Apprentissage profond et modèles de langage visuel pour l'automatisation des inspections de ponts

Soutenance de stage

Coërchon Colin

ENS Paris-Saclay – ENSIIE
Master MVA
Stage de fin d'étude

31 octobre 2025

Informations clefs

- Tuteur académique : Yann Gousseau
- Maître de stage : Pierre Charbonnier & Philippe Foucher
- Organisme d'accueil : Cerema - Agence de Strasbourg (équipe **ENDSUM**)
- Période de stage : du **10 juin** 2025 au **28 novembre** 2025

Présentation du Cerema

Un établissement public de référence

- Placé sous la tutelle du ministère de la Transition écologique et du ministère de l'Aménagement du territoire.
- Présent sur tout le territoire : **26 implantations**, environ **2 400 agents**.
- Accompagne l'État et les collectivités dans la transition écologique, la mobilité durable et la cohésion des territoires.

L'équipe de recherche ENDSUM

- Présente sur les sites de Rouen et **Strasbourg**.
- Travaille sur la **diagnostic, pronostic et visualisation** de l'état des structures.

Axes de recherche

1. Caractérisation physique des milieux.
2. Développement d'outils de mesure et de traitement de données.
3. **Méthodes de diagnostic et d'intelligence artificielle** appliquées au génie civil.

Le Programme National Ponts (PNP)

Un grand recensement national

- Lancé en **2020** et piloté par le Cerema.
- Objectif : recenser et évaluer les **ouvrages d'art communaux** (ponts et murs).
- Environ **300 000 ouvrages** en France.
  ↪ 63 000 recensés, dont 46 000 ponts et 17 000 murs.
- Outil numérique dédié : **SIRO**.
  ↪ Les « carnets de santé » remplis pour chaque ouvrage.



La grande problématique

Comment exploiter efficacement ces milliers d'images et de rapports pour aider les collectivités à suivre leur patrimoine ?

La grande problématique

Comment exploiter efficacement ces milliers d'images et de rapports pour aider les collectivités à suivre leur patrimoine ?

$\Longrightarrow$ Le projet ApPLAuDI !

Le projet ApPLAuDI

- Porté par le Cerema et récemment financé par la fondation FEREC.
- Objectif : développer un assistant d'inspection intelligent pour favoriser l'automatisation de l'inspection des ouvrages d'art.
  ↪ Et ceci pour toute la chaîne de traitement !
- Approche exploratoire fondée sur les **modèles Vision–Langage (VLM)** récents.



Objectifs du stage

Une première brique du projet ApPLAuDI

- Étude bibliographique sur les modèles **VLM** :
  - Exploration et compréhension de ces modèles complexes.
  - Identification des architectures les plus pertinentes.
- **Exploration d'un cas d'usage** : obtenir la description textuelle d'un ouvrage.

Objectifs du stage

Une première brique du projet ApPLAuDI

- Étude bibliographique sur les modèles **VLM** :
  - Exploration et compréhension de ces modèles complexes.
  - Identification des architectures les plus pertinentes.
- **Exploration d'un cas d'usage** : obtenir la description textuelle d'un ouvrage.

Exemple : « Pont à tablier, porté par des culées maçonnées, franchissant une rivière, et portant une piste cyclable. »

Ma démarche

1. Comprendre les modèles VLM.
2. Explorer la base du PNP.
3. Adopter une stratégie **logique** et **progressive**.

Ma démarche

Les sous-objectifs créés

- Classification sur les images de la base : détecter la typologie de l'ouvrage.
  $\hookrightarrow$ **5 classes** : buse / cadre et portique / mur / pont à tablier / pont voûte.
- Modèle connu pour commencer : `ResNet`.
  $\hookrightarrow$ Se créer une *baseline*.
- Mono-vue puis multi-vues.
- Puis les modèles VLM...

Sommaire

1. Contexte et objectifs du stage
2. Présentation des données
3. État de l'art
4. Chaîne de traitement pour l'apprentissage
5. Résultats et analyses
6. Conclusion et perspectives

1 Contexte et objectifs du stage

2 Présentation des données
- Données textuelles
- Données images

3 État de l'art

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses

6 Conclusion et perspectives

Des exemples de variables

- Des données de localisation : `Nom de la commune`, `Département`, `Code INSEE`, etc.
- Des métadonnées.
- Des données descriptives : `PTAC limite`, `Type de l'obstacle franchi`, etc.
- Des données sur son état : `Niveau de défaut général`, `Problème de sécurité`, etc.
- Beaucoup de données selon **la nature de l'ouvrage**.
  $\hookrightarrow$ **Plus de la moitié** en réalité.

En quelques chiffres

63 457 ouvrages recensés, avec **148** variables chacun :

Catégorie	**Nombre de variables**
Variables 100% manquantes	24
Variables binaires strictes	6
Variables quasi binaires	14
Variables enums (3–15 modalités)	47
Variables jouant un rôle de clef	2
Variables restantes	55
Variables 100% présentes	41
Variables avec $<1\%$ de NaN	1
Total de variables analysées	**148**

Note : « quasi binaire » $\Longrightarrow$ Oui / Non / [autre chose]

Pour notre étude

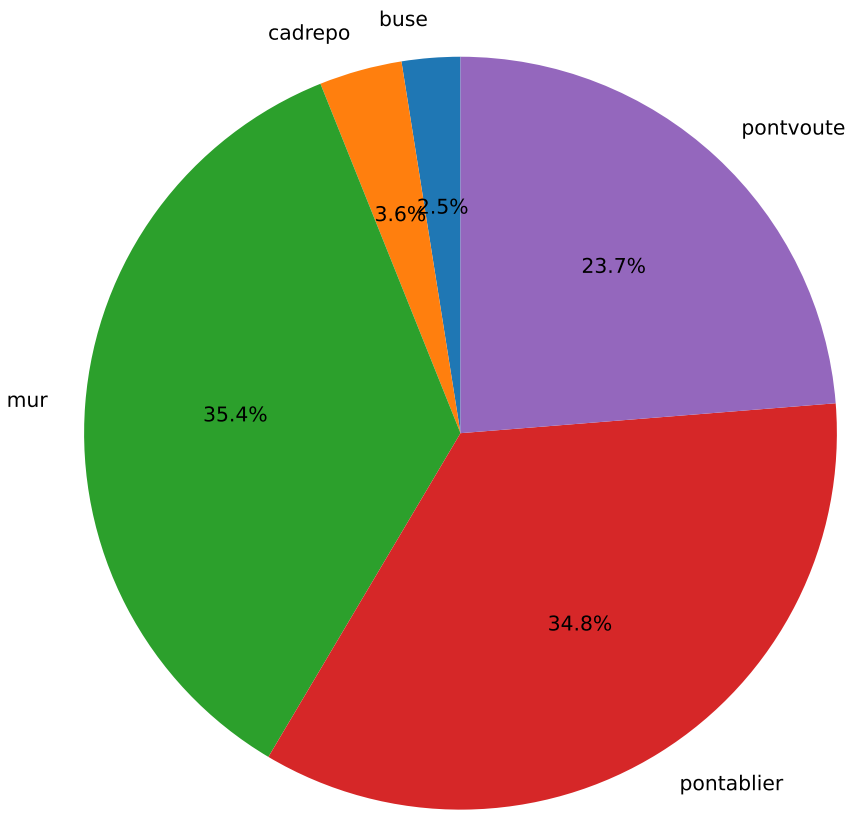


Figure – Répartition des données selon la nature de l'ouvrage

1 Contexte et objectifs du stage

2 Présentation des données
- Données textuelles
- Données images

3 État de l'art

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses

6 Conclusion et perspectives

Mise en contexte

Les données photographiques

- **350 549** images associées aux 63 457 ouvrages recensés.
- 148 variables $\rightsquigarrow$ **21 variables « photos »**
- 5,3 images en moyenne par ouvrage.
- Stockées par commune. On retrouve :
  - L'identifiant de l'ouvrage : `globalid`.
  - Les **codes** (`X1`, `X2`, etc.) présents dans les noms des fichiers.
  - Les **dossiers** dans lesquels les images sont stockées.



Élévation amont



Élévation aval



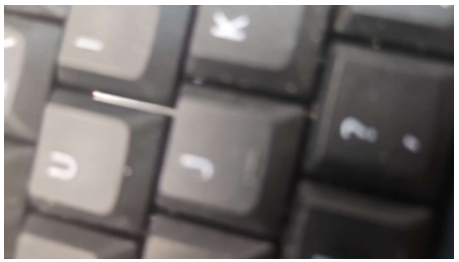
Vue générale



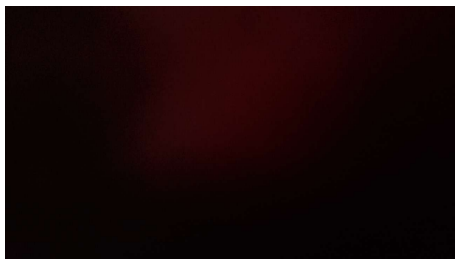
Défaut de structure

Figure – Exemple de pont voûte — Cercié (69)

Certaines images difficilement exploitables





Quelques bizarreries





Ceci est en réalité un mur.





Ceci est un pont à tablier, mais aussi un cadre et portique.

