Université Paris Cité

# Identification des Relations Spatiales dans les Images à l'aide de Méthodes Deep Learning

## Relations Spatiales SpatialSense++


**Réalisé par:**

- ABED Nada-Fatima Zohra

Université Paris Cité



20/06/2025

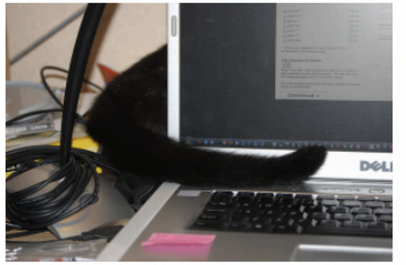
Université Paris Cité

20/06/2025

# Introduction

1

- Compréhension des relations spatiales : clé pour l'interprétation d'images.
- Applications : robotique, navigation visuelle, description automatique.

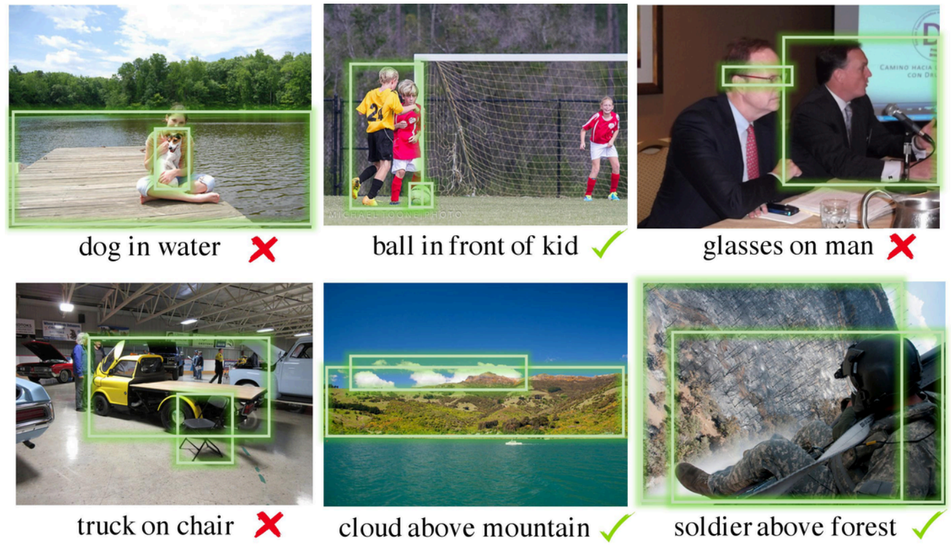
| *The cat is behind the laptop*. (True) | *The horse is left of the person*. (False) |
| --- | --- |
|  |  |

# Problématique

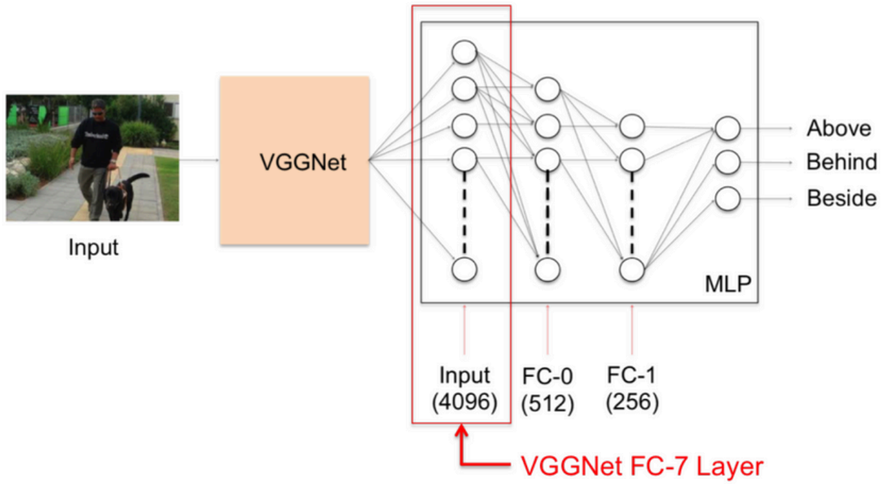
2

- **Problème global :** les relations spatiales sont complexes à modéliser car elles dépendent à la fois de la position, de l'échelle, de la perspective et du contexte sémantique.

- **Limite de l'approche classique :** elle repose uniquement sur l'image brute sans exploiter explicitement les informations géométriques ou relationnelles entre objets.



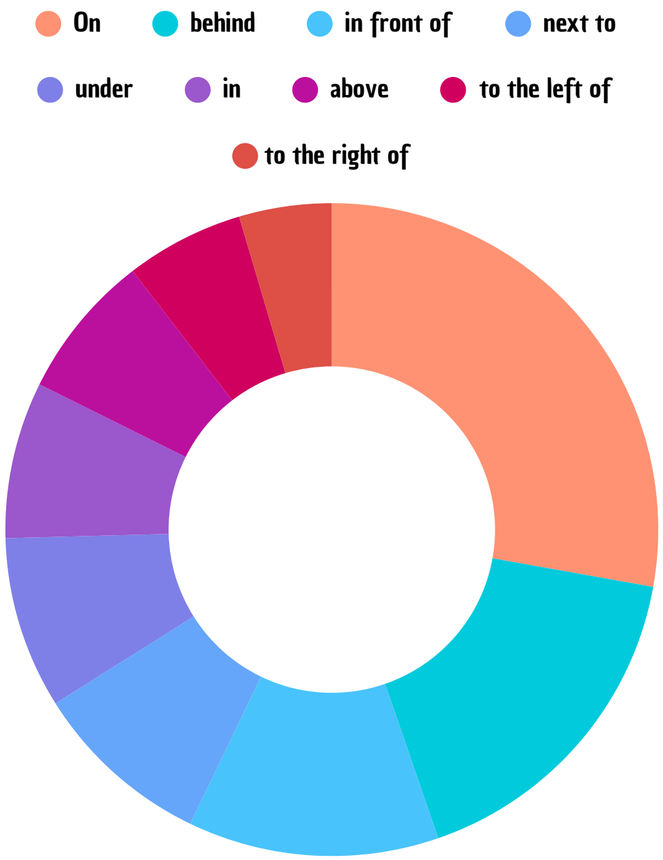
# Objectif




- Reproduire la méthode Haldekar (CNN sur image complète).
- Développer et évaluer des variantes multimodales (image, texte, features spatiales).

