Université Paris Cité · LIPADE Laboratoire d'Informatique PAris DEscartes · SIP Systèmes Intelligents de Perception
sylvainlobry.com
sip.mi.parisdescartes.fr

ACADEMIC POSITION · MASTER 2 INTERNSHIP

Self-Supervised Change Captioning in Remote Sensing

ChanGen Project

Computer Vision | Remote Sensing | Change Captioning | Generative Models | Self-Supervised Learning | Multi-modality

POSITION DETAILS

Duration	6 months (Feb. - Apr. 2027)
Gratification	Approx. 682.50 €/month
Institute	Université Paris Cité — LIPADE, Team SIP (*Systèmes Intelligents de Perception*)
Location	45 rue des Saints-Pères, 75006 Paris, France
Supervision	Sylvain Lobry
Status	Open until filled, first interviews in Dec.

APPLICATION PROCEDURE

Send an email to **sylvain.lobry@u-paris.fr** with subject:

`[MSc ChanGen] FirstName LastName`

Required documents:

- Updated CV
- Cover letter
- M1 & M2 transcripts (and/or contact info of a referee)

01 Introduction

Detecting and describing changes in remote sensing (RS) images is important for monitoring environmental evolution, tracking urban development, and supporting disaster response. In addition to detecting where changes occur, models that generate natural-language descriptions make these outputs easier to interpret for non-specialists and allow automated reporting in time-critical situations. Such textual descriptions can also serve as inputs to Large Language Models (LLMs) and conversational interfaces for querying geospatial changes.

However, current change captioning methods face two main limitations. First, they rely heavily on manually written text annotations, which restricts both the level of detail in the descriptions and the scale of available datasets. Second, standard benchmarks are mostly built from scarce and expensive Very High Resolution (VHR) images limited to bi-temporal (before-and-after) pairs. In emergency response scenarios, recent pre-event VHR imagery is often unavailable, and post-event acquisitions may be partially hidden by clouds or smoke. Lower-resolution time series, such as Synthetic Aperture Radar (SAR, e.g. Sentinel-1) and multi-spectral imagery (e.g. Sentinel-2), provide complementary temporal coverage to fill these gaps.

In the **ChanGen** project, we investigate a collaborative self-supervised framework in which a generated change description (i.e., the caption) is trained using the generation of an up-to-date VHR image as the supervision signal (Figure 1). By using the text caption as an information bottleneck, the goal is to learn a description that captures enough semantic information to reconstruct the target VHR image. The reconstructed image is then compared to the ground-truth VHR acquisition to provide a training signal in a cooperative optimization loop.

This internship will be carried out in the **SIP** team at **LIPADE** (Université Paris Cité), in collaboration with **Google** (with regular project meetings with Google's team). Computing resources are available through Université Paris Cité. The internship focuses on the first two stages of the project (Stage 1: dataset construction, and Stage 2: a bi-temporal baseline), with the aim of submitting a dataset and baseline paper to a remote sensing conference or workshop (e.g., IEEE IGARSS, CVPR EarthVision).