1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art
- Classification *single-view*
- Classification *multi-view*
- Application des VLM

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses

6 Conclusion et perspectives

Les articles importants

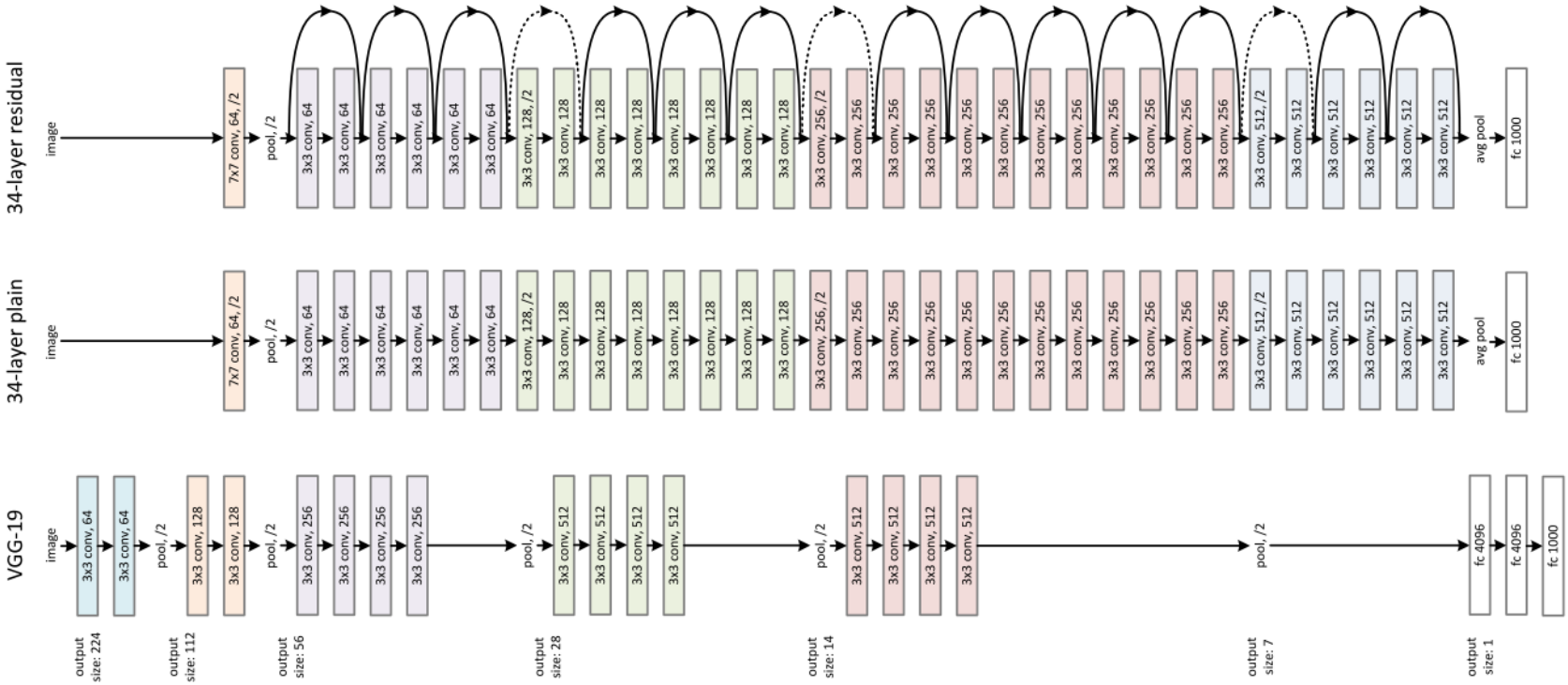
Articles pour l'apprentissage en *Single-view*

- Présentation générale du modèle ResNet - *Deep Residual Learning for Image Recognition* [1]
- Transfert des modèles préentraînés sur ImageNet - *Do Better ImageNet Models Transfer Better ?* [2]
- *Advanced Deep Learning* de Guillaume Charpiat au MVA - *Advanced Deep Learning* (cours MVA) [3]

Pourquoi commencer avec ResNet ?

- **Architecture de référence** en vision par ordinateur depuis **2016**.
- **Préentraîné sur ImageNet** $\rightsquigarrow$ transfert efficace vers d'autres tâches.
- Structure modulaire : adaptée pour l'expérimentation *single-view* puis *multi-view*.

Architecture ResNet



Source : figure de [1]

1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art
- Classification *single-view*
- Classification *multi-view*
- Application des VLM

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses

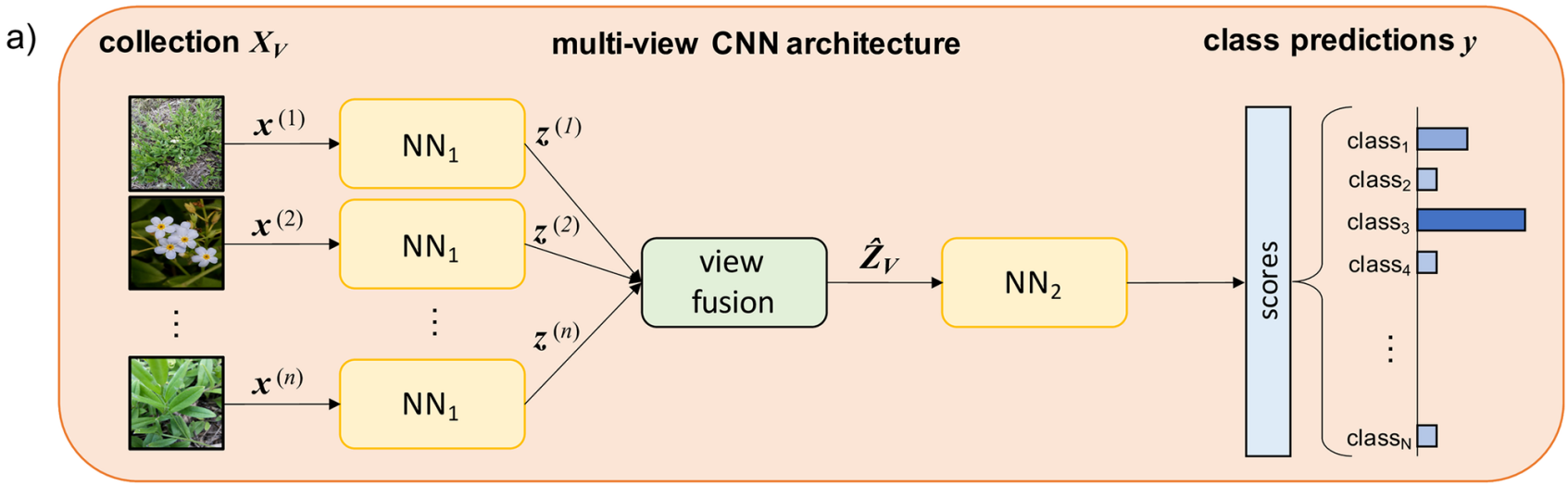
6 Conclusion et perspectives

Les articles importants

Articles pour l'apprentissage en *Multi-view*

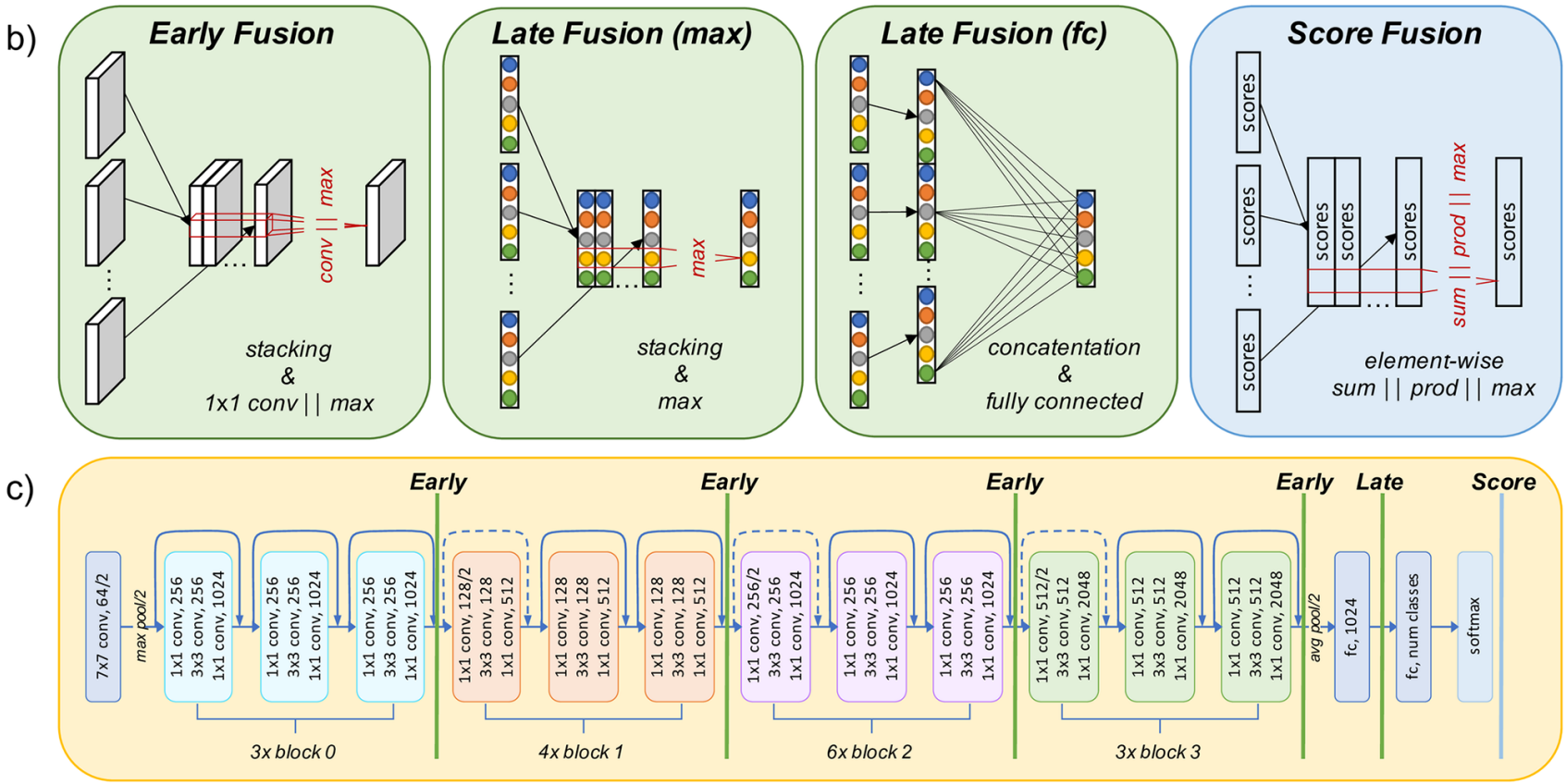
- Principes et méthodes de la classification multi-vues - *Multi-view Classification with Convolutional Neural Networks* [4]
  ↪ Ma référence principale.
- Analyse mathématique et revue approfondie des approches multi-vues - *Deep Multi-View Learning Methods : A Review* [5]
- Application concrète à la reconnaissance d'ouvrages à partir d'images de rue - *Two-Step Deep Learning Approach for Classifying Bridges using Street Imagery* [6]
  ↪ Méthodologie très similaire à notre étude.

