better.

### 8.2.2 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. However, these approaches assume the problems are well-posed, limiting de facto their application to ill-posed problems. In [37], we focus on the differentiation of optimal control solvers widely used in robotics. We notably introduce a differentiable proximal

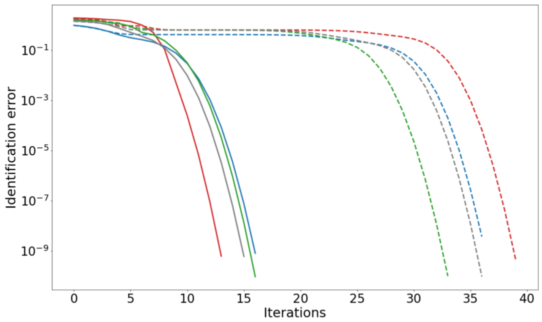


Figure 13: Identification error on identifying the dynamics matrices. Pairs of same color curves are identification experiments on the same problem parameters solved using different solvers: diff-mpc in dashed-lines, and ours in solid lines.

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 equality-constrained 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 (Fig. 13).

### 8.2.3 Assembly Planning from Observations under Physical Constraints

**Participants:** Thomas Chabal, Robin Strudel, Etienne Arlaud, Jean Ponce, Cordelia Schmid.

Our paper [26] addresses the problem of copying an unknown assembly of primitives with known shape and appearance, such as the one shown in Figure 14, using information extracted from a single photograph by an off-the-shelf procedure for object detection and pose estimation. The proposed algorithm uses a simple combination of physical stability constraints, convex optimization and Monte Carlo tree search to plan assemblies as sequences of pick-and-place operations represented by STRIPS operators. It is efficient and, most importantly, robust to the errors in object detection and pose estimation unavoidable in any real robotic system. The proposed approach is demonstrated with thorough experiments on a UR5 manipulator.

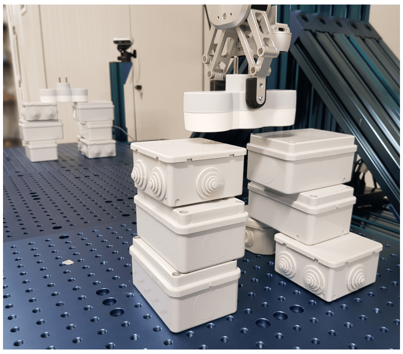


Figure 14: Example result of our approach. A UR5 manipulator assembles a configuration of known primitives (right - foreground). The configuration is specified by a single photograph of the assembly (camera on the left - background).

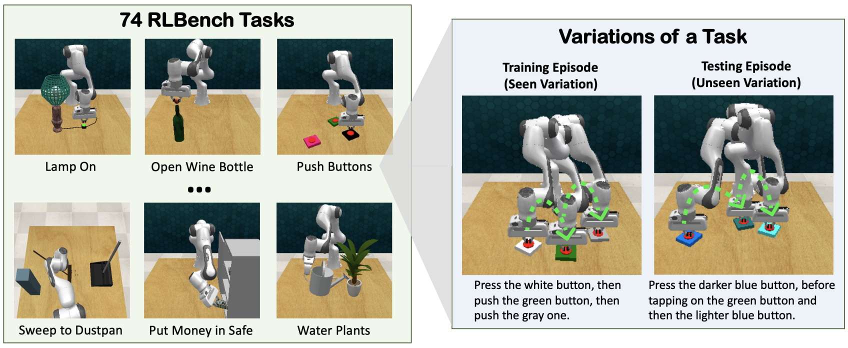

74 RLBench Tasks
Lamp On
Open Wine Bottle
Push Buttons
Sweep to Dustpan
Put Money in Safe
Water Plants
Variations of a Task
Training Episode (Seen Variation)
Testing Episode (Unseen Variation)
Press the white button, then push the green button, then push the gray one.
Press the darker blue button, before tapping on the green button and then the lighter blue button.

Figure 15: Hiveformer can adapt to simultaneously perform 74 tasks from RLBench given language instructions. Note that tasks can have multiple variations, such as the push buttons task. We test our model on unseen variations on such tasks.

### 8.2.4 Instruction-driven history-aware policies for robotic manipulations

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

In human environments, robots are expected to accomplish a variety of manipulation tasks given simple natural language instructions. Yet, robotic manipulation is extremely challenging as it requires fine-grained motor control, long-term memory as well as generalization to previously unseen tasks and environments. To address these challenges, we propose a unified transformer-based approach that takes into account multiple inputs in [27]. In particular, our transformer architecture integrates (i) natural language instructions and (ii) multi-view scene observations while (iii) keeping track of the full history of observations and actions. Such an approach enables learning dependencies between history and instructions and improves manipulation precision using multiple views. We evaluate our method on the challenging RLBench benchmark and on a real-world robot. Notably, our approach scales to 74 diverse RLBench tasks and outperforms the state-of-the-art, as illustrated on Figure 15. We also address instruction-conditioned tasks and demonstrate excellent generalization to previously unseen variations.

### 8.2.5 ProxNLP: a primal-dual augmented Lagrangian solver for nonlinear programming in Robotics and beyond

**Participants:** Wilson Jallet, Antoine Bambade, Justin Carpentier, Nicolas Mansard.

Mathematical optimization is the workhorse behind several aspects of modern robotics and control. In these applications, the focus is on constrained optimization, and the ability to work on manifolds (such as the classical matrix Lie groups), along with a specific requirement for robustness and speed. In recent years, augmented Lagrangian methods have seen a resurgence due to their robustness and flexibility, their connections to (inexact) proximal-point methods, and their interoperability with Newton or semismooth Newton methods. In our work [18], we present a primal-dual augmented Lagrangian method for inequality-constrained problems on manifolds, which we introduced in our recent work, as well as an efficient C++ implementation suitable for use in robotics applications and beyond.

### 8.2.6 Constrained Differential Dynamic Programming: A primal-dual augmented Lagrangian approach

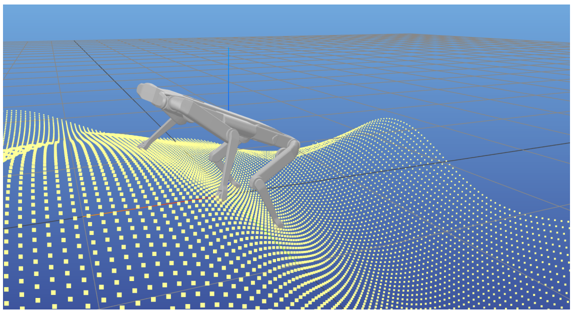


Figure 16: Static pose generated on the Solo robot using ProxNLP.