## SpatialSense++

- **Composition générale :**
  - Total d'images : 10 440
  - Total de relations annotées : 17 498
- **Vocabulaire spatial :**
  - Prédicats uniques : 9 relations spatiales
  - Objets uniques : 20 catégories d'objets
- **Système d'annotation :**
  - Label binaire : True/False pour chaque relation
  - Relations validées (label=True) : 8 749 (50%)
  - Relations rejetées (label=False) : 8 749 (50%)
- **Structure des données :**
  - Chaque relation contient :
    - Bounding boxes sujet/objet
    - Prédicat spatial
    - Label de validité
    - Coordonnées de centres



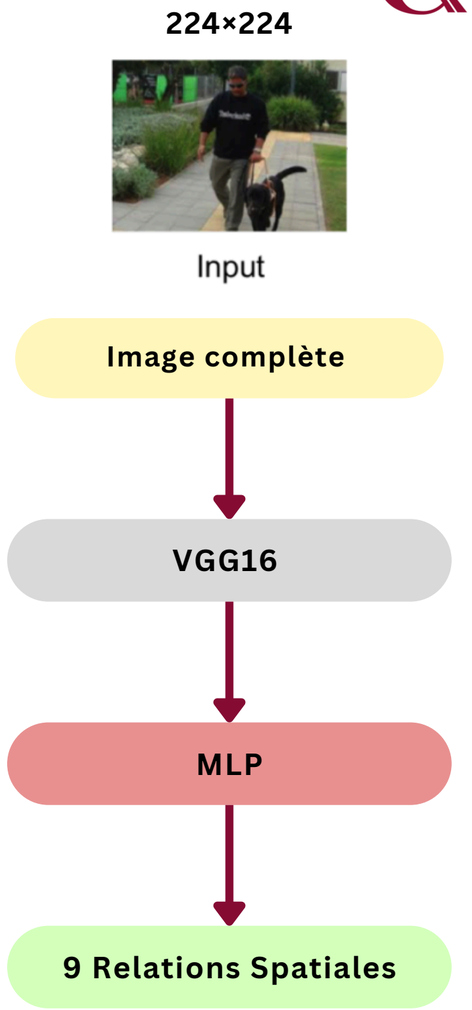

*Répartition des* relations spatiales après filtrage

# MÉTHODE 1: Haldekar (VGG16)

Université Paris Cité

5

- Entrée : Image complète 224×224
- Backbone : VGG16
- Features : 4096 dimensions
- MLP : 4096 -> 512 -> 256 → 9

