Avis de Soutenance

Monsieur Alexandre CHAPIN

Soutiendra publiquement ses travaux de thèse intitulés

Apprentissage latent détaché pour la manipulation précise en robotique

Travaux dirigés par Monsieur Liming CHEN et Emmanuel DELLANDREA

Soutenance prévue le ***mercredi 04 novembre 2026*** à 10h00

Lieu : Ecole Centrale de Lyon, 36 Av. Guy de Collongue, 69130 Écully

Salle : Amphi 1bis

Composition du jury proposé

M. Liming CHEN	Professeur	Ecole Centrale de Lyon	Directeur de thèse
M. Ngoc Son VU	Professeur	Université de Technologie de Troyes	Rapporteur
M. Eric MARCHAND	Professeur	Université de Rennes	Examinateur
M. Emmanuel DELLANDREA	Maître de conférences HDR	Ecole Centrale de Lyon	Co-directeur de thèse
Mme Céline TEULIÈRE	Professeure	Université Clermont Auvergne	Examinatrice
M. Cédric DEMONCEAUX	Professeur des universités	IUT Le Creusot, Université Bourgogne Europe	Rapporteur

Mots-clés : apprentissage profond, vision par ordinateur, manipulation robotique, espace latent

Résumé :

Cette thèse étudie le rôle des représentations visuelles dans les capacités d'apprentissage et de généralisation des politiques de manipulation robotique. Alors que les méthodes de l'état de l'art s'appuient généralement sur des caractéristiques globales ou denses extraites de modèles de fondation visuelle à grande échelle, ces représentations mélangent souvent les informations pertinentes avec le bruit issu de l'arrière-plan. Pour pallier cette limite, nous proposons une transition vers les représentations centrées sur l'objet (OCR), une alternative structurée qui décompose les scènes visuelles en « slots » discrets et interactifs. Nos recherches établissent une progression en trois étapes vers une perception robotique généralisable. Premièrement, nous démontrons d'abord les avantages fondamentaux des biais inductifs centrés sur l'objet par rapport aux représentations holistiques dans les scènes multi-objets, en soulignant leur robustesse supérieure face aux couleurs et textures perturbatrices. Deuxièmement, nous introduisons les représentations centrées sur l'objet basées sur des slots en robotique réelle, à travers une comparaison systématique à grande échelle sur des environnements simulés et réels. Nous démontrons que le pré-entraînement de ces modèles sur des ensembles de données robotiques établit une base perceptive robuste, surpassant considérablement les caractéristiques conventionnelles. Cependant, nous identifions un compromis critique entre capacité et généralisation lors de l'introduction de nouveaux objets distracteurs, nécessitant un équilibre précis entre le nombre de slots alloués et les performances finales. Pour surmonter ce compromis, nous proposons trois cadres spécifiques. Premièrement, nous présentons STORM, un module léger qui ancre sémantiquement les slots à l'aide d'instructions en langage naturel via une stratégie d'adaptation multi-phases. Deuxièmement, nous introduisons Slot-RAE, un modèle génératif de diffusion centré sur l'objet utilisant des auto-encodeurs de représentation directs. Cette approche permet non seulement de simplifier le processus d'apprentissage en évitant les lourds biais génératifs, mais elle débloque également de véritables capacités compositionnelles directement au sein de l'espace des caractéristiques continues des modèles de fondation visuelle (VFM). Enfin, à titre de travail exploratoire, nous présentons ARMOR, un pipeline conceptuel d'augmentation de données dans l'espace latent qui extrait des tokens invariants aux permutations à partir d'images extérieures et les injecte dynamiquement pour renforcer la robustesse des politiques robotiques face au bruit visuel. Ces travaux posent les bases d'une perception structurée et symbolique dans l'apprentissage robotique, comblant ainsi le fossé entre les entrées visuelles de bas niveau et le raisonnement neuro-symbolique de haut niveau.

Summary:

This thesis investigates the role of visual representations in the learning and generalization capabilities of robotic manipulation policies. While state-of-the-art methods typically rely on global or dense features extracted from large-scale visual foundation models, these representations often entangle task-relevant information with irrelevant background clutter. We address this limitation by proposing a shift toward Object-Centric Representations (OCR), a structured alternative that decomposes visual scenes into discrete, interactive "slots". Our research establishes a three-stage progression toward generalizable robotic perception. First, we demonstrate the foundational benefits of object-centric inductive biases over holistic representations in multi-object scenes, highlighting their superior robustness to distractor colors and textures. Second, we introduce object-centric representations to real-world robotics, conducting a large-scale systematic comparison across both simulated and real-world benchmarks. We demonstrate that pre-training these models on large-scale robotic datasets establishes a robust perceptual foundation. However, we reveal new insights regarding the behavior of these models when introducing novel distractor objects, identifying a critical capacity-generalization trade-off: a delicate balance must be struck between the number of allocated slots and the final performance in both in-domain and generalized settings. To overcome this trade-off, we propose three specific frameworks. We present STORM, a lightweight module that semantically grounds slots using natural language instructions via a novel multi-phase adaptation strategy. We also introduce Slot-RAE, a generative object-centric diffusion model utilizing direct representation auto-encoders. This framework not only streamlines the learning process by bypassing heavy generative biases, but it also unlocks true compositional capacities natively within the continuous feature space of visual foundation models (VFMs). Finally, as an ongoing exploratory direction, we outline ARMOR, a compute-efficient latent-space data augmentation pipeline that extracts permutation-invariant tokens from diverse, task-irrelevant images and dynamically injects them into the state representations to inoculate control policies against visual noise. This work lays the groundwork for structured, symbol-aware perception in robot learning, bridging the gap between low-level visual input and high-level neuro-symbolic reasoning.