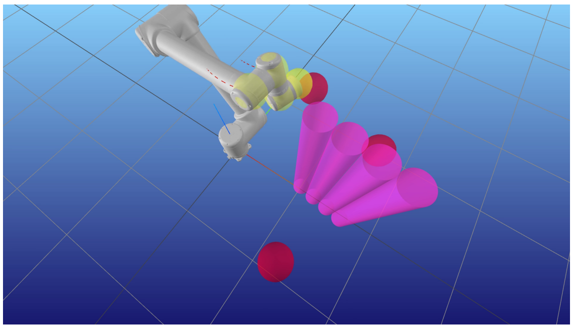


Figure 17: Pick-and-place task on the UR5 arm with obstacle avoidance.

**Participants:** Wilson Jallet, Antoine Bambade, Justin Carpentier, Nicolas Mansard.

Trajectory optimization is an efficient approach for solving optimal control problems for complex robotic systems. It relies on two key components: first the transcription into a sparse nonlinear program, and second the corresponding solver to iteratively compute its solution. On one hand, differential dynamic programming (DDP) provides an efficient approach to transcribe the optimal control problem into a finite-dimensional problem while optimally exploiting the sparsity induced by time. On the other hand, augmented Lagrangian methods make it possible to formulate efficient algorithms with advanced constraint-satisfaction strategies. In our work [17], we propose to combine these two approaches into an efficient optimal control algorithm accepting both equality and inequality constraints. Based on the augmented Lagrangian literature, we first derive a generic primal-dual augmented Lagrangian strategy for nonlinear problems with equality and inequality constraints. We then apply it to the dynamic programming principle to solve the value-greedy optimization problems inherent to the backward pass of DDP, which we combine with a dedicated globalization strategy, resulting in a Newton-like algorithm for solving constrained trajectory optimization problems. Contrary to previous attempts of formulating an augmented Lagrangian version of DDP, our approach exhibits adequate convergence properties without any switch in strategies. We empirically demonstrate its interest with several case-studies from the robotics literature.

#### 8.2.7 Implicit Differential Dynamic Programming

**Participants:** Wilson Jallet, Justin Carpentier, Nicolas Mansard.

Over the past decade, the Differential Dynamic Programming (DDP) method has gained in maturity and popularity within the robotics community. Several recent contributions have led to the integration of constraints within the original DDP formulation, hence enlarging its domain of application while making it a strong and easy-to-implement competitor against alternative methods of the state of the art such as collocation or multiple-shooting approaches. Yet, and similarly to its competitors, DDP remains unable to cope with high-dimensional dynamics within a receding horizon fashion, such as in the case of online generation of athletic motion on humanoid robots. In our work [19], we propose to make a step towards this objective by reformulating classical DDP as an implicit optimal control problem, allowing the use of more advanced integration schemes such as implicit or variational integrators. To that end, we introduce a primal-dual proximal Lagrangian approach capable of handling dynamical and path constraints in a unified manner, while taking advantage of the time sparsity inherent to optimal control problems. We show that this reformulation enables us to relax the dynamics along the optimization process by solving it inexactly: far from the optimality conditions, the dynamics are only partially fulfilled, but continuously enforced as the solver gets closer to the local optimal solution. This inexactness enables our approach to robustly handle large time steps (100 ms or more), unlike other DDP solvers of the state of the art, as experimentally validated through different robotic scenarii.

### 8.2.8 Stagewise Newton Method for Dynamic Game Control with Imperfect State Observation

**Participants:** Armand Jordana, Bilal Hammoud, Justin Carpentier, Ludovic Righetti.

In our work [1], we study dynamic game optimal control with imperfect state observations and introduce an iterative method to find a local Nash equilibrium. The algorithm consists of an iterative procedure combining a backward recursion similar to minimax differential dynamic programming and a forward recursion resembling a risksensitive Kalman smoother. A coupling equation renders the resulting control dependent on the estimation. In the end, the algorithm is equivalent to a Newton step but has linear complexity in the time horizon length. Furthermore, a merit function and a line search procedure are introduced to guarantee convergence of the iterative scheme. The resulting controller reasons about uncertainty by planning for the worst case disturbances. Lastly, the low computational cost of the proposed algorithm makes it a promising method to do output-feedback model predictive control on complex systems at high frequency. Numerical simulations on realistic robotic problems illustrate the risk-sensitive behavior of the resulting controller.

### 8.2.9 On the Derivation of the Contact Dynamics in Arbitrary Frames: Application to Polishing with Talos

**Participants:** Sébastien Kleff, Justin Carpentier, Nicolas Mansard, Ludovic Righetti.

Contact dynamics relies on the simultaneous satisfaction of constraints at the robot body level and at the contact level. At both levels, various formulations can be chosen that all must lead to the same results, given the same hypothesis, hence the little importance of their details. Yet when using it in an optimal control problem, a particular formulation is often imposed by the task to be performed by the robot. In our work [28], we detail the formulation of the contact quantities (force, movement) in an arbitrary frame imposed by the task. In that case, we will show that we are typically not interested in working in the local frame (attached to the robot contact point), nor in the world frame, but in a user-defined frame centered at the contact location with a fixed orientation in the world. The derivations can then be used for 6D, 3D or normal (pure-sliding) contact. We implemented the corresponding derivatives on top of the contact dynamics of Pinocchio in the optimal control solver Crocoddyl. We show that a unique formulation is able to handle several operational orientations by achieving several surfacing tasks in model predictive control with the robot Talos.

### 8.2.10 Single-view robot pose and joint angle estimation via render & compare

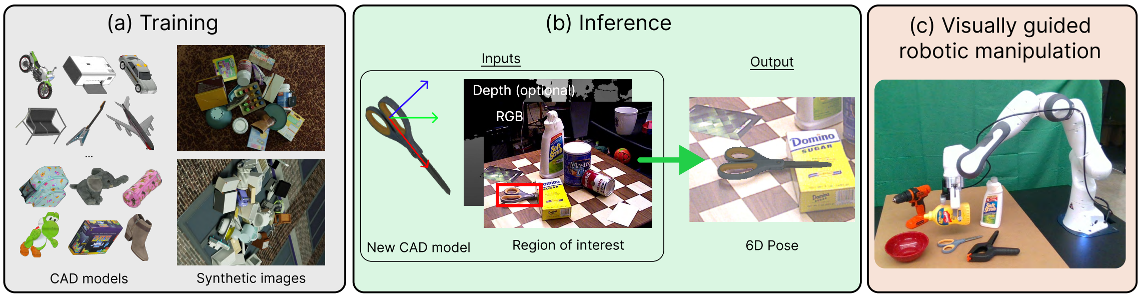


Figure 18: **MegaPose** is a 6D pose estimation approach (a) that is trained on millions of synthetic scenes with thousands of different objects and (b) can be applied *without re-training* to estimate the pose of any novel object, given a CAD model and a region of interest displaying the object. It can thus be used to rapidly deploy visually guided robotic manipulation systems in novel scenes containing novel objects (c).