Fonctionnement général



Source : figure de [4]

Les 3 grandes stratégies



Source : figure de [4]

Stratégies choisies

Late Fusion

Concaténation des différents *embeddings* suivie d'une couche entièrement connectée :

$$\hat{z}_V = \phi(W[z^{(1)}, \ldots, z^{(n_V)}] + b)$$

où ϕ est une activation non linéaire (ReLU), $V = 5$ images par ouvrage.

Stratégies choisies

Score Fusion

Chaque vue est classée indépendamment et les scores de sortie $y^{(v)}$ sont ensuite agrégés selon différentes règles :

$$\text{(a) somme :} \quad \hat{y}_{V,\oplus} = \sum_v y^{(v)}$$

$$\text{(b) produit :} \quad \hat{y}_{V,\otimes} = \prod_v y^{(v)}$$

$$\text{(c) maximum :} \quad \hat{y}_{V,\max} = \max_v y^{(v)}$$

1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art
- Classification *single-view*
- Classification *multi-view*
- Application des VLM

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses

6 Conclusion et perspectives

Les articles importants

Articles pour les modèles *Vision–Langage (VLM)*

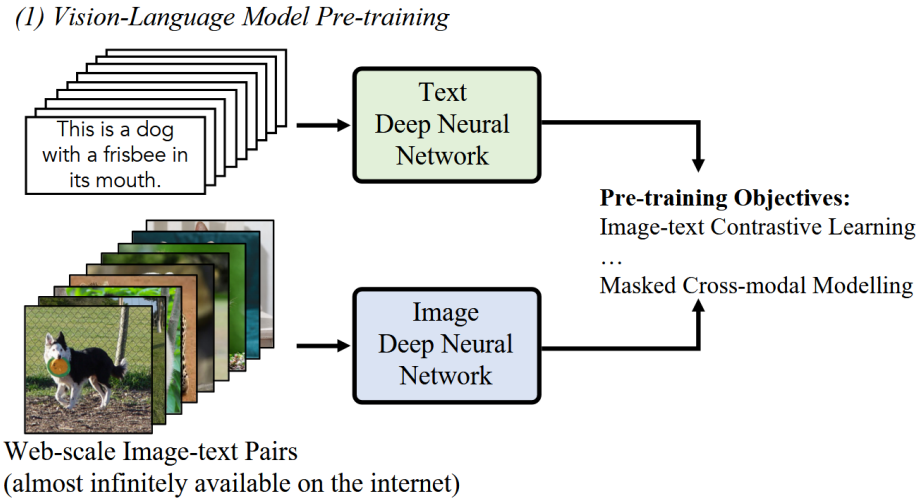
- **Théorie générale et panorama des VLM** - *Vision–Language Models for Vision Tasks : A Survey* [7] et *An Introduction to Vision–Language Modeling* [8].
- **Méthodes d'adaptation et de transfert** - *LoRA : Low-Rank Adaptation of Large Language Models* [9] et *Learning to Prompt for Vision–Language Models* [10].
- **Applications à l'inspection d'ouvrages d'art** - *Bridge Inspection Using a Multi-Modal Vision Language Model* [11] et *Automatic Bridge Inspection Database Construction Through Hybrid Information Extraction and Large Language Models* [12].

Principe général des VLM

- On dispose de paires **image + texte** $\left(x^I, x^T\right)$
- Chaque élément est encodé séparément dans un espace latent pour former des **embeddings** :

$$x^I \longrightarrow z^I \quad \text{et} \quad x^T \longrightarrow z^T$$

- L'objectif du pré-entraînement $\rightsquigarrow$ faire en sorte que : $\langle z^I, z^T \rangle$ soit maximal si l'image et le texte vont ensemble.



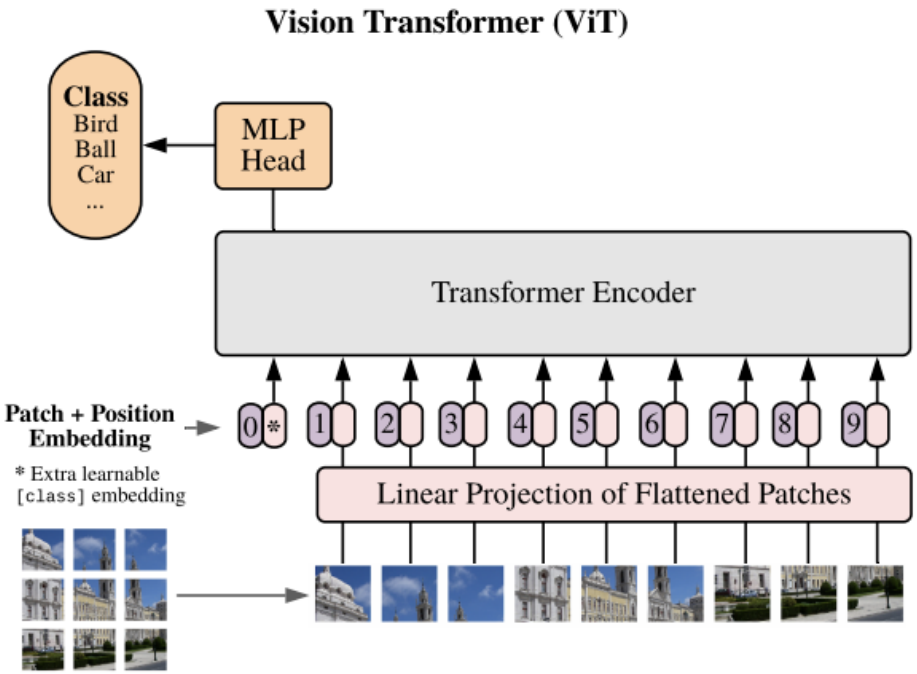
Méthode générale

Images

- **CNN** : `ResNet`, `ConvNeXt`
- **ViT** : Vision Transformer (patchs + self-attention)

Texte

- **Transformers** : architecture dominante
- Exemples : `BERT`, `RoBERTa`, etc...



Exemple d'encodeur image : `ViT`

Source : extrait adapté de [7]

Objectifs de pré-entraînement

Plusieurs objectifs existent :

- Objectifs contrastifs.
- Objectifs génératifs.
- Objectifs d'alignement.

↪ Ils sont décrits plus en détail dans une autre présentation, et dans le rapport de stage.

Objectifs contrastifs : ITC

Image-Text Contrastive Learning (ITC)

- On apprend à rapprocher les paires (x^I, x^T) positives, et à éloigner les autres
- Structure **symétrique** : encodeur image + encodeur texte

$$\mathcal{L}_{I\to T} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\left\langle z_i^I, z_i^T \right\rangle / \tau\right)}{\sum_{j=1}^{B} \exp\left(\left\langle z_i^I, z_j^T \right\rangle / \tau\right)}$$

$$\mathcal{L}_{T\to I} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\left\langle z_i^T, z_i^I \right\rangle / \tau\right)}{\sum_{j=1}^{B} \exp\left(\left\langle z_i^T, z_j^I \right\rangle / \tau\right)}$$

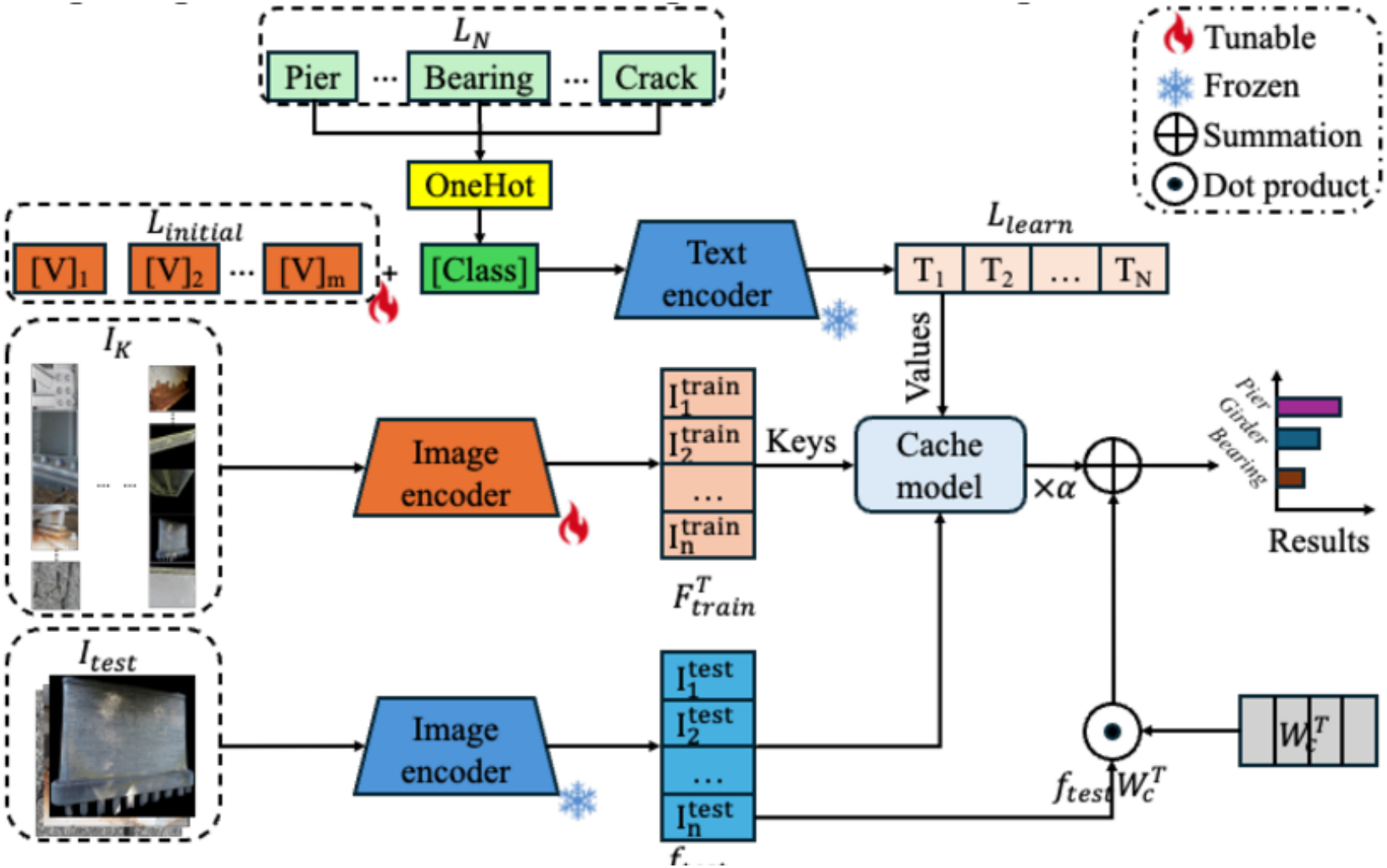
$$\mathcal{L}_{\text{ITC}} = \mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}$$

↪ CLIP, ALIGN, FILIP, etc...