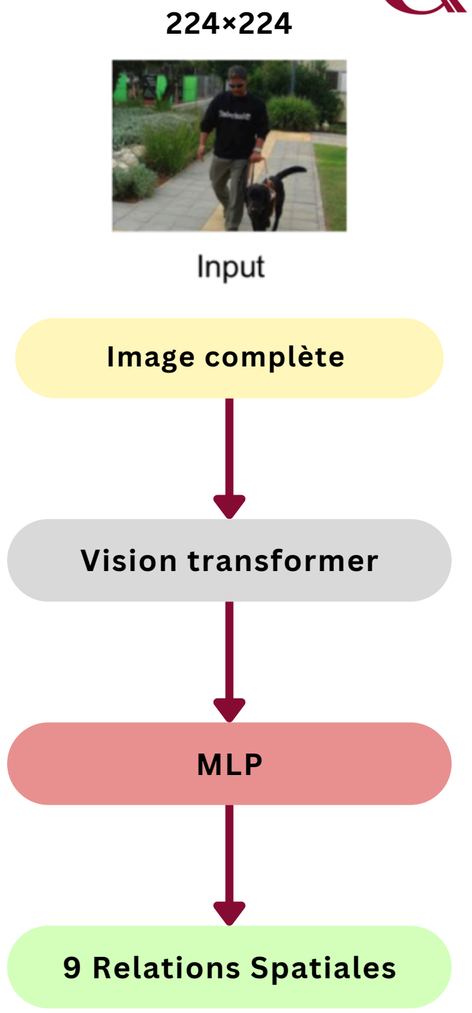


# MÉTHODE 2: Vision Transformer

Université Paris Cité

6

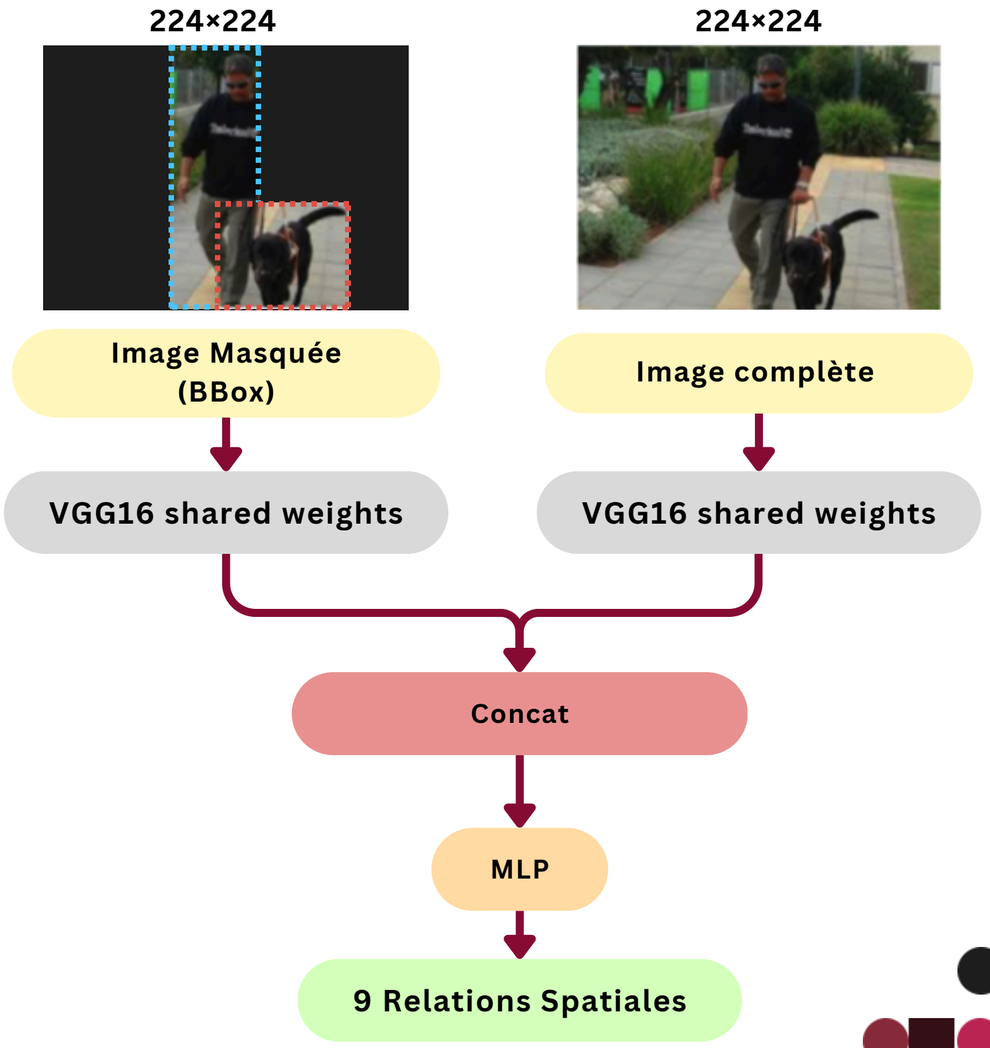
- Entrée : Image complète 224×224
- Tokenisation : Patches 16×16 (196 patches + CLS token)
- Backbone : Vision Transformer (ViT-Base)
- Features : 768 dimensions
- Classifier : MLP (768 → 512 → 256→9)



# MÉTHODE 3: Architecture Duale
## Image complète + Masking

7

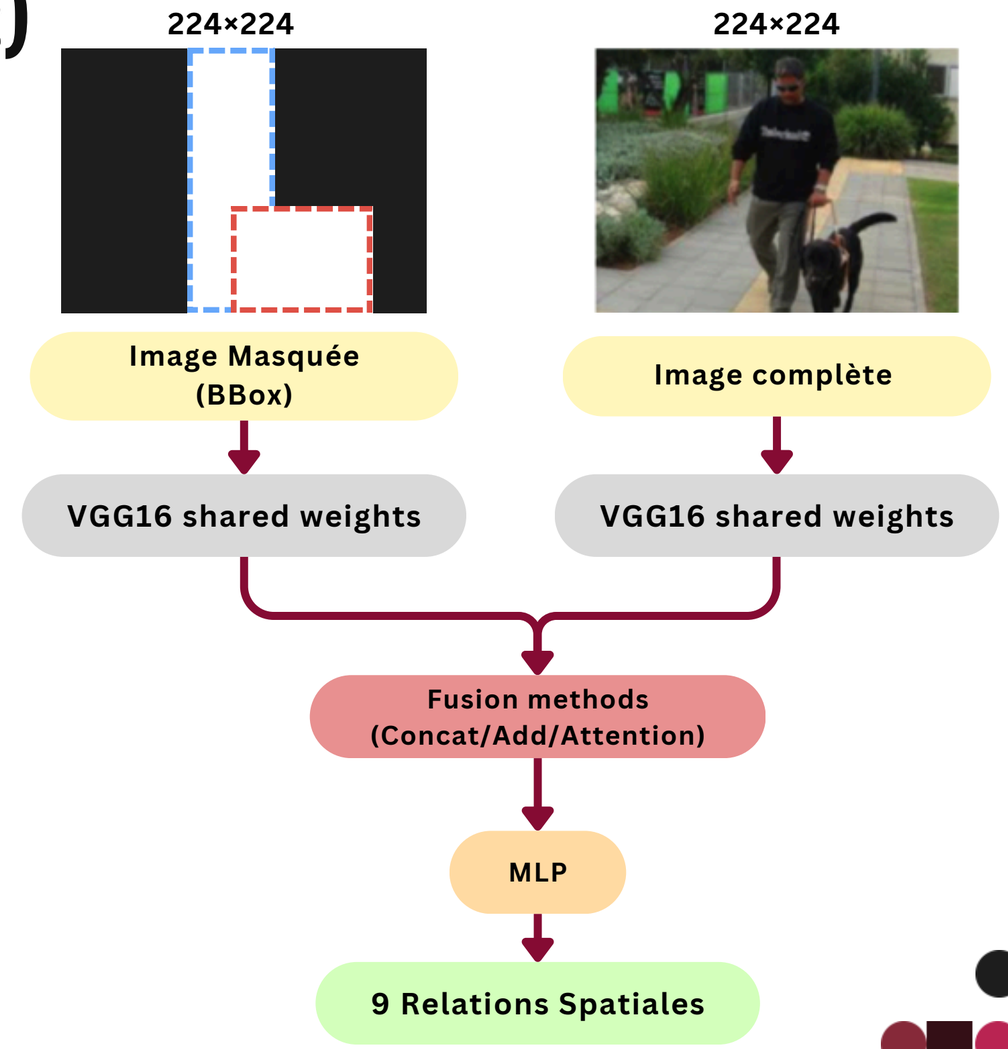
- Entrée 1: Image complète (RGB)
- Entrée 2: Masque la regios hors des bounding boxes en noir
- F usion: Concaténation des features VGG (4096 + 4096 = 8192)
- MLP 8192->512>256->9