**Participants:** Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, Josef Sivic.

In [29], we introduce MegaPose, a method to estimate the 6D pose of novel objects, that is, objects unseen during training. At inference time, the method only assumes knowledge of (i) a region of interest displaying the object in the image and (ii) a CAD model of the observed object. The contributions of this work are threefold. First, we present a 6D pose refiner based on a render-and-compare strategy, as we also explored in [33], which can be applied to novel objects. The shape and coordinate system of the novel object are provided as inputs to the network by rendering multiple synthetic views of the object's CAD model. Second, we introduce a novel approach for coarse pose estimation which leverages a network trained to classify whether the pose error between a synthetic rendering and an observed image of the same object can be corrected by the refiner. Third, we introduce a large-scale synthetic dataset of photorealistic images of thousands of objects with diverse visual and shape properties and show that this diversity is crucial to obtain good generalization performance on novel objects. We train our approach on this large synthetic dataset and apply it without retraining to hundreds of novel objects in real images from several pose estimation benchmarks. Our approach achieves state-of-the-art performance on the ModelNet and YCB-Video datasets. An extensive evaluation on the 7 core datasets of the BOP challenge demonstrates that our approach achieves performance competitive with existing approaches that require access to the target objects during training. Code, dataset and trained models are made available on the project webpage. The method is illustrated in figure 18.

#### 8.2.11 Augmenting differentiable physics with randomized smoothing

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

In the past few years, following the differentiable programming paradigm, there has been a growing interest in computing the gradient information of physical processes (e.g., physical simulation, image rendering). However, such processes may be non-differentiable or yield uninformative gradients (i.d., null almost everywhere). When faced with the former pitfalls, gradients estimated via analytical expression or numerical techniques such as automatic differentiation and finite differences, make classical optimization schemes converge towards poor quality solutions. Thus, relying only on the local information provided by these gradients is often not sufficient to solve advanced optimization problems involving such physical processes, notably when they are subject to non-smoothness and non-convexity issues. In our work [30], inspired by the field of zero-th order optimization, we leverage randomized smoothing

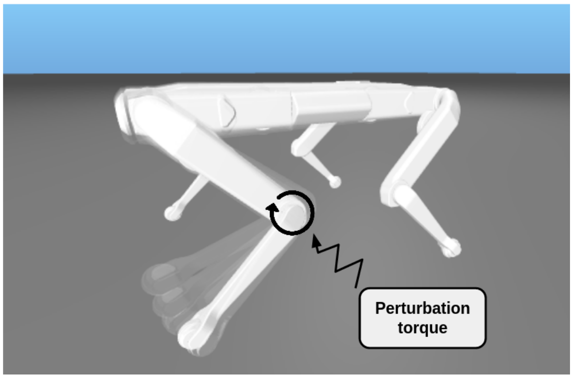


Figure 19: Illustration of randomized smoothing effects on the front left leg of the Solo robot.

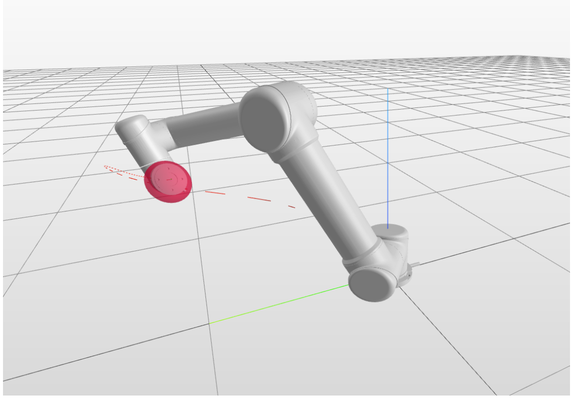


Figure 20: After training, a rollout of the policy leads to precise control on a test problem.

to augment differentiable physics by estimating gradients in a neighborhood. Our experiments suggest that integrating this approach inside optimization algorithms may be fruitful for tasks as varied as mesh reconstruction from images or optimal control of robotic systems subject to contact and friction issues.

#### 8.2.12 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 [20] 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.

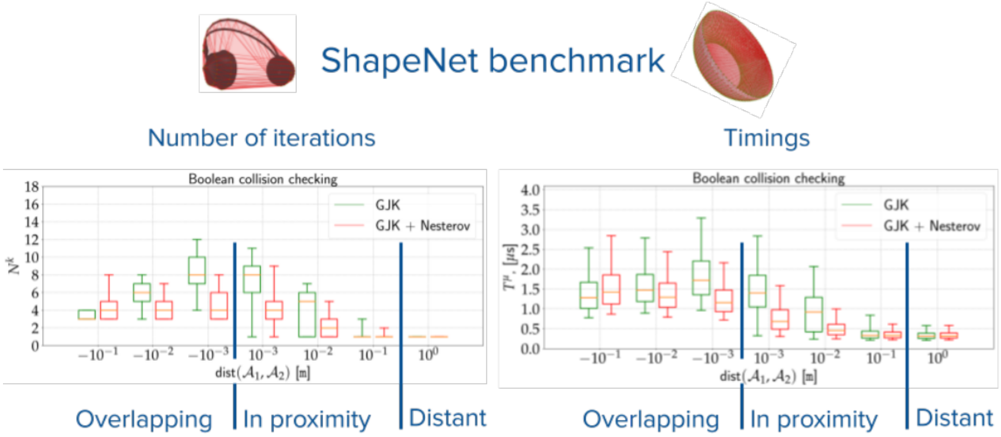


Figure 21: Nesterov-accelerated GJK vs. vanilla GJK on the ShapeNet benchmark. The ShapeNet benchmark is made of 1000 meshes for a total of 12 millions collision problems. We compare both the number of iterations and timings of our approach, Nesterov-accelerated GJK and the vanilla GJK algorithm (y-axis). The x-axis represents the distance between the objects. Our results show that Nesterov-accelerated GJK is up to two times faster than vanilla GJK. The acceleration is especially significative for scenarios important in physics simulations: when the objects of a collision pair are in proximity or in shallow interpenetration. (Left) Number of iterations. (Right) Timings.