Utilisation dans notre étude

Plusieurs méthodes d'ajustement d'un modèle pré-entraîné

- *Fine-tuning* : complet, partiel ou « efficace en paramètre » (PEFT) comme LoRA [9].
- *Transfer learning* :
  - ***Prompt tuning*** : Ajustement sur les données textuelles.
    ↪ La méthode la plus populaire.
  - ***Feature adaptation*** : Ajustement sur *features* des données images.
- Distillation de connaissances : souvent utilisé pour la **détection d'objets** ou la **segmentation sémantique**.



Source : figure de [11]

1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage
- Plan d'expérience
- Mise en qualité
- Les datasets
- Configuration et mesure de performance

5 Résultats et analyses

6 Conclusion et perspectives

Plan d'expérience

1. **Mise en qualité de la base de données**
   ↪ Fusion, typage, et nettoyage.
2. **Classification *single-view***
   ↪ `ResNet-18` et `ResNet-50`.
3. **Classification *multi-vues***
   ↪ Deux méthodes de fusion : *Score Fusion* et *Late Fusion*.
4. **Modèles vision–langage (VLM)**
   - **CLIP** en *zero-shot learning*.
   - **Fine-tuning** et **transfer learning** sur des modèles préentraînés. Méthodologie par [11].

1 Contexte et objectifs du stage

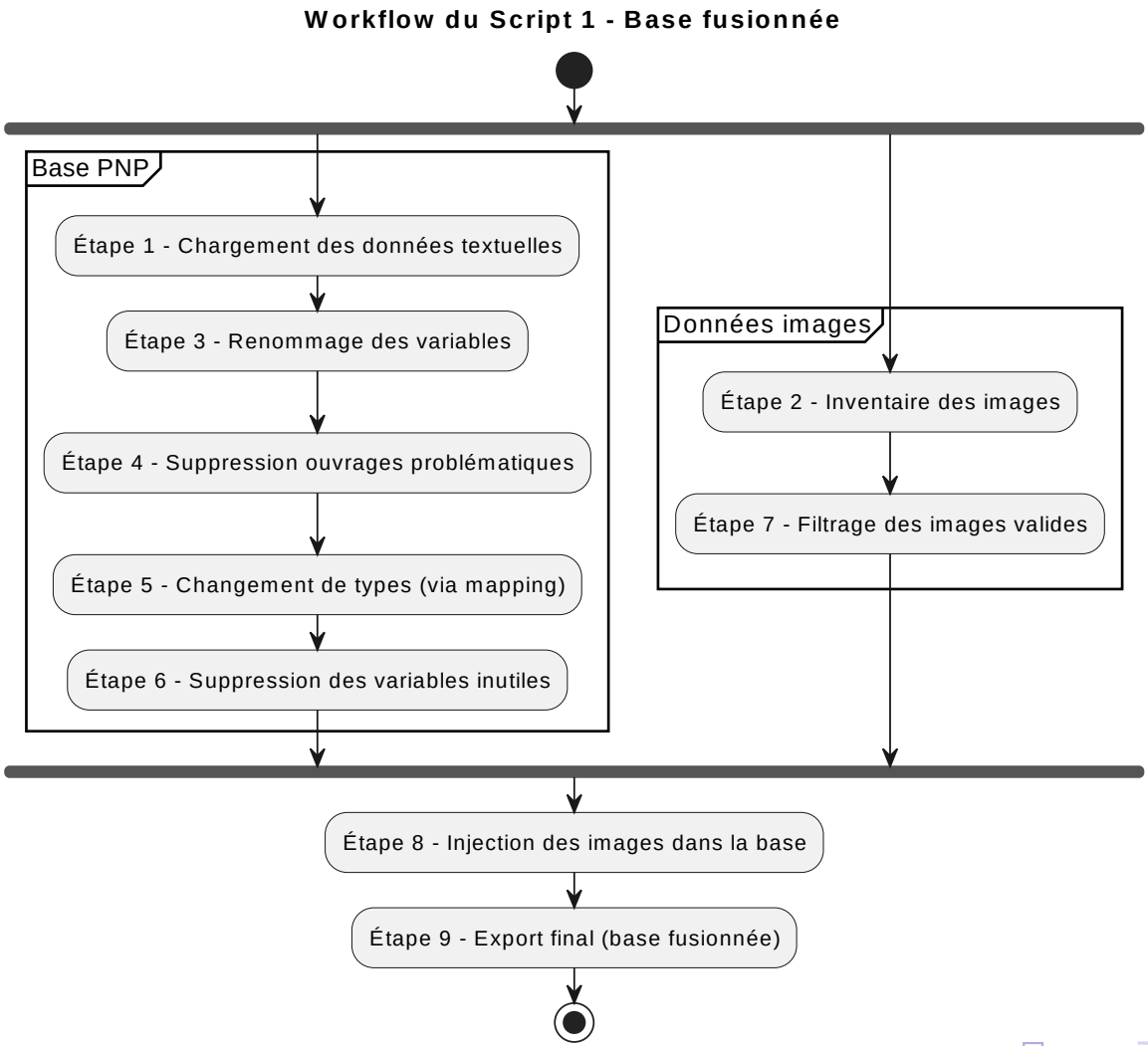
2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage
- Plan d'expérience
- Mise en qualité
  - Création d'une base exploitable
  - Nettoyage des photos
- Les datasets
- Configuration et mesure de performance

5 Résultats et analyses

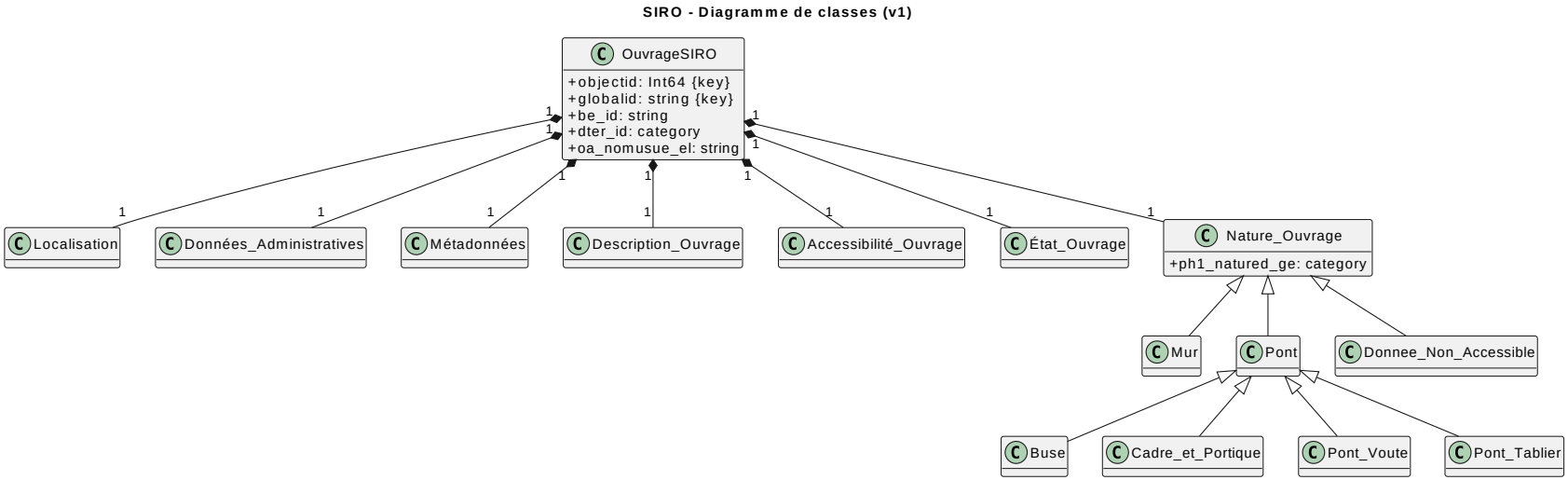
Script de la mise en qualité



Résultat

On obtient une **base exploitable** pour du *machine learning* et *deep learning* :

- Variables filtrées, renommées et typées.
- Base fusionnée : 21 champs de photos correctement instanciés.
- Suppression de quelques ouvrages problématiques.



1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage
- Plan d'expérience
- Mise en qualité
  - Création d'une base exploitable
  - Nettoyage des photos
- Les datasets
- Configuration et mesure de performance

5 Résultats et analyses

Détection des images vertes

Méthodologie

1. Conversion des images en format HSV.
2. Un pixel est considéré comme « vert » si :

$$70^\circ \leq H \leq 170^\circ, \quad S \geq 0.3, \quad V \geq 0.5$$

3. Calcul du taux de pixels **verts dans la zone centrale** (zone fovéale).
4. Suppression des images contenant plus de **50 % de pixels verts**.

Zone fovéale

Image originale (disque central)



Pixels verts détectés (dans le disque)

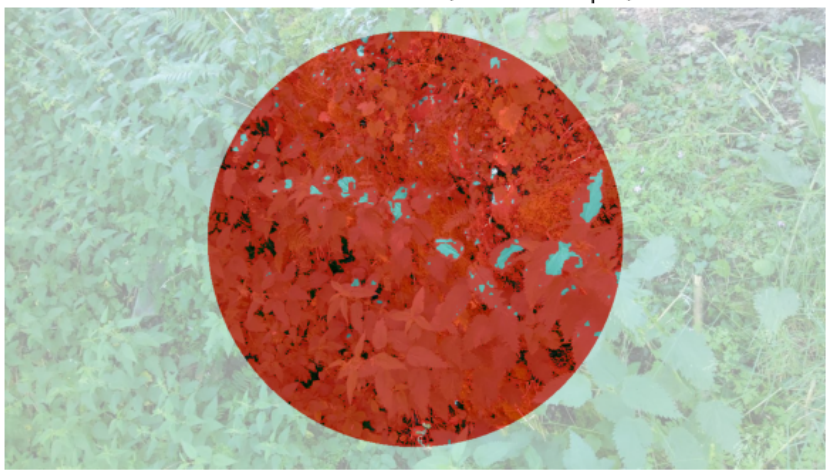


Image originale (disque central)



Pixels verts détectés (dans le disque)



Résultats

5 597 images dominées par la végétation supprimées.