# MÉTHODE 4: Architecture Duale
## Image complète + Masking (Noir ⊊ Blanc)

7

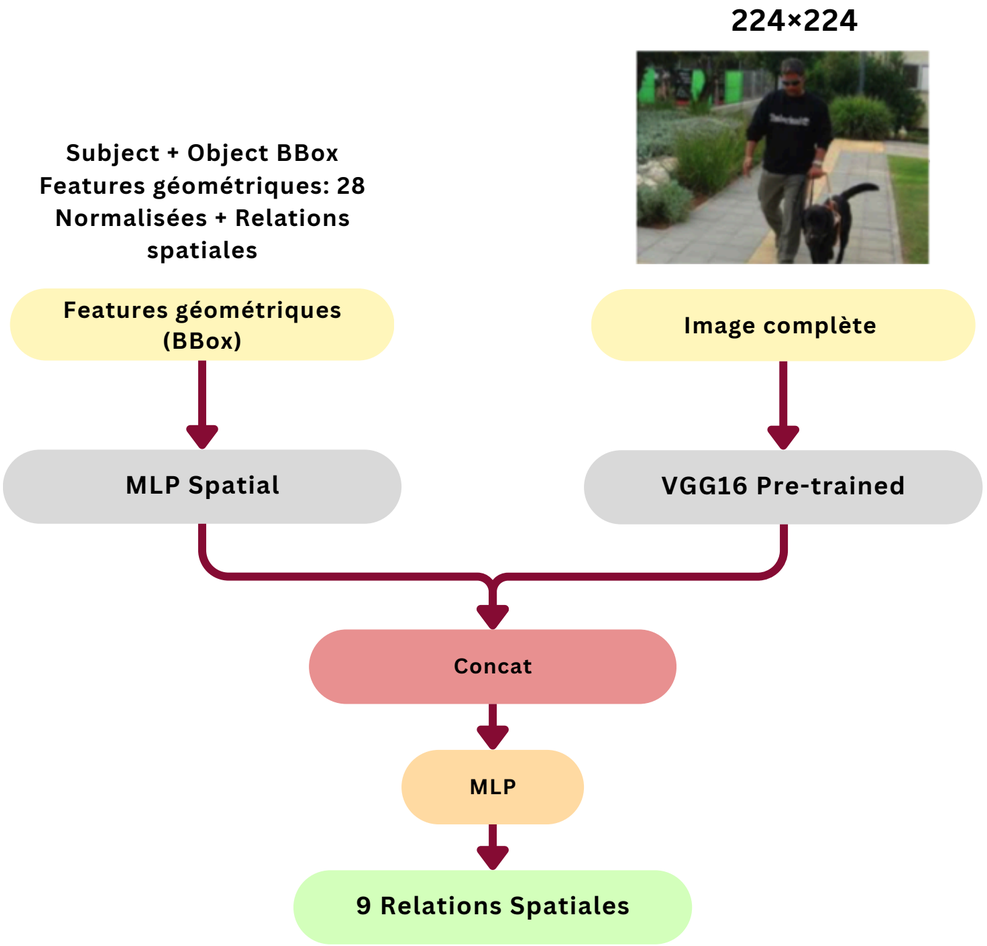
- Entrée 1: Image complète (RGB)")
- Entrée 2: Masque binaire des bounding boxes (Blanc=objets, Noir=fond)
- Fusion: Concaténation des features VGG (4096 + 4096 = 8192)
- MLP 8192->512>256->9



# MÉTHODE 5: Architecture Duale Image + Features géométriques

8

- Approche 2 Modalités:
  - Image complète → VGG16 pré-entraîné (4096 features)
  - Features géométriques → MLP spatial (28 → 256 → 512)
- Fusion ҭ Classification:
  - Concaténation : 4096 + 512 = 4608 features
  - MLP : 4096 → 512 → 256→ 9 relations