### 8.2.13 Collision Detection Accelerated: An Optimization Perspective

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

Collision detection between two convex shapes is an essential feature of any physics engine or robot motion planner. It has often been tackled as a computational geometry problem, with the Gilbert, Johnson and Keerthi (GJK) algorithm being the most common approach today. In our work [32], we leverage the fact that collision detection is fundamentally a convex optimization problem. In particular, we establish that the GJK algorithm is a specific sub-case of the well-established Frank-Wolfe (FW) algorithm in convex optimization. We introduce a new collision detection algorithm by adapting recent works linking Nesterov acceleration and Frank-Wolfe methods. We benchmark the proposed accelerated collision detection method on two datasets composed of strictly convex and non-strictly convex shapes. Our results show that our approach significantly reduces the number of iterations to solve collision detection problems compared to the state-of-the-art GJK algorithm, leading to up to two times faster computation times.

### 8.2.14 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 [44], 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

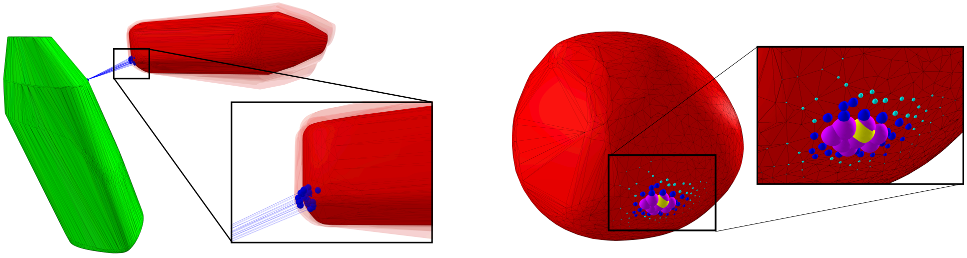


Figure 22: Illustration of randomized smoothing approximation on meshes from the YCB dataset. Left: $0^{th}$ order estimator, using a Gaussian distribution with 25 samples. Right: $1^{st}$ order estimator using a Gumbel distribution.

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.

### 8.2.15 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 [24], 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.

### 8.2.16 Learning to Manipulate Tools by Aligning Simulation to Video Demonstration

**Participants:** Kateryna Zorina, Justin Carpentier, Josef Sivic, Vladimír Petrík.

A seamless integration of robots into human environments requires robots to learn how to use existing human tools. Current approaches for learning tool manipulation skills mostly rely on expert demonstrations provided in the target robot environment, for example, by manually guiding the robot manipulator or by teleoperation. In our work [6], we introduce an automated approach that replaces an expert demonstration with a Youtube video for learning a tool manipulation strategy. The main contributions are twofold. First, we design an alignment procedure that aligns the simulated environment with the realworld scene observed in the video. This is formulated as an optimization problem that finds a spatial alignment of the tool trajectory to maximize the sparse goal reward given by the environment.

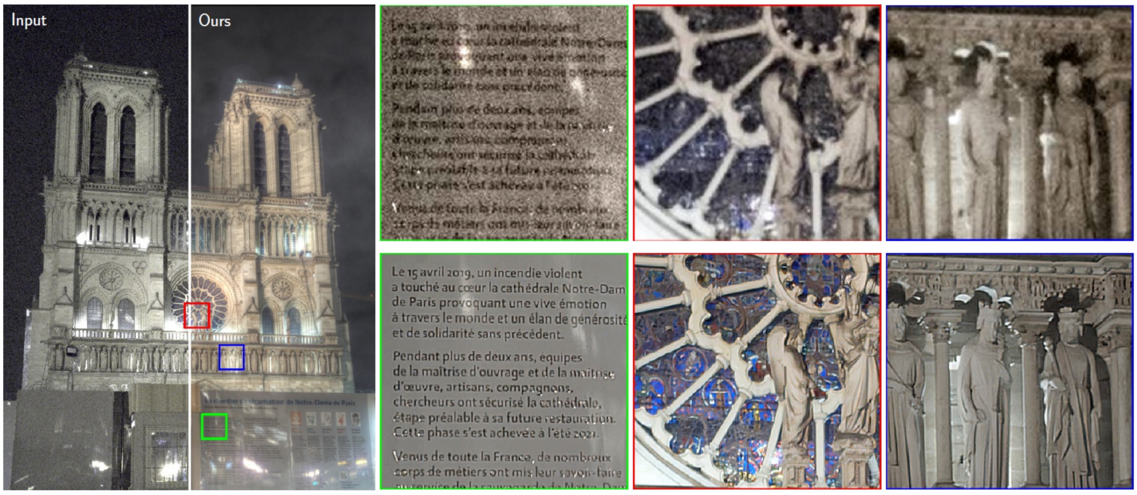

Input
Ours
Le 15 avril 2019, un incendie violent
a touché au cœur la cathédrale Notre-Dam
de Paris provoquant une vive émotion
à travers le monde et un élan de générosité
et de solidarité sans précédent.
Pendant plus de deux ans, équipes
de la maîtrise d'ouvrage et de la maîtrise
d'œuvre, artisans, compagnons,
chercheurs ont sécurisé la cathédrale,
étape préalable à sa future restauration.
Cette phase s'est achevée à l'été 2021.
Venus de toute la France, de nombreux
corps de métiers ont mis leur savoir-faire

Figure 23: An example of joint super-resolution (SR) and high-dynamic range (HDR) imaging. Left: An 18-photo burst was shot at night from a hand-held Pixel 4a smartphone at 12MP resolution with an exposure time varying from 1/340s to 1/4s. The left half of the central image from the burst is shown along with the right half of the 192MP HDR image reconstructed by our algorithm with a super-resolution factor of ×4 (after tone mapping). Right: Three small crops of the two images corresponding to the colored square regions on the left.

Second, we devise an imitation learning approach that focuses on the trajectory of the tool and how it interacts with the environment, rather than the motion of the human. We demonstrate the proposed approach on spade, scythe and hammer tools in simulation, and show the effectiveness of the trained policy for the spade on a real Franka Emika Panda robot demonstration.

## 8.3 Image restoration and enhancement

### 8.3.1 High Dynamic Range and Super-Resolution from Raw Image Bursts

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

Photographs captured by smartphones and mid-range cameras have limited spatial resolution and dynamic range, with noisy response in underexposed regions and color artefacts in saturated areas. In [2] we introduce the first approach (to the best of our knowledge) to the reconstruction of high-resolution, high-dynamic range color images from raw photographic bursts captured by a handheld camera with exposure bracketing. This method uses a physically-accurate model of image formation to combine an iterative optimization algorithm for solving the corresponding inverse problem with a learned image representation for robust alignment and a learned natural image prior. The proposed algorithm is fast, with low memory requirements compared to state-of-the-art learning-based approaches to image restoration, and features that are learned end to end from synthetic yet realistic data. Extensive experiments demonstrate its excellent performance with super-resolution factors of up to ×4 on real photographs taken in the wild with hand-held cameras, and high robustness to low-light conditions, noise, camera shake, and moderate object motion.