Exemples de clichés d'élévation supprimés car largement obstrués par la végétation.

1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage
- Plan d'expérience
- Mise en qualité
- Les datasets
- Configuration et mesure de performance

5 Résultats et analyses

6 Conclusion et perspectives

Nos 3 jeux de données d'images

Initialement, dans la base SIRO :

- **63 457** ouvrages.
- **350 549** images dont 103 386 d'élévation.

Nom du dataset	Description principale	Nombre d'images
Dataset 1	Vues d'élévation uniques par ouvrage	92 859
Dataset 2	Sous-échantillonnage du Dataset 1	25 004
Dataset 3	Multi-vues (toutes les images)	306 038

Nos 3 jeux de données d'images

Initialement, dans la base SIRO :

- **63 457** ouvrages.
- **350 549** images dont 103 386 d'élévation.

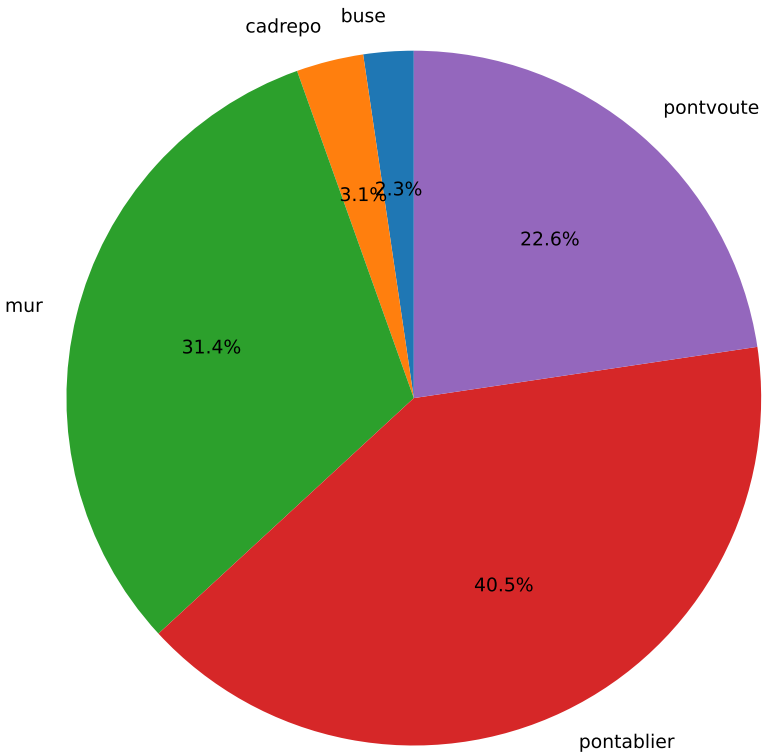
Nom du dataset	**Description principale**	**Nombre d'images**
Dataset 1	Vues d'élévation uniques par ouvrage	92 859
Dataset 2	Sous-échantillonnage du Dataset 1	25 004
Dataset 3	Multi-vues (toutes les images)	306 038

Séparation données `train`, `validation` et `test` : 70% / 15% / 15%.
↪ Selon **le nombre d'ouvrages**, et non le nombre d'images !

Répartition des types d'ouvrage - Dataset 1

Nombre d'images par nature

- **Ponts** : 2 vues d'élévation.
- **Murs** : 1 vue d'élévation.



1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage
- Plan d'expérience
- Mise en qualité
- Les datasets
- Configuration et mesure de performance

5 Résultats et analyses

6 Conclusion et perspectives

Fonction de perte finale

Cross Entropy pondérée avec *label smoothing*

$$\mathcal{L}_{\text{finale}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} \widetilde{w}_c \, y_{i,c}^{\delta} \log(p_{i,c})$$

- $p_{i,c}$: probabilité prédite par le modèle pour la classe c (via *softmax*).
- $\widetilde{w}_c$: pondération de la classe selon sa fréquence (corrige le déséquilibre).
- $y_{i,c}^{\delta} = (1-\delta)y_{i,c} + \frac{\delta}{K}$: cibles lissées (*label smoothing*).
- $\delta \in \{0.05, 0.1\}$: niveau de lissage appliqué aux labels.

Mesures de performances

Métriques utilisées

- **Accuracy** : proportion d'échantillons correctement classés.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(\hat{y}_i = y_i)}$$

- **F1-score macro** : moyenne des F1-scores calculés pour chaque classe.

$$\text{F1}_{\text{macro}} = \frac{1}{K} \sum_{c=1}^{K} 2 \times \frac{\text{Précision}_c \times \text{Rappel}_c}{\text{Précision}_c + \text{Rappel}_c}$$

1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses
- Les modèles single-view
- Les modèles multi-view
- Les modèles VLM
- Synthèse des résultats

6 Conclusion et perspectives

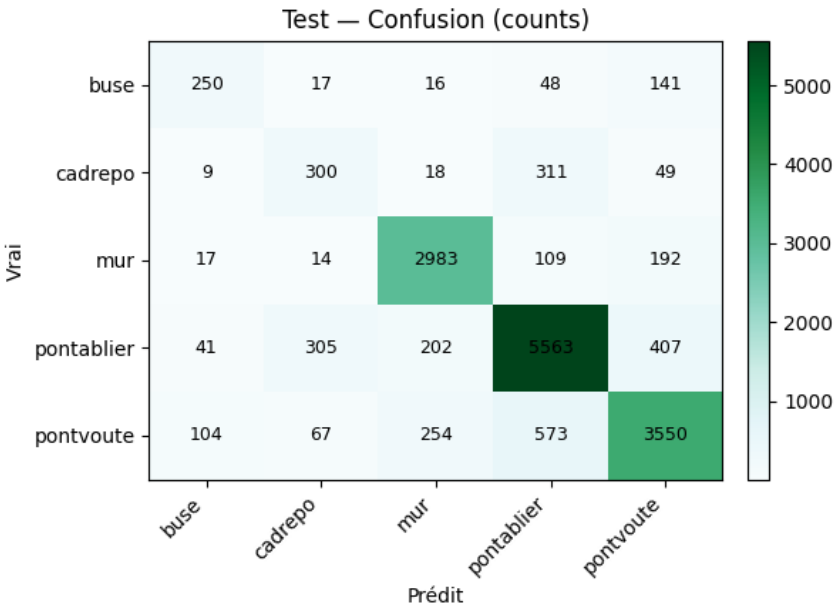
La *baseline* - ResNet-18 - 30 *epoch*

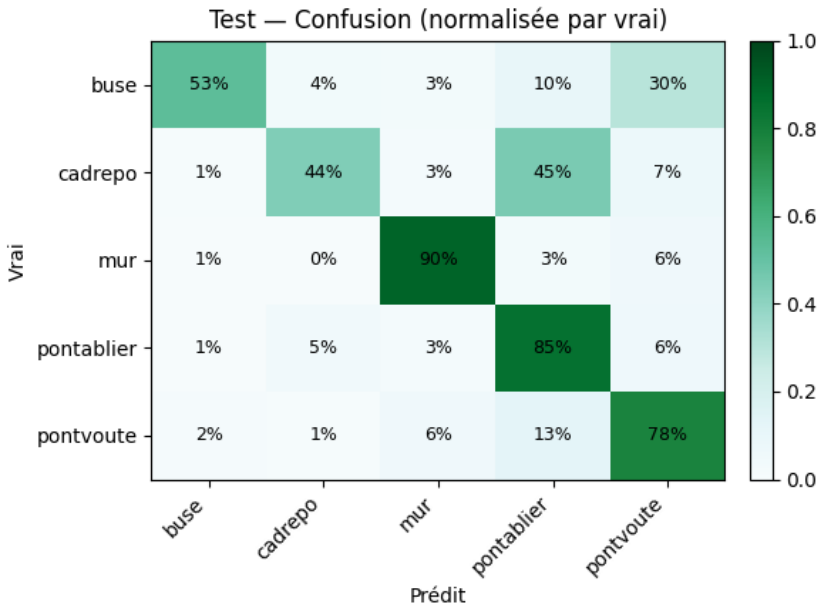
Classe	Précision	Rappel	F1-score	Support
buse	0.594	0.530	0.560	472
cadrepo	0.427	0.437	0.432	687
mur	0.859	0.900	0.879	3315
pontablier	0.842	0.853	0.848	6518
pontvoute	0.818	0.781	0.799	4548
Moyenne (macro)	0.708	0.700	0.703	15540
Moyenne pondérée	0.813	0.814	0.813	15540

	Accuracy	**F1-macro**
Top-1	0.8138	0.7035

Table – Résultats globaux de l'expérience *Baseline* (ResNet-18).