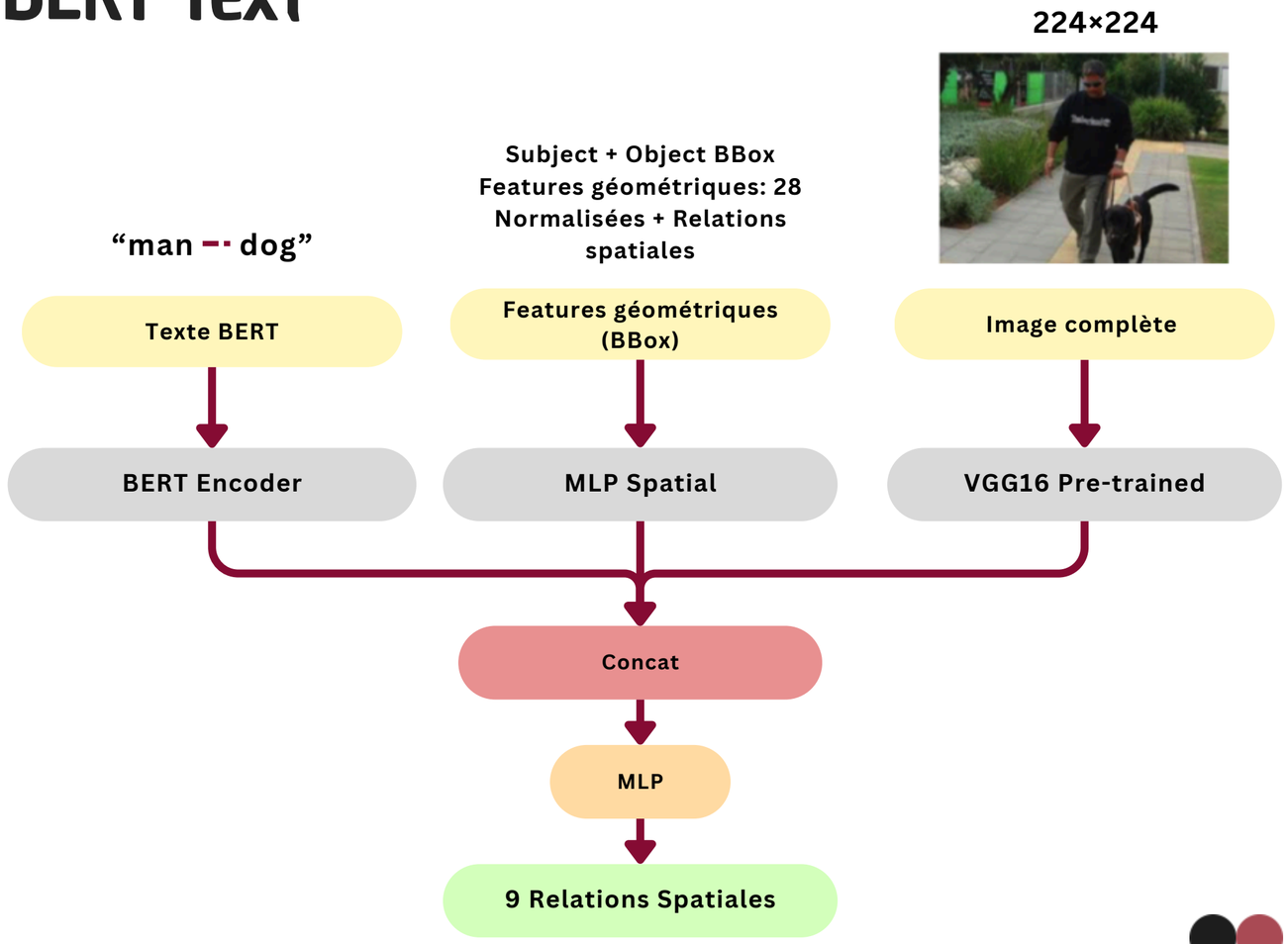


# MÉTHODE 6: Architecture Multimodal
# Image + Features géométriques + BERT Text

Université Paris Cité

9

- Approche 3 Modalités:
  - Image complète → VGG16 pré-entraîné (4096 features)
  - Coordonnées BBox → MLP spatial (28 → 256 → 512)
  - Texte "subject object" → BERT encoder (768 )
- Fusion et Classification:
  - Concaténation : 4096 + 512 + 768= 5376features
  - MLP: 5376→ 512→ 256→ 9 relations