# 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.

## 9.2 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 Associate team GAYA

**Participants:** Jean Ponce, Cordelia Schmid.

GAYA is a joint research team bringing together two Inria project-teams (Thoth, Grenoble and WILLOW, Paris) and Carnegie Mellon University, USA. It focuses on two research themes: (i) semantic structured interpretation of videos, and (ii) studying the geometric properties of object shapes to enhance state-of-the-art object recognition approaches.

Interpreting videos semantically in a general setting, involving various types of video content like home video clips, news broadcasts, feature films, which contain a lot of clutter, non-rigid motion, many “actors” performing actions, person-object and person-person interactions, varying viewpoints, is challenging. This task is being examined increasingly over the past decade, with the availability of large video resources, e.g., YouTube. Despite this progress, an effective video representation for recognizing actions is still missing. To address this critical challenge, we propose a joint optimization framework, wherein we learn the video representation and also develop models for action recognition. Specifically, we aim to exploit the spatio-temporal relations among pixels in a video through graphical models and novel deep learning feature representations.

The second research theme explores geometric aspects of computer vision, in particular how to model three-dimensional objects from their two-dimensional projections, and how the appearance of these objects evolves with changes in viewpoint. Beyond its theoretical interest, this work is critical for developing object recognition algorithms that take into account the three-dimensional nature of the visual world and go beyond the template-matching approaches dominant today. Duality is an important concept in this area, and we are investigating its application to the construction of visual hulls as well as the characterization of the topology of image contours using the Gauss map. Existing results are essentially limited to the Euclidean setting, and we are investigating their generalization to the general projective case.

Partners: CMU (Deva Ramanan, Martial Hebert, Abhinav Gupta, Pavel Tokmakov), INRIA Thoth (Karteek Alahari).

## 10.2 International research visitors

Ludovic Righetti (NYU) was a visiting researcher at Willow during July 2022. Sébastien Kleff and Armand Jordana were visiting PhD students from NYU in August and September respectively. Moreover, J. Ponce spends half of his time at New York University. Willow PhD students Oumayma Bounou, Bruno Lecouat and Thomas Chabal as well as Alexandre Araujo visited NYU during 2022. We also continue international collaboration with Y. LeCun (Meta/NYU), A. Miech and J.-B. Alayrac (Deepmind) and J. Sivic (CTU Prague).

## 10.3 European initiatives

### 10.3.1 IMPACT: Intelligent machine perception

**Participants:** Josef Sivic, Jean Ponce, Ivan Laptev.

IMPACT is a collaborative project with Czech Technical University, Center for Robotics, Informatics and Cybernetics (CIIRC) (2017-2023). The IMPACT project focuses on fundamental and applied research in computer vision, machine learning and robotics to develop machines that learn to perceive, reason, navigate and interact with complex dynamic environments. For example, people easily learn how to change a flat tire of a car or perform resuscitation by observing other people doing the same task. This involves advanced visual intelligence abilities such as interpreting sequences of human actions that manipulate objects to achieve a specific task. Currently, however, there is no artificial system with a similar level of cognitive visual competence. Breakthrough progress in intelligent machine perception will have profound implications on our everyday lives as well as science and commerce, with smart assistive robots that automatically learn new skills from the Internet, safer cars that autonomously navigate in difficult changing conditions, or intelligent glasses that help people navigate never seen before environments.

### 10.3.2 AGIMUS: Next generation of AI-powered robotics for agile production

**Participants:** Justin Carpentier.

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.

AGIMUS is collaborative project with CNRS (France), AIRBUS (France), KLEEMANN HELLAS SA (Greece), PAL ROBOTICS (Spain), Q-PLAN INTERNATIONAL (Greece), TOWARD SAS (France), THIMM OBALY, K.S. (Czechia) and CVUT (Czech Republic).

## 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

# 11 Dissemination

## 11.1 Promoting scientific activities

### 11.1.1 Scientific events: organisation

**General chair, scientific chair**

- General chair for the International Conference on Computer Vision 2023 (J. Ponce, C. Schmid).
- Program chair for the International Conference on Computer Vision 2023 (I. Laptev).

- Workshop in honor of Jean-Paul Laumond, Collège de France, July 2022 (J. Ponce).

**Member of the organizing committees**

- Co-organization of the 7th International Workshop on Recovering 6D Object Pose, ECCV 2022 (Y. Labbé).

### 11.1.2 Scientific events: selection

**Chair of conference program committees**

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 (I. Laptev, AC).
- European Conference on Computer Vision (ECCV), 2022 (I. Laptev, C. Schmid, AC).
- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 (C. Schmid, senior AC).

**Member of the conference program committees**

- ACM Multimedia, 2022 (S. Chen).

**Reviewer**

- AAAI Conference on Artificial Intelligence (AAAI), 2022 (S. Chen).
- Conference on Empirical Methods in Natural Language Processing (EMPL), 2022 (S. Chen).
- Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2022 (S. Chen).
- European Conference on Computer Vision (ECCV), 2022 (S. Chen, Z. Chen, P.-L. Guhur, Y. Labbé, J. Sivic, A. Yang).
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022 (S. Chen, P.-L. Guhur, Y. Labbé, J. Sivic, A. Yang).
- NeurIPS 2022 Track Datasets and Benchmarks (J. Sivic).
- IEEE-RAS International Conference on Robotics and Automation, 2022 (M. Alakuijala, Y. Labbé).
- IEEE-RAS International Conference on Humanoid Robots (W. Jallet).
- IEEE/RSJ International Conference on Intelligent Robots, 2022 (W. Jallet, Y. Labbé).
- International Conference on Machine Learning, 2022 (H. Cisneros).

### 11.1.3 Journal

**Member of the editorial boards**

- Associate Editor, IEEE Transactions on Robotics (J. Carpentier).
- Associate Editor, IEEE Robotics and Automation Letters (J. Carpentier).
- Associate Editor, International Journal of Computer Vision (I. Laptev).
- Senior Editor-in-Chief, International Journal of Computer Vision (J. Ponce).

**Reviewer - reviewing activities**

- IEEE Robotics and Automation Letters (M. Alakuijala, W. Jallet, Y. Labbé).
- IEEE Transactions on Robotics (W. Jallet, S. Chen, Y. Labbé).
- IEEE Transactions on Multimedia (S. Chen).
- International Journal of Computer Vision (A. Bardes, Y. Labbé).
- Machine Learning, Springer (A. Bardes).