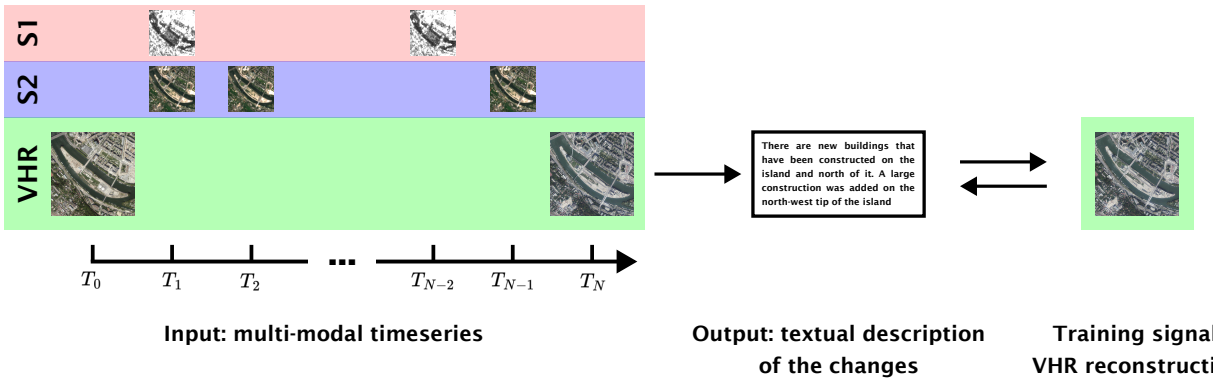

Figure 1: Overview of the ChanGen framework. Multi-sensor time series (Sentinel-1, Sentinel-2, and VHR) serve as input to generate a textual change description, which acts as an information bottleneck to guide synthetic VHR reconstruction.

02 Background

Multi-modal change detection has been studied in recent works [1, 2], mostly by optimizing change detection between two VHR modalities in a bi-temporal setting. Other recent approaches use flow-based generative models to capture changes [**rolih2026generative**, 3], but remain limited to bi-temporal, single-modality inputs and do not generate text. The *2Player* framework [4] uses a similar cooperative learning setup, though it outputs binary change detection maps rather than textual descriptions. Finally, *GroundSet* [5] showed that cadastral vector records can provide accurate annotations without manual labeling, making it possible to scale training data to large areas.

03 Objectives

The internship is structured around two main stages:

- **Stage 1 — Dataset creation:** Because no existing dataset supports this task, we will first collect and assemble the data needed for training and evaluation. The dataset will combine low-frequency VHR aerial imagery (**BD ORTHO**, acquired every three years) with **Sentinel-1** (SAR, every 6 days) and **Sentinel-2** (optical, every 5 days) time series to bridge the temporal gaps, together with ground-truth annotations from the French cadaster's **BD TOPO**. We will start from an existing codebase used to build GroundSet [5] and adapt it to collect, align, and structure this multi-modal dataset.
- **Stage 2 — Bi-temporal baseline:** Once a first version of the dataset is ready, the intern will develop and evaluate a first baseline demonstrating the feasibility of the cooperative change captioning framework in a simplified bi-temporal setting.

04 Desired Background for the Candidate

We are looking for a Master 2 (or final-year engineering) student in Computer Science, Machine Learning, or a related field. The ideal candidate should have a solid background in computer vision and deep learning, and be comfortable programming in Python (PyTorch). Prior experience with remote sensing data, generative models, or natural language processing is a plus.

About LIPADE, the SIP Team, and Sylvain Lobry's Group

The intern will join **Sylvain Lobry's research group** in the **SIP (*Systèmes Intelligents de Perception*)** team at **LIPADE** (*Laboratoire d'Informatique de Paris Descartes*), Université Paris Cité (45 rue des Saints-Pères, 75006 Paris). Founded in 1990 and led by Prof. Camille Kurtz, SIP works on computer vision, image analysis, and AI for remote sensing, biomedical imaging, and document analysis.

References

[1] A. Oubara et al. "Adversarial learning network for multi-source change detection in built-up areas: integrating VHR RGB and multispectral imagery". In: *International Journal of Digital Earth* 18.1 (2025), p. 2503443.

[2] H. Li et al. "Integrating bi-temporal VHR optical and long-term SAR images for built-up area change detection". In: *International Journal of Digital Earth* 17.1 (2024), p. 2316109.

[3] R. Kazoom, G. Leifman, and G. Beryozkin. "FM-ChangeNet: Learning Change through Pathwise Feature Transport". In: *MORSE Workshop @ CVPR*. 2026.

[4] M. Béchaz et al. "2Player: A general framework for self-supervised change detection via cooperative learning". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 232 (2026).

[5] R. Ferrod et al. "GroundSet: A Cadastral-Grounded Dataset for Spatial Understanding with Vector Data". In: *ECCV* (2026).