Matrices de confusion





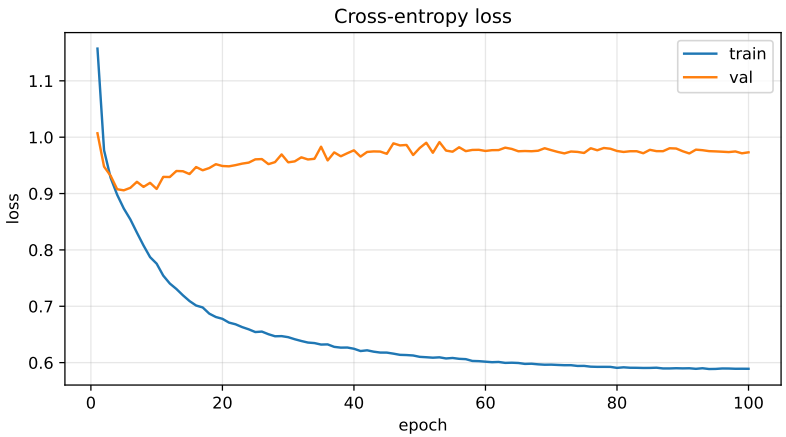
Expérience finale - ResNet-50 - 100 *epochs*

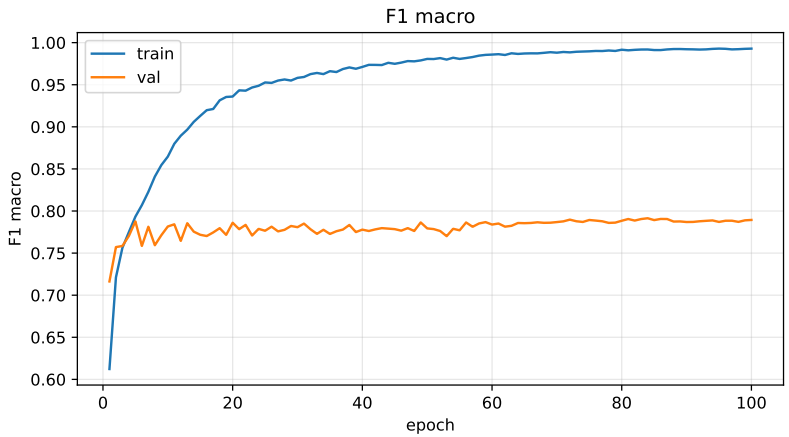
Classe	Précision	Rappel	F1-score	Support
buse	0.714	0.630	0.669	408
cadrepo	0.606	0.441	0.510	610
mur	0.920	0.927	0.924	2999
pontablier	0.872	0.912	0.891	5891
pontvoute	0.871	0.854	0.862	4001
Moyenne (macro)	0.797	0.753	0.771	13909
Moyenne pondérée	0.866	0.870	0.867	13909

	Accuracy	**F1-macro**
Top-1	0.8695	0.7714
Top-2	0.9420	0.8533

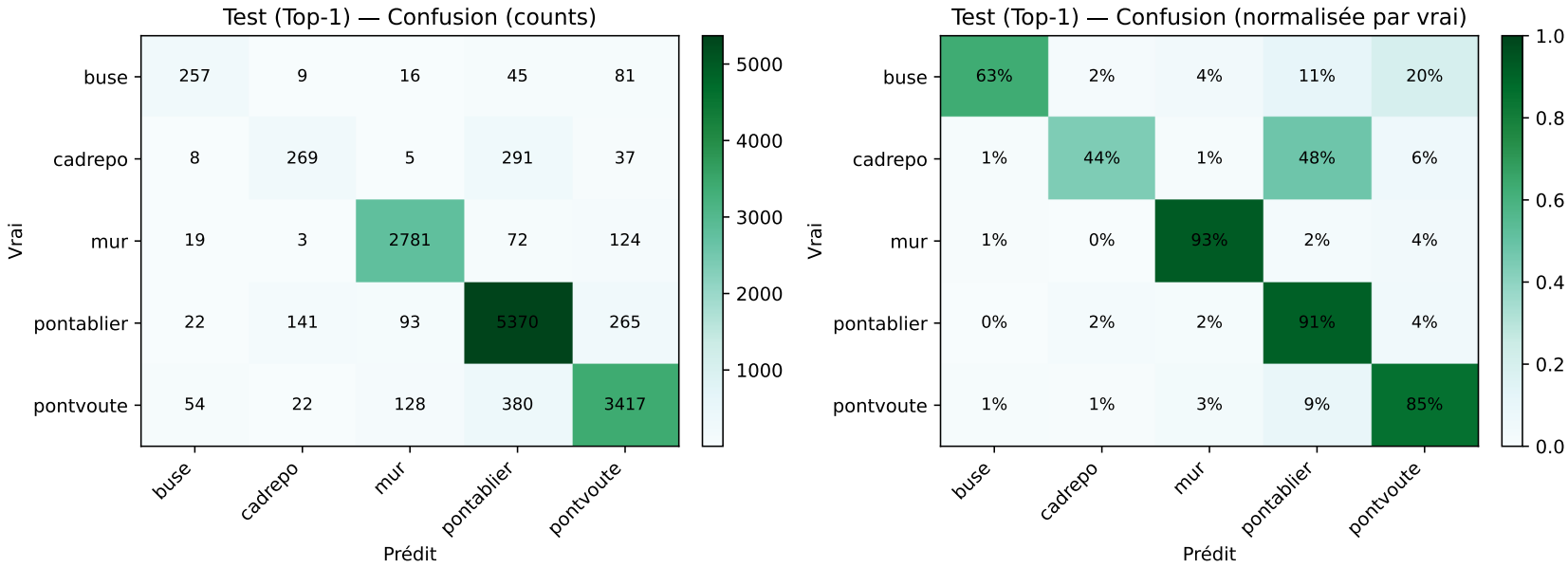
Table – Résultats globaux de l'expérience finale (ResNet-50, scheduler et élagage complet).

Les courbes d'apprentissage

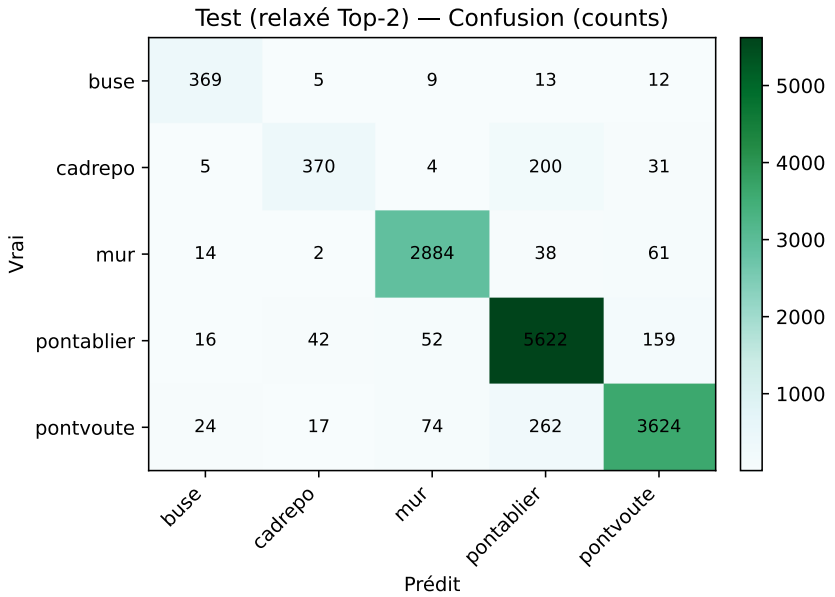


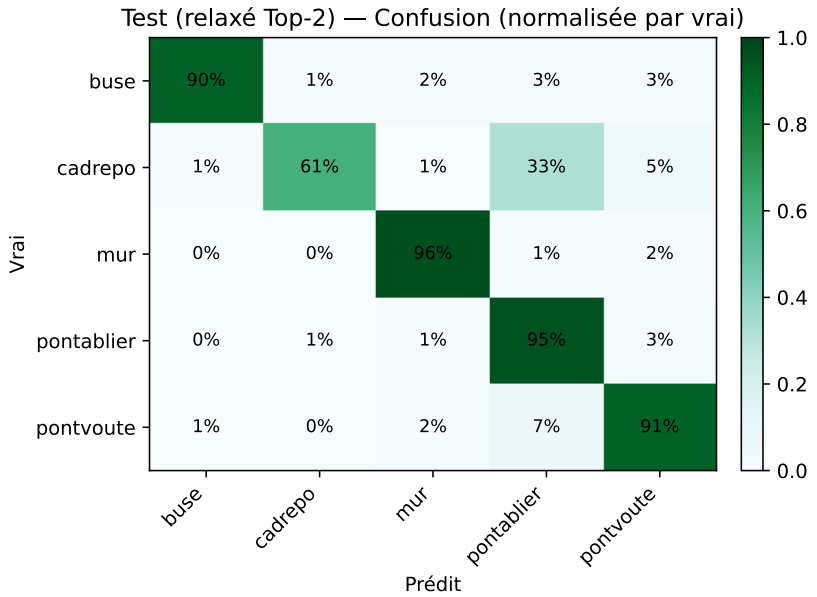


Matrices de confusion - Top-1

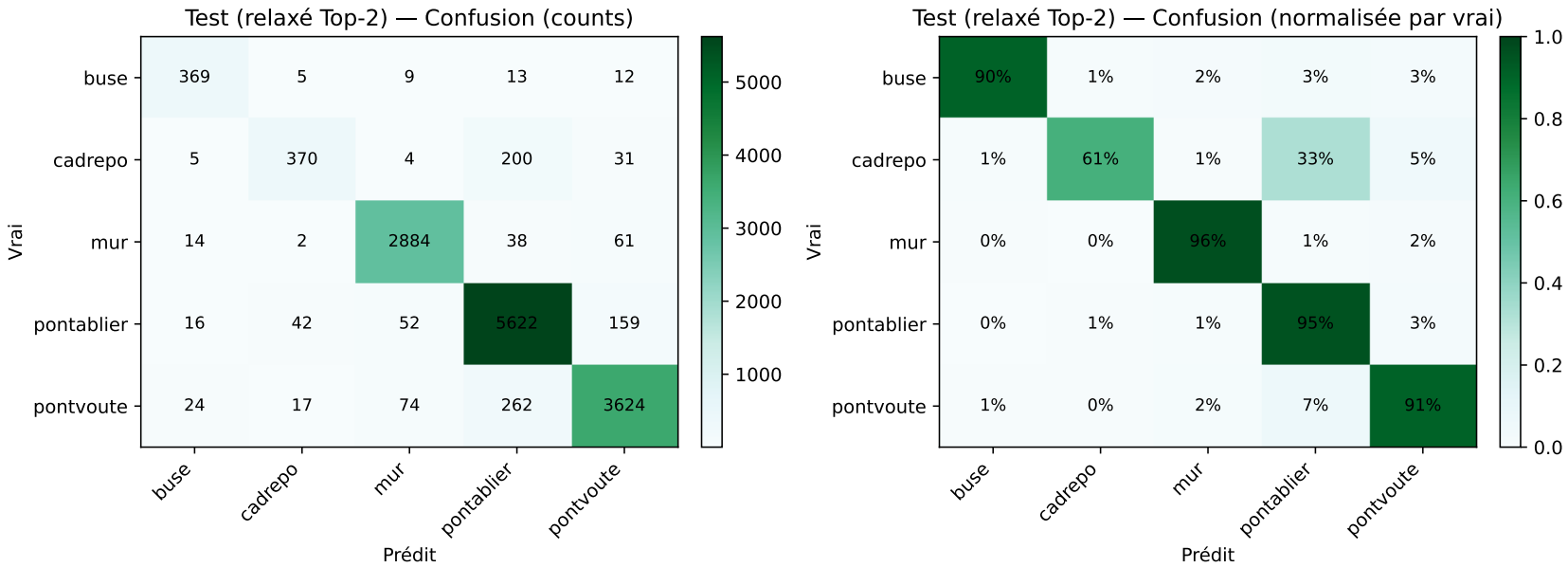


Matrices de confusion - Top-2





Matrices de confusion - Top-2



Conclusion

Beaucoup de difficultés à différencier les **pont tablier** des **cadres et portiques**.