#### 11.1.4 Invited talks

- O. Bounou, NYU Tandon School of Engineering, 2022.
- O. Bounou, NYU Center for Data Science, 2022.
- J. Carpentier, ENNSD Workshop, Toulouse, France.
- J. Carpentier, NAFEMS online workshop.
- J. Carpentier, Humanoids 2022 Tutorial on Challenge-driven Learning of Humanoid Robot Control in Virtual Environments, remote presentation (Japan), December 2022.
- T. Chabal, LAAS-CNRS (Gepetto team), Toulouse, April 2022.
- S. Chen, AIMind Group Renmin University of China, July 2022.
- S. Chen, Microsoft Research Asia. March 2022.
- S. Chen, IMAGINE/LIGM École des Ponts ParisTech, December 2022.
- S. Chen, Stanford Vision and Learning Lab, September 2022.
- Z. Chen, Human Body, Hands, and Activities from Egocentric and Multi-view Cameras workshop (HBHA@ECCV22).
- Z. Chen, Observing and Understanding Hands in Action workshop (HANDS@ECCV22).
- W. Jallet, RAINBOW team, Inria Rennes, 2022.
- W. Jallet, Journées nationales de la robotique humanoïde, Angers, 2022.
- W. Jallet, NYU Tandon School of Engineering, 2022.
- I. Laptev, Int. Workshop on AI for Visual Computing, Pohang, 2022
- I. Laptev, Journée Visage, Gestes, Actions et Comportement, Paris, 2022
- I. Laptev, Andrew Zisserman Festschrift, Oxford, 2022.
- I. Laptev, Video Understanding Symposium, Amsterdam, 2022.
- I. Laptev, Machine and Deep Learning for Plenoptics, Sundsvall, 2022.
- J. Ponce, Andrew Zisserman Festschrift, 2022.
- J. Ponce, NYU Robotics Lab, 2022.
- C. Schmid, NeurIPS 2022 Workshop on Vision Transformers: Theory and Applications, virtual, December 2022.
- C. Schmid, TrecVid 2022, Keynote talk, virtual, December 2022.

- C. Schmid, ACCV 2022, Keynote talk, virtual, December 2022.
- C. Schmid, 2nd edition of the 3IA Doctoral Workshop, Keynote talk, Grenoble, November 2022.
- C. Schmid, Seminar at MPI Tubingen, October 2022.
- C. Schmid, Czech-French AI Workshop, Prague, September 2022.
- C. Schmid, BIFOLD opening event, Berlin, September 2022.
- C. Schmid, Long-form Video Understanding Workshop, in conjunction with CVPR'22, virtual, June 2022.
- C. Schmid, 5th Multimodal Learning and Applications Workshop, in conjunction with CVPR'22, virtual, June 2022
- C. Schmid, ICLR 2022, Keynote talk, virtual, April 2022.
- C. Schmid, Stanford University HAI Spring Conference, virtual, April 2022.
- J. Sivic, European Big Data Value Association Forum, Prague, 2022.
- J. Sivic, CZ-US workshop on manufacturing, Prague, 2022.
- J. Sivic, Czech-French AI workshop, Prague, 2022.
- J. Sivic, Salesforce Research, US, 2022.
- A. Yang, NeurIPS 2022, Poster presentation, November 2022.
- A. Yang, NeurIPS@Paris 2022, November 2022.
- A. Yang, Seminar of the Computer Science department of École Normale Supérieure (Mûr-de-Bretagne, France), June 2022.

#### 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).
- Director of Ellis program on Computer Vision and Machine Learning, (C. Schmid).
- Global Member of the Bavarian AI Council (J. Sivic).
- Member of the advisory board, Computer Vision Foundation (J. Sivic).
- Member of the board of directors of the Computer Vision Foundation (CVF) (C. Schmid).
- Member of the Milner award committee (C. Schmid).
- Member of the PAMI-TC executive committee (C. Schmid).
- Member of the PAMI-TC awards committee (C. Schmid).
- Member of the scientific advisory board for the German Competence Centers for AI Research (C. Schmid).
- Member of the Scientific Advisory Committee of the Helmholtz AI Cooperation Unit (C. Schmid).
- Member of the Technical Activities Board for International Foundation of Robotics Research (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: Data science and business analytics, École des Mines de Lyon, 2022 (P.-L. Guhur).
- Master: Computer vision and time series analysis for Physics and Engineering 2022, 6h (I. Laptev, A. Yang).
- Master: Introduction à la vision artificielle, M1, École normale supérieure Paris (J. Ponce, O. Bounou, I. Laptev).
- Master: Introduction to computer vision to medical students, 15h (J. Ponce).
- 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: Reinforcement learning, EPITA, Paris, 2022 (P.-L. Guhur).
- Master: Robotics courses at ENS-DI, 30h (J. Carpentier).
- Master: Robotics courses at Formation des Inge‘nieurs de l'Automobile, PSL, 10h (J. Carpentier).
- Master: Three lectures (3 x 1.5h) in the 3D computer vision class of V. Hlavac at Charles University in Prague (J. Sivic).
- Tutorial: Practical courses in deep learning, Spring School on Data Science, Ecole Centrale Casablanca, June 2022, 6h (T. Chabal).
- Bachelor: Oral examinations (colles) at Lycée Chaptal (O. Bounou).
- Bachelor: Oral examinations (colles) at Lycée Marcelin Berthelot (A. Yang).

### 11.2.2 Supervision

- PhD in progress: Minttu Alakuijala, graduated in Dec 2022 [34], J. Ponce and C. Schmid.
- PhD in progress: Alaaeldin Ali, started in Aug. 2020, I. Laptev and H. Jégou (Meta).
- 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 in progress: Elliot Chane-Sane, started in Oct. 2020, 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 in progress: Pierre-Louis Guhur, started in Oct 2019, I. Laptev and C. Schmid.
- PhD in progress: Wilson Jallet, started in Oct 2021, J. Carpentier and N. Mansard.
- PhD in progress: Yann Labbe, started in Oct 2018, 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 in progress: Bruno Lecouat, started in Sept 2019, J. Ponce and J. Mairal (Inria Grenoble).
- PhD in progress: Zongmian Li, started in Oct 2017, 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 in progress: Robin Strudel, started in Oct 2018, I. Laptev, C. Schmid and J. Sivic.
- PhD in progress: Vo Van Huy, graduated in Nov 2022 [35], J. Ponce.
- 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 in progress: Antoine Yang, started in Oct. 2020, I. Laptev, C. Schmid and J. Sivic.
- Mentor at the Doctoral Consortium, in conjunction with CVPR 2022, C. Schmid.

#### 11.2.3 Juries

- Minttu Alakuijala, École Normale Supérieure, November 2022 (J. Ponce, C. Schmid, PhD committee).
- Eloïse Berthier, PSL University, (J. Carpentier, PhD committee)
- Rui Dai, Côte d'Azur University, September 2022 (I. Laptev, PhD committee, président du jury).
- Valentin Gabeur, Université Grenoble Alpes (J. Sivic, PhD committee, rapporteur).
- Shruti Palaskar, Carnegie Mellon University, April 2022 (C. Schmid, PhD committee).
- Pierre Schegg, Inria Lille, May 2022 (J. Carpentier, PhD committee).
- Hugo Touvron, Sorbonne Université, September 2022 (C. Schmid, PhD committee).
- Huy Vo, École Normale Supérieure, November 2022 (J. Ponce, C. Schmid, PhD committee).

