www.onera.fr

ONERA
THE FRENCH AEROSPACE LAB

## PROPOSITION DE SUJET DE THESE

**Intitulé : Développement d'un modèle fondamental Vision-Langage pour la compréhension automatique d'images SAR multi-sources**

Référence : **Domaine-Département-2026-12**
*(à rappeler dans toute correspondance)*

| **Début de la thèse : 01/2027** | **Date limite de candidature :** |
|---|---|

**Deep learning, Imagerie SAR, Visual-language model, Image synthétique, multimodalité**

**Profil et compétences recherchées :**

Master 2 ou Diplôme d'ingénieur avec spécialisation en traitement du signal et des images, intelligence artificielle, radar.
Compétences : Python, PyTorch, mathématiques appliquées, traitement d'images ; la connaissance de l'imagerie radar constitue un atout supplémentaire
Bon niveau rédactionnel, bon niveau d'anglais, expérience et goût pour la recherche appliqué.

**Développement d'un modèle fondamental Vision-Langage pour la compréhension automatique d'images SAR multi-sources**

Le département DEMR de l'ONERA est reconnu pour son expertise dans l'acquisition, la simulation et l'exploitation des données radar, en particulier en imagerie SAR (Synthetic Aperture Radar). Grâce à leur capacité à fonctionner de jour comme de nuit et indépendamment des conditions météorologiques, les capteurs SAR sont aujourd'hui utilisés dans de nombreux domaines civils et militaires, tels que la surveillance environnementale, la gestion de crises, la cartographie ou encore la détection et l'analyse d'activités d'intérêt. Cependant, l'interprétation des images SAR demeure délicate du fait de leurs mécanismes physiques de formation, qui compliquent l'établissement d'un lien direct entre l'information radiométrique et une représentation sémantique intuitive des scènes observées.

Ces dernières années, les modèles fondamentaux de type *Vision-Language Models* (VLM) [1], capables de relier automatiquement des images à des descriptions en langage naturel, ont montré des performances remarquables en imagerie optique. Ces modèles permettent non seulement la génération de descriptions textuelles, mais aussi une compréhension plus globale des scènes. Ils ouvrent la voie à de nouvelles formes d'interaction homme-machine et d'analyse automatique. Plus récemment, ces modèles ne se limitent plus à une simple traduction image–texte, mais utilisent le langage comme une représentation intermédiaire structurante permettant d'enrichir la compréhension sémantique des images [2]. Cette approche unifiée relie le contenu visuel à un large éventail de tâches telles que la génération et la modification d'images par le texte, la segmentation sémantique ou encore l'interprétation fine des scènes.

En revanche, de tels modèles restent quasiment inexistants pour l'imagerie SAR. Les approches actuelles se limitent majoritairement à des modèles spécialisés et mono-tâche (détection, classification, segmentation), entraînés sur des jeux de données restreints et fortement dépendants d'un capteur ou d'un contexte d'acquisition donné, ce qui limite fortement leur capacité de généralisation.

Cette thèse a pour objectif de développer un modèle vision-langage multimodal pour l'imagerie SAR, capable de produire et d'exploiter des descriptions en langage naturel pour une compréhension globale des images radar. L'une des principales limitations scientifiques réside dans la rareté de jeux de données SAR de grande taille, annotés, diversifiés et non redondants, adaptés à l'apprentissage de modèles vision-langage restent limitées. Les bases de données existantes sont aujourd'hui très riches mais présentent des limites [3][4]. Par ailleurs, les modèles récents proposés dans la littérature pour le SAR (SUMMIT [7], SARATR-X [6], etc.) ne considèrent pas la modalité langage et ne permettent pas une compréhension sémantique riche des scènes observées.

Au sein de l'ONERA, l'unité SEM du DEMR a développé le simulateur physique EMPRISE [8], capable de générer des images SAR de haute fidélité à partir de scènes numériques paramétrables, en respectant les lois de propagation électromagnétique. Ces données synthétiques offrent un contrôle précis sur les configurations géométriques, radiométriques et contextuelles, et constituent un levier majeur pour enrichir les jeux de données existants. Néanmoins, l'intégration conjointe de données réelles multi-capteurs et de données synthétiques pose des questions fondamentales liées à l'écart de domaine, à l'équilibre entre réalisme et diversité, et aux stratégies d'apprentissage adaptées pour éviter le sur-apprentissage du domaine synthétique.

Le travail de doctorat visera ainsi à étudier et formaliser des stratégies d'apprentissage vision–langage robustes pour l'imagerie SAR, en s'appuyant sur un mélange contrôlé de données réelles et synthétiques. Il s'articulera autour des axes suivants :

- Revue de littérature et état de l'art des travaux existants sur les différents VLM en SAR et les bases de données existantes.
- Mise en place d'un vaste jeu de données diversifié, combinant des données réelles (SETHI, UMBRA) et des données synthétiques issues d'EMPRISE ou de modèles génératifs récents, avec une standardisation géométrique et radiométrique rigoureuse.
- Etude de différentes stratégies hybrides d'entraînement des VLM SAR, telles que le pré-entraînement sur données synthétiques suivi d'un ajustement fin sur données réelles, le mélange pondéré de bases de données, ou encore l'apprentissage multi-domaines et multi-résolutions.
- Evaluations des modèles vision-langage appliqués au SAR. Les métriques classiques issues du traitement automatique du langage (BLEU, ROUGE, etc.) se révélant insuffisantes pour juger la pertinence de descriptions complexes de scènes radar, de nouvelles méthodologies d'évaluation seront étudiées.


[1] W. Wang et al. "CogVLM: Visual Expert for Pretrained Language Models," arXiv: 2311.03079, 2024.

[2] A. Yang et al. "Qwen3 Technical Report," arXiv: 2505.09388, 2025.

[3] Y. Wei et al., "Sarlang-1m: A benchmark for vision-language modeling in sar image understanding," arXiv:2504.03254, 2025.

[4] Z. Ma et al., "Sarchat-bench-2m: A multi-task vision-language benchmark for sar image interpretation," arXiv:2502.08168, 2025.

[5] Y. He et al., "Sar-text: A large-scale sar image-text dataset built with sar-narrator and progressive learning," arXiv:2507.18743, 2025.

[6] W. Li et al., "Saratr-x: Towards building a foundation model for sar target recognition," IEEE Trans. Image Process., 2025.

[7] Y. Du et al., "Summit: A sar foundation model with multiple auxiliary tasks," ISPRS J. Photogramm. Remote Sens., 2025.

[8] E. Everaere *et al*., "SAR imaging of complex environments with EMPRISE® simulation software," *EUSAR 2024; 15th European Conference on Synthetic Aperture Radar*, Munich, Germany, 2024, pp. 503-508.


**Collaborations envisagées**


| **Laboratoire d'accueil à l'ONERA** | **Directeur de thèse** |
|---|---|
| Département : DEMR | Nom : Trouvé |
| Lieu (centre ONERA) : Palaiseau | Laboratoire : ONERA/DEMR |
| **Contact** : Nathan Letheule | Tél. : +33 1 80 38 62 95 |
| Tél. : +33 1 80 38 72 23 | Email : nicolas.trouve@onera.fr |
| Email : nathan.letheule@onera.fr | |


Pour plus d'informations : https://www.onera.fr/rejoindre-onera/la-formation-par-la-recherche