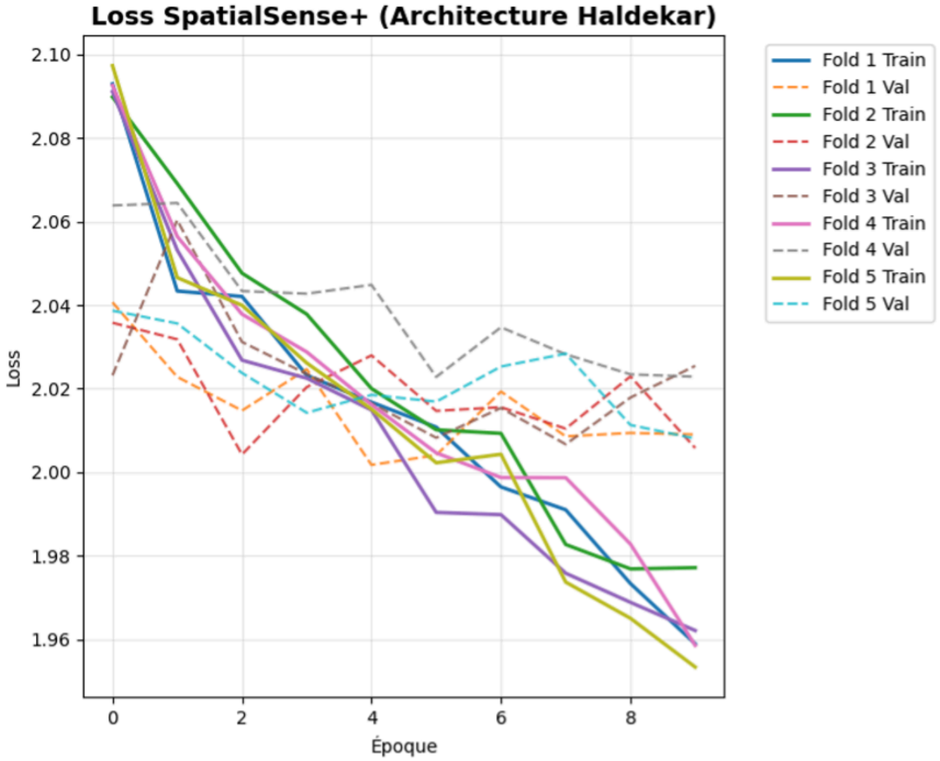
# Expérimentations et Résultats

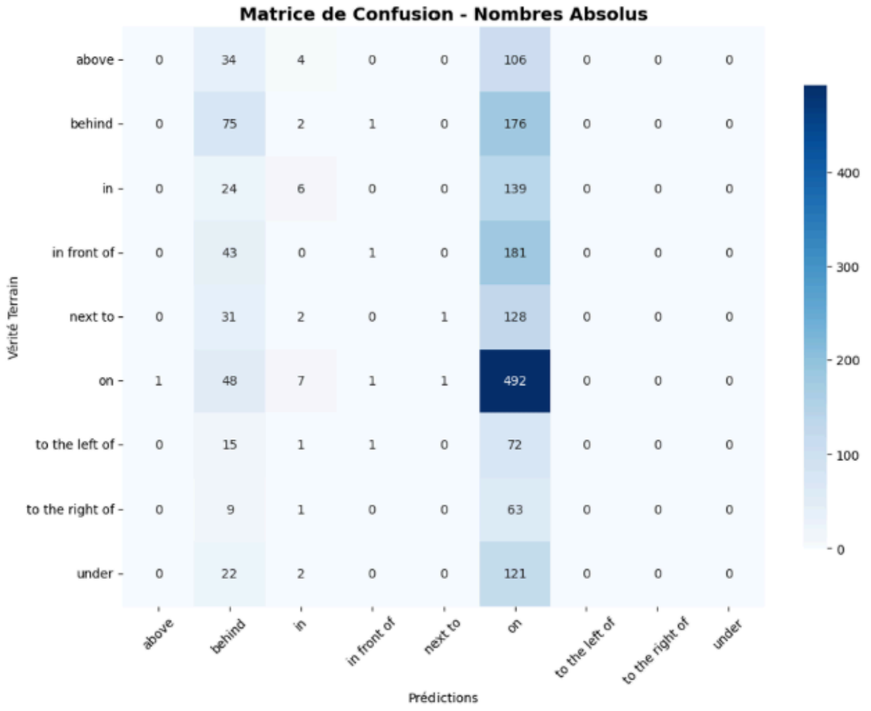
## Méthode 1 Haldekar (VGG16)

Université Paris Cité



**Accuracy du meilleur fold 30.7%, accuracy moyenne de tout les folds 29,8% parametre pour toutesles configs : k_folds=5,early_stopping_patience=5, min_delta=0.001 et batch 8**



# Expérimentations et Résultats

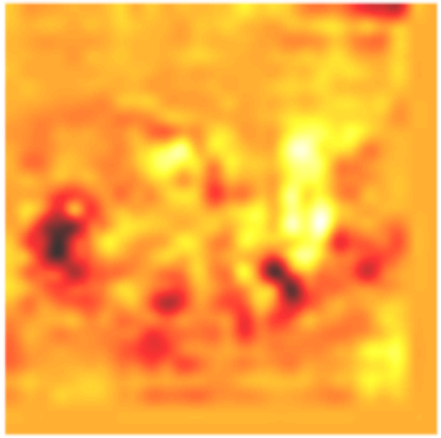
## Méthode 1 Haldekar (VGG16)

Université Paris Cité



boy - boat

Heatmap Haldekar-style



Superposition

SpatialSense+ Sample:

Vérité: next to
Prédiction: on
Original: next to

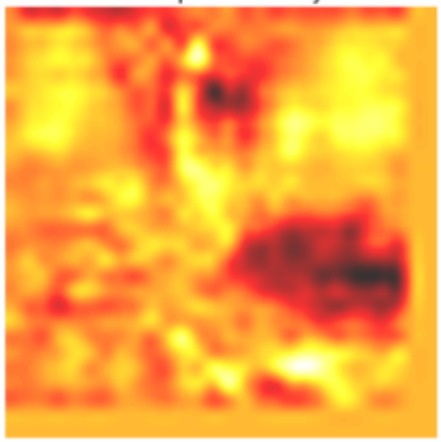
Confiance prédiction: 18.79%
Confiance vérité: 9.28%

Résultat: ❌ INCORRECT

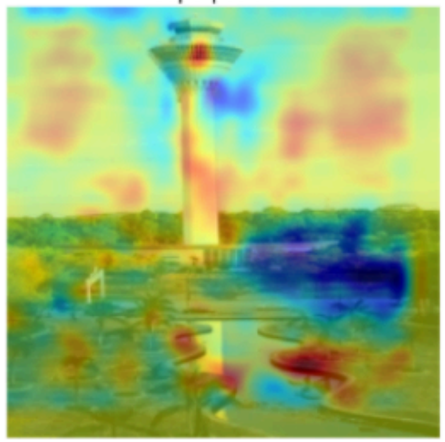
Heatmap (influence):
Max: 0.003
Min: -0.004

tower - water

Heatmap Haldekar-style



Superposition



SpatialSense+ Sample:

Vérité: behind
Prédiction: behind
Original: behind

Confiance prédiction: 18.34%
Confiance vérité: 18.34%

Résultat: ✅ CORRECT

Heatmap (influence):
Max: 0.007
Min: -0.012

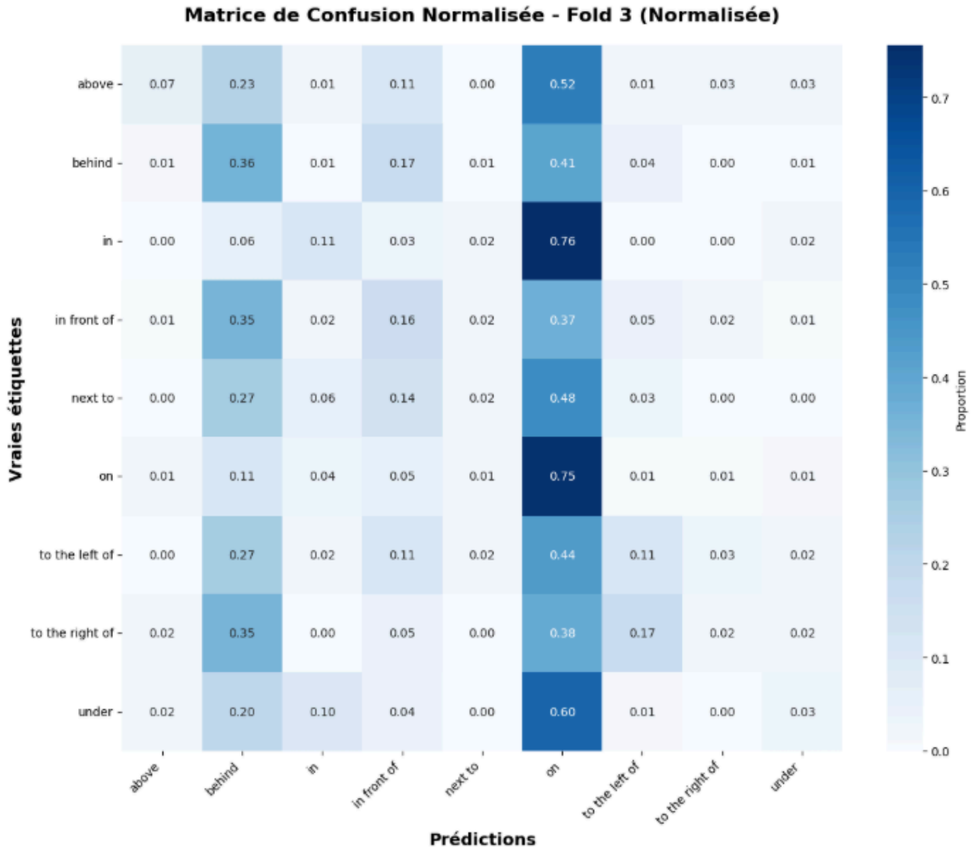
# Expérimentations ET Résultats

## Méthode 4  Image complète + Masking (Noir & Blanc)

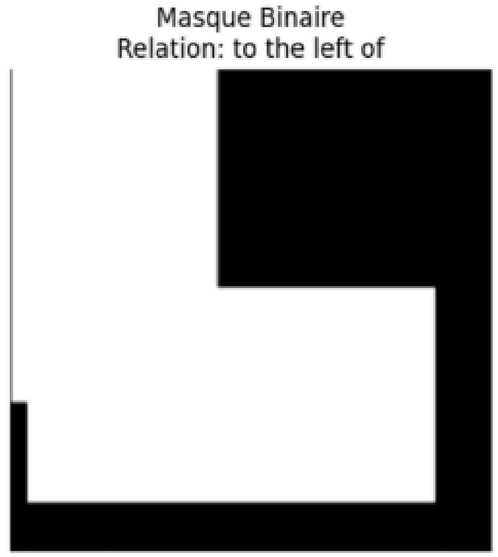
Université Paris Cité



**Accuracy du meilleur fold 33.19%, accuracy moyenne de tout les folds  32.8%**







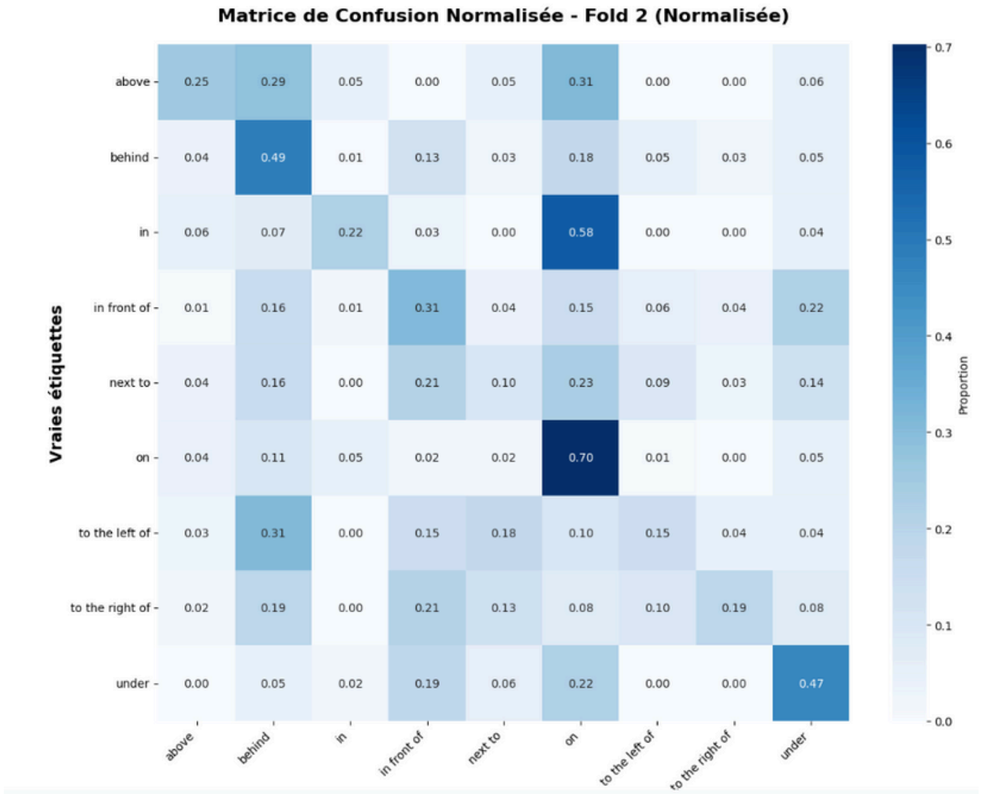
# Expérimentations ET Résultats

## Méthode 5 Image + Features géométriques

Université Paris Cité



**Accuracy du meilleur fold 42.35%, accuracy moyenne de tout les folds 40.33%**




20/06/2025

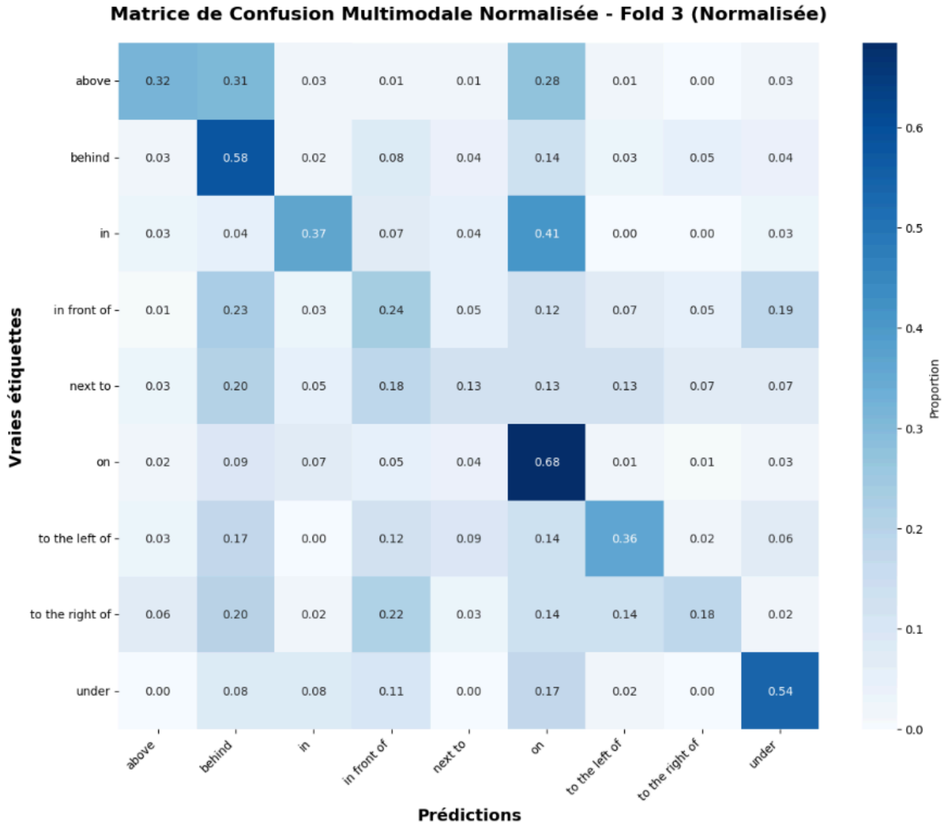
# Expérimentations ET Résultats

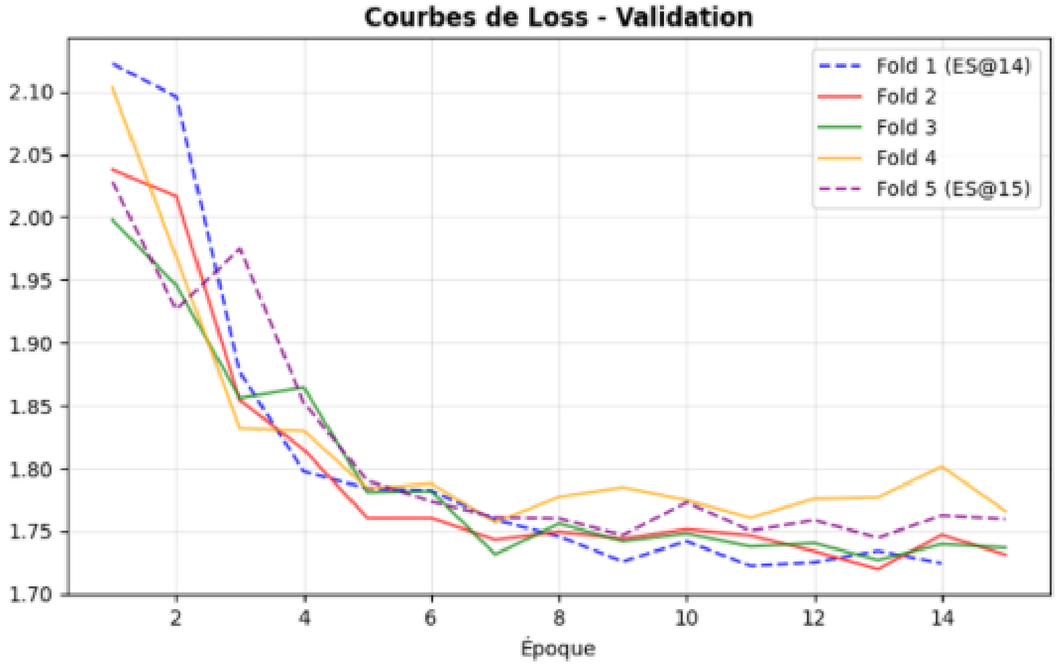
## Méthode 6 Image + Features géométriques + BERT Text

Université Paris Cité



**Accuracy du meilleur fold 45.5%, accuracy moyenne de tout les folds  43.3%**



# Expérimentations ET Résultats

Université Paris Cité

| Méthode | Meilleur Fold | Accuracy Moyenne |
|---|---|---|
| Méthode 1. Haldekar (VGG16) | 30.7% | 29.8% |
| Méthode 2. Vision Transformer | 31.02% | 30.4% |
| Méthode 3. Image complète + Masking | 31.6% | 30.59% |
| Méthode 4. Image complète + Masking(N/B) | 33.19% | 32.8% |
| Méthode 5 Image + Features géométriques | 42.35% | 40.33% |
| Méthode 6. Image + Features géométriques + BERT Text | 45.5% | 43.3% |

## Top 5 erreurs

14

- in → on
- behind → on
- in front of → behind
- on → behind
- above → on

# Conclusion et Perspectives futures

Université Paris Cité




- **Conclusion :**
  - Approche multimodale (Image + geometrie+ BERT) surpasse significativement les méthodes classiques (+13.5%)
  - Features géométriques sont cruciales pour identifier les relations spatiales
  - Confusions persistantes entre relations sémantiquement proches (in/on, behind/on)

- **Perspectives Futures :**
  - Attention mechanisms pour améliorer la fusion multimodale
  - Graph Neural Networks pour modéliser explicitement les relations spatiales