## 11.3 Popularization

### 11.3.1 Internal or external Inria responsibilities

- P.-L. Guhur, "Mathématiques, numériques et climat", Inria, 2022.
- P.-L. Guhur, Semaine du climat at Inria, September 2022.
- P.-L. Guhur, Workshop facilitation for the Climate Fresk (x15) and MyCO2 (x4), 2022.

### 11.3.2 Articles and contents

- "L'intelligence artificielle est imparfaite, et alors ?", Carte blanche au *Monde*, 12 October 2022 (J. Ponce).
- Interviews for L'Express (J. Carpentier).
- Interviews for Usine Nouvelle (J. Carpentier).

# 12 Scientific production

## 12.1 Publications of the year

### International journals

[1] A. Jordana, B. Hammoud, J. Carpentier and L. Righetti. 'Stagewise Newton Method for Dynamic Game Control with Imperfect State Observation'. In: *IEEE Control Systems Letters* (20th June 2022). DOI: 10.1109/LCSYS.2022.3184657. URL: https://hal.inria.fr/hal-03705557.

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

[3] Z. Li, J. Sedlar, J. Carpentier, I. Laptev, N. Mansard and J. Sivic. 'Estimating 3D Motion and Forces of Human-Object Interactions from Internet Videos'. In: *International Journal of Computer Vision* 130.2 (Feb. 2022), pp. 363–383. DOI: 10.1007/s11263-021-01540-1. URL: https://hal.science/hal-03420419.

[4] X. Shen, R. Champenois, S. Ginosar, I. Pastrolin, M. Rousselot, O. Bounou, T. Monnier, S. Gidaris, F. Bougard, P.-G. Raverdy, M.-F. Limon, C. Bénévent, M. Smith, O. Poncet, K. Bender, B. Joyeux-Prunel, E. Honig, A. Efros and M. Aubry. 'Spatially-consistent Feature Matching and Learning for Heritage Image Analysis'. In: *International Journal of Computer Vision* (25th Mar. 2022). DOI: 10.1007/s11263-022-01576-x. URL: https://hal.science/hal-03620996.

[5] A. Yang, A. Miech, J. Sivic, I. Laptev and C. Schmid. 'Learning to Answer Visual Questions from Web Videos'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022). DOI: 10.1109/tpami.2022.3173208. URL: https://hal.inria.fr/hal-03664182.

[6] K. Zorina, J. Carpentier, J. Sivic and V. Petrík. 'Learning to Manipulate Tools by Aligning Simulation to Video Demonstration'. In: *IEEE Robotics and Automation Letters* (3rd Jan. 2022). DOI: 10.48550/arXiv.2111.03088. URL: https://hal.inria.fr/hal-03478117.

### International peer-reviewed conferences

[7] 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. URL: https://hal.inria.fr/hal-03997549.

[8] A. Bardes, J. Ponce and Y. Lecun. 'VICReg: Variance-Invariance-Covariance Regularization For Self-Supervised Learning'. In: ICLR 2022 - International Conference on Learning Representations. Online, United States, 25th Apr. 2022. URL: https://hal.inria.fr/hal-03541297.

[9] A. Bardes, J. Ponce and Y. Lecun. 'VICRegL: Self-Supervised Learning of Local Visual Features'. In: 36th Conference on Neural Information Processing Systems (NeurIPS 2022). New Orleans, United States, 28th Nov. 2022. URL: https://hal.inria.fr/hal-03893126.

[10] E. Berthier, J. Carpentier, A. Rudi and F. Bach. 'Infinite-Dimensional Sums-of-Squares for Optimal Control'. In: 61st IEEE Conference on Decision and Control. Cancun, Mexico, 6th Dec. 2022. URL: https://hal.science/hal-03377120.

[11] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid and I. Laptev. 'Language Conditioned Spatial Relation Reasoning for 3D Object Grounding'. In: NeurIPS 2022 - 36th Conference on Neural Information Processing Systems. New Orleans, United States, 28th Nov. 2022. URL: https://hal.inria.fr/hal-03890174.

[12] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid and I. Laptev. 'Learning from Unlabeled 3D Environments for Vision-and-Language Navigation'. In: ECCV 2022 - European Conference on Computer Vision. Tel Aviv-Jaffa, Israel, 23rd Oct. 2022. URL: https://hal.inria.fr/hal-03890196.

[13] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid and I. Laptev. 'Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation'. In: CVPR 2022 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, United States, 19th June 2022. URL: https://hal.inria.fr/hal-03696868.

[14] Z. Chen, Y. Hasson, C. Schmid and I. Laptev. 'AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction'. In: ECCV 2022 - European Conference on Computer Vision. Tel Aviv-Jaffa, Israel, 23rd Oct. 2022. URL: https://hal.inria.fr/hal-03761124.

[15] H. Cisneros, J. Sivic and T. Mikolov. 'Benchmarking Learning Efficiency in Deep Reservoir Computing'. In: CoLLAs 2022 - Conference on Lifelong Learning Agents. Montreal, Canada, 22nd Aug. 2022. URL: https://hal.inria.fr/hal-03790477.

[16] O. Flasseur, T. Bodrito, J. Mairal, J. Ponce, M. Langlois and A.-M. Lagrange. 'Exoplanet detection in angular differential imaging: combining a statistics-based learning with a deep-based learning for improved detections'. In: Adaptive Optics Systems VIII. Montréal, Canada: SPIE, 17th July 2022, p. 139. DOI: 10.1117/12.2629849. URL: https://hal.science/hal-03988124.

[17] W. Jallet, A. Bambade, N. Mansard and J. Carpentier. 'Constrained Differential Dynamic Programming: A primal-dual augmented Lagrangian approach'. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, Japan, Oct. 2022. URL: https://hal.science/hal-03597630.

[18] W. Jallet, A. Bambade, N. Mansard and J. Carpentier. 'ProxNLP: a primal-dual augmented Lagrangian solver for nonlinear programming in Robotics and beyond'. In: *Proceedings of the 6th Legged Robots Workshop*. 6th Legged Robots Workshop. Philadelphia, Pennsylvania, United States, 27th May 2022. URL: https://hal.science/hal-03680510.

[19] W. Jallet, N. Mansard and J. Carpentier. 'Implicit Differential Dynamic Programming'. In: International Conference on Robotics and Automation (ICRA 2022). Philadelphia, United States, May 2022. DOI: 10.1109/ICRA46639.2022.9811647. URL: https://hal.science/hal-03351641.