↪ Si hésitation : préférence pour les ponts tabliers.

Des exemples



1 Contexte et objectifs du stage

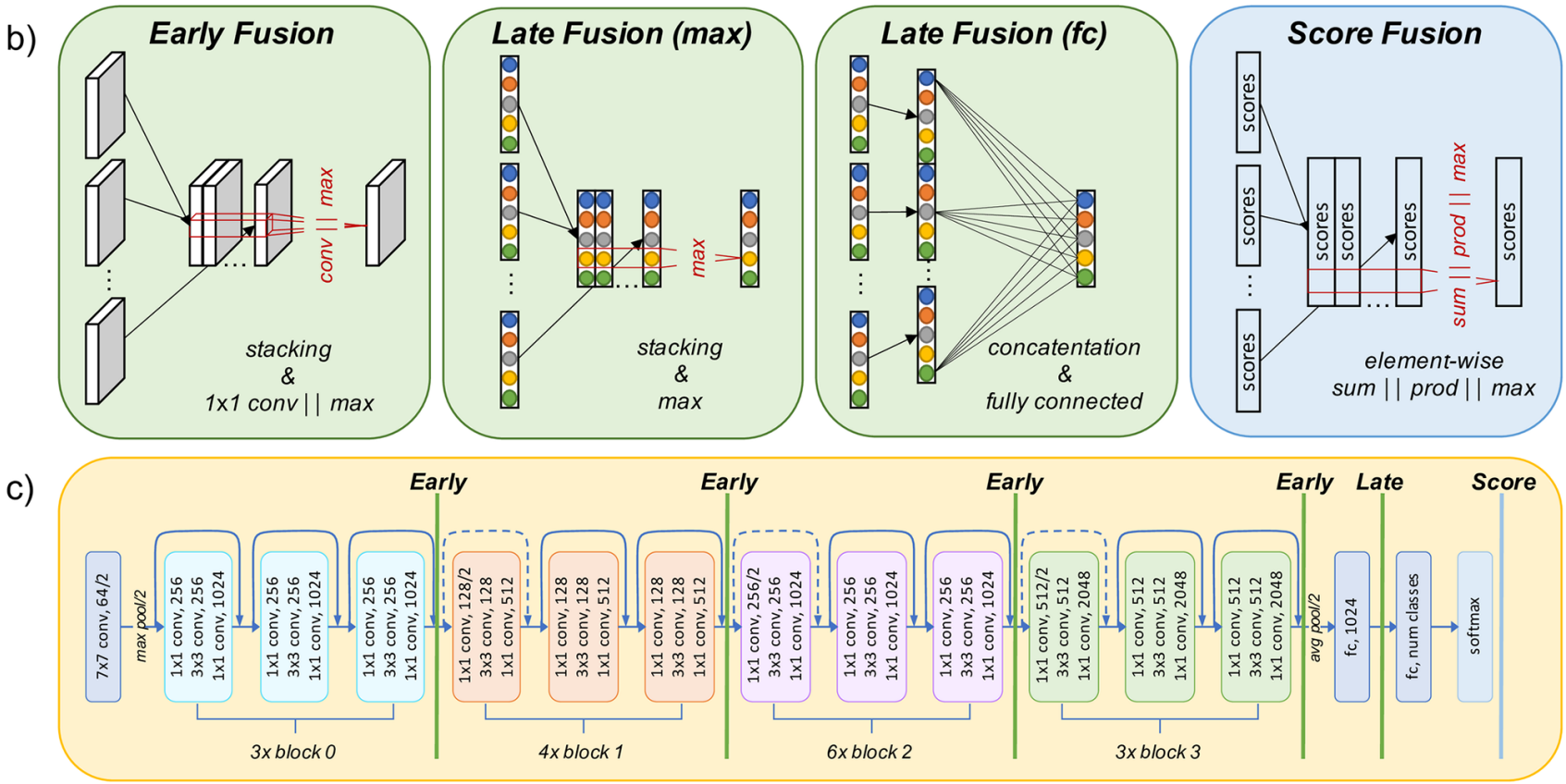
2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage

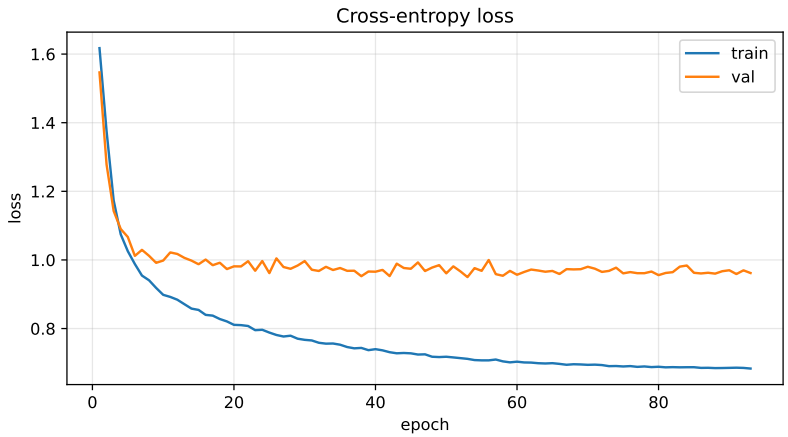
5 Résultats et analyses
- Les modèles single-view
- Les modèles multi-view
  - Late fusion
  - Score fusion
- Les modèles VLM
- Synthèse des résultats

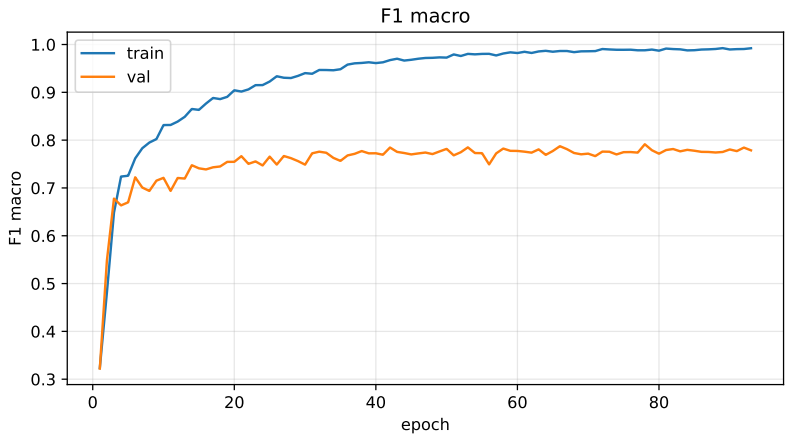
Rappel des stratégies



Source : figure de [4]

Courbes d'apprentissage





↪ Un apprentissage **beaucoup plus lent**, malgré le *backbone* de `ResNet` gelé.

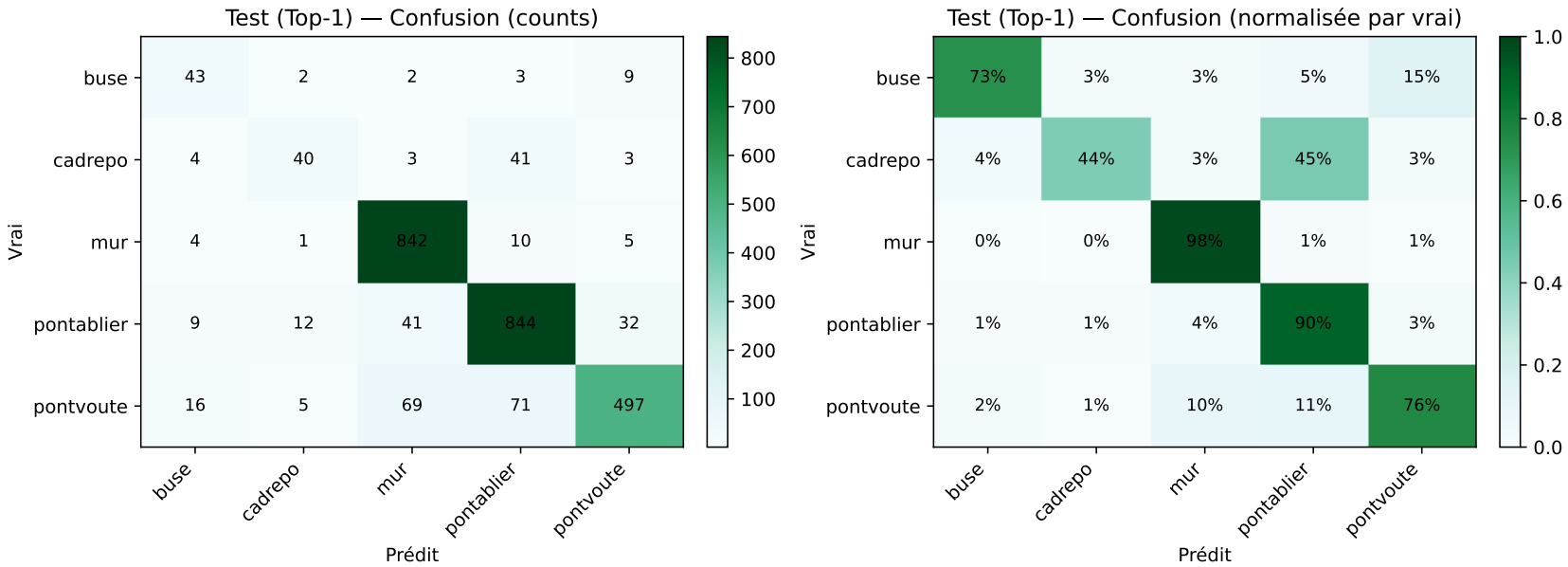
Late fusion - ResNet-50 - 89 *epochs*

Classe	Précision	Rappel	F1-score	Support
buse	0.566	0.729	0.637	59
cadrepo	0.667	0.440	0.530	91
mur	0.880	0.977	0.926	862
pontablier	0.871	0.900	0.885	938
pontvoute	0.910	0.755	0.826	658
Moyenne (macro)	0.779	0.760	0.761	2608
Moyenne pondérée	0.870	0.869	0.866	2608

	Accuracy	**F1-macro**
Top-1	0.8689	0.7607
Top-2	0.9467	0.9015

Table – Résultats globaux — Late Fusion (ResNet-50).