[20] 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, 29th May 2023. URL: https://hal.science/hal-03780392.

[21] T. Souček, J.-B. Alayrac, A. Miech, I. Laptev and J. Sivic. 'Look for the Change: Learning Object States and State-Modifying Actions from Untrimmed Web Videos'. In: CVPR 2022 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, United States, 19th June 2022. URL: https://hal.inria.fr/hal-03996825.

[22] A. Yang, A. Miech, J. Sivic, I. Laptev and C. Schmid. 'TubeDETR: Spatio-Temporal Video Grounding with Transformers'. In: CVPR 2022 - IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, United States, 19th June 2022. URL: https://hal.inria.fr/hal-03625586.

[23] A. Yang, A. Miech, J. Sivic, I. Laptev and C. Schmid. 'Zero-Shot Video Question Answering via Frozen Bidirectional Language Models'. In: NeurIPS 2022 - 36th Conference on Neural Information Processing Systems. New Orleans, United States, 28th Nov. 2022. URL: https://hal.inria.fr/hal-03807016.

[24] 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://hal.laas.fr/hal-03945110.

### Conferences without proceedings

[25] A. Bambade, S. El-Kazdadi, A. Taylor and J. Carpentier. 'PROX-QP: Yet another Quadratic Programming Solver for Robotics and beyond'. In: RSS 2022 - Robotics: Science and Systems. New York, United States, June 2022. URL: https://hal.inria.fr/hal-03683733.

[26] T. Chabal, R. Strudel, E. Arlaud, J. Ponce and C. Schmid. 'Assembly Planning from Observations under Physical Constraints'. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, Japan, 23rd Oct. 2022. URL: https://hal.science/hal-03647973.

[27] 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.

[28] S. Kleff, J. Carpentier, N. Mansard and L. Righetti. 'On the Derivation of the Contact Dynamics in Arbitrary Frames: Application to Polishing with Talos'. In: Humanoids 2022 - IEEE-RAS International Conference on Humanoid Robots. Okinawa, Japan, 28th Nov. 2022. URL: https://hal.science/hal-03758989.

[29] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox and J. Sivic. 'MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare'. In: CoRL 2022 - Conference on Robot Learning. Auckland, New Zealand, 14th Dec. 2022. URL: https://hal.science/hal-03910329.

[30] Q. Le Lidec, L. Montaut, C. Schmid, I. Laptev and J. Carpentier. 'Augmenting differentiable physics with randomized smoothing'. In: RSS 2022 - Robotics Science and Systems, Workshop on Differentiable Simulation For Robotics. New York, United States, 27th June 2022. URL: https://hal.science/hal-03703324.

[31] R. Loiseau, B. Bouvier, Y. Teytaut, E. Vincent, M. Aubry and L. Landrieu. 'A Model You Can Hear: Audio Identification with Playable Prototypes'. In: ISMIR 2022 - 23rd International Society for Music Information Retrieval Conference. Bengaluru, India, 4th Dec. 2022. URL: https://hal.science/hal-03794815.

[32] L. Montaut, Q. Le Lidec, V. Petrik, J. Sivic and J. Carpentier. 'Collision Detection Accelerated: An Optimization Perspective'. In: RSS 2022 - Robotics: Science and Systems. New York, United States, 27th June 2022. URL: https://hal.science/hal-03662157.

[33] G. Ponimatkin, Y. Labbé, B. Russell, M. Aubry and J. Sivic. 'Focal Length and Object Pose Estimation via Render and Compare'. In: CVPR 2022 - IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, United States, 19th June 2022. URL: https://hal.science/hal-03847201.

### Doctoral dissertations and habilitation theses

[34] M. Alakuijala. 'Self-taught Robots: Autonomous and Weakly-Supervised Learning for Robotic Manipulation'. ENS Paris - Ecole Normale Supérieure de Paris, 13th Dec. 2022. URL: https://hal.science/tel-04001370.

[35] H. V. Vo. 'Annotation-efficient learning for object discovery and detection'. Ecole normale supérieure - ENS PARIS, 28th Nov. 2022. URL: https://hal.science/tel-03919952.

### Reports & preprints

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

[37] O. Bounou, J. Carpentier and J. Ponce. *Leveraging Proximal Optimization for Differentiating Optimal Control Solvers*. 23rd Sept. 2022. URL: https://hal.science/hal-03786820.

[38] C. Debeunne, M. Fourmy, Y. Labbé, P.-A. Léziart, G. Saurel, J. Solà and N. Mansard. *CosySlam: investigating object-level SLAM for detecting locomotion surfaces*. 3rd Mar. 2022. URL: https://hal.science/hal-03351438.

[39] M. Futeral, C. Schmid, I. Laptev, B. Sagot and R. Bawden. *Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation*. 20th Dec. 2022. URL: https://hal.inria.fr/hal-03977982.

[40] Q. Garrido, Y. Chen, A. Bardes, L. Najman and Y. Lecun. *On the duality between contrastive and non-contrastive self-supervised learning*. 2nd Oct. 2022. DOI: 10.48550/arXiv.2206.02574. URL: https://hal.science/hal-03685169.

[41] Q. Le Lidec, L. Montaut, C. Schmid, I. Laptev and J. Carpentier. *Leveraging Randomized Smoothing for Optimal Control of Nonsmooth Dynamical Systems*. 11th Mar. 2022. DOI: 10.48550/arXiv.2203.03986. URL: https://hal.science/hal-03480419.

[42] G. L. Moing, J. Ponce and C. Schmid. *WALDO: Future Video Synthesis using Object Layer Decomposition and Parametric Flow Prediction*. 8th Dec. 2022. URL: https://hal.inria.fr/hal-03889664.

[43] 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://hal.inria.fr/hal-03856704.

[44] L. Montaut, Q. Le Lidec, A. Bambade, V. Petrík, J. Sivic and J. Carpentier. *Differentiable Collision Detection: a Randomized Smoothing Approach*. 29th Sept. 2022. URL: https://hal.science/hal-03780482.

[45] A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jégou and E. Grave. *Are Large-scale Datasets Necessary for Self-Supervised Pre-training?* 14th Feb. 2022. URL: https://hal.science/hal-03572721.

[46] A. El-Nouby, N. Neverova, I. Laptev and H. Jégou. *Training Vision Transformers for Image Retrieval*. 14th Feb. 2022. URL: https://hal.science/hal-03572734.

[47] C. Pieters, E. Damblon, P. Souères and J.-P. Laumond. *Talking about moving machines: an argumentative perspective*. 3rd Sept. 2022. URL: https://hal.laas.fr/hal-03768385.