Matrices de confusion - Top-1



1 Contexte et objectifs du stage

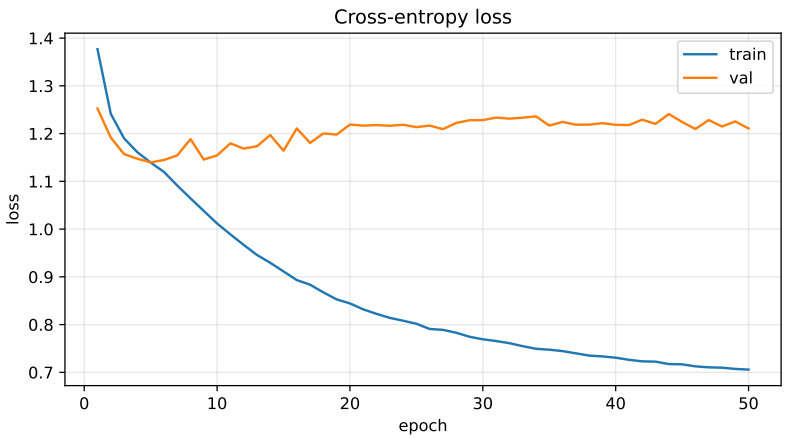
2 Présentation des données

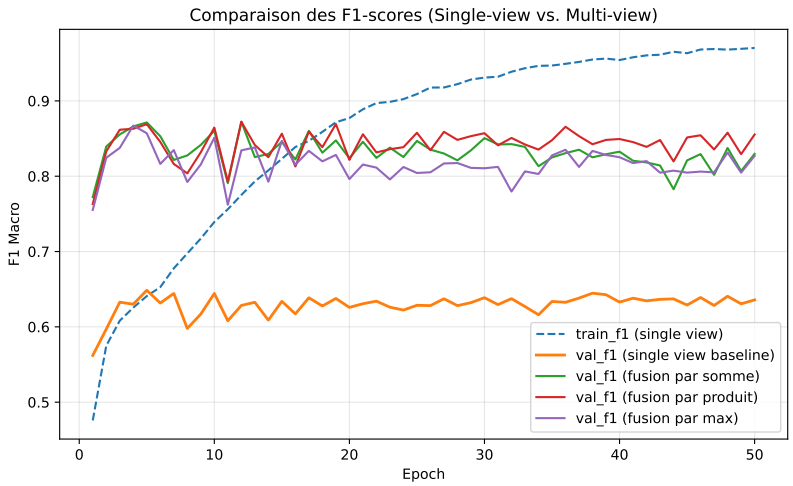
3 État de l'art

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses
- Les modèles single-view
- Les modèles multi-view
  - Late fusion
  - Score fusion
- Les modèles VLM
- Synthèse des résultats

Courbes d'apprentissage



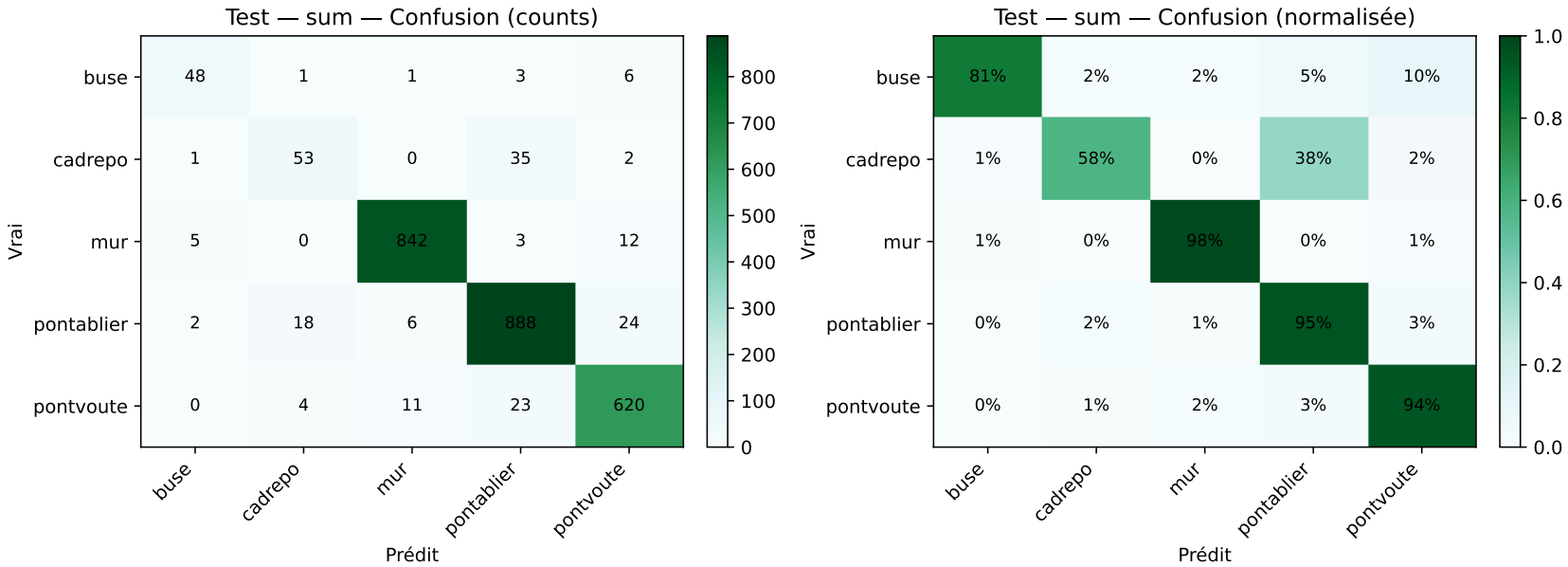


La courbe F1 inclut les performances de validation selon les **trois stratégies de fusion**.

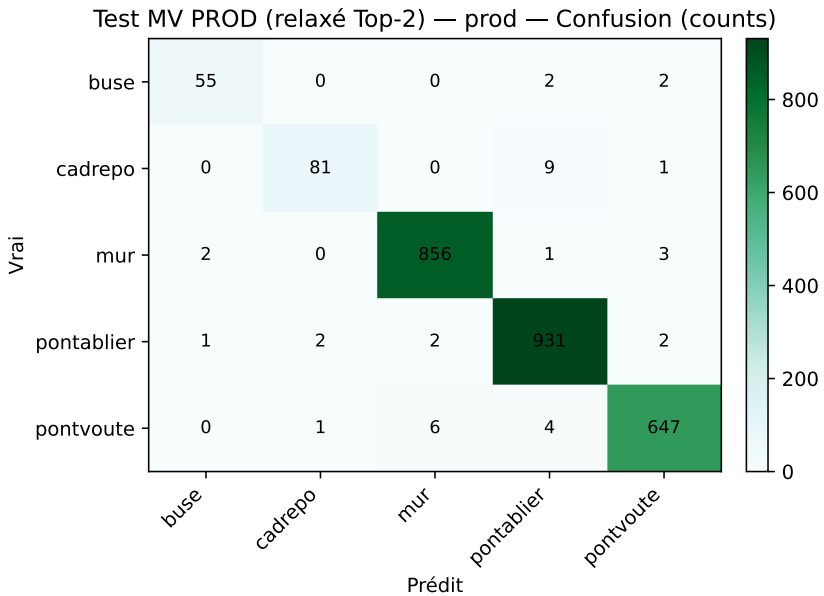
Résultats

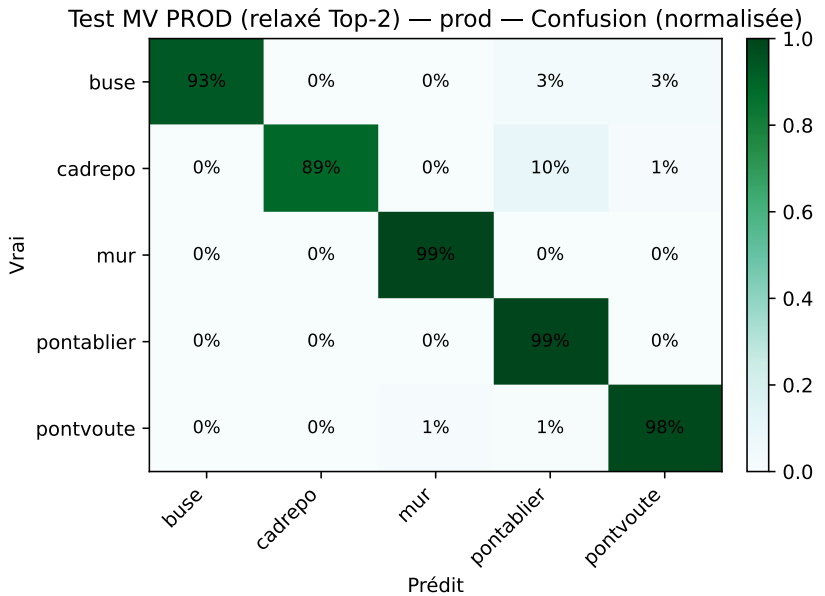
Expérience	Fusion	Top-1 Acc.	Top-1 F1	Top-2 Acc.	Top-2 F1
ResNet-18	SUM	0.9310	0.8463	0.9635	0.8951
ResNet-18	PROD	0.9282	0.8281	0.9587	0.8816
ResNet-18	MAX	0.9047	0.8272	0.9458	0.8743
ResNet-50	SUM	**0.9398**	**0.8650**	**0.9682**	**0.9183**
ResNet-50	PROD	**0.9398**	**0.8650**	**0.9682**	**0.9183**
ResNet-50	MAX	0.9260	0.8390	0.9541	0.8938

Matrices de confusion - Top-1 - Méthode *PROD*



Matrices de confusion - Top-2 - Méthode *PROD*





1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses
- Les modèles single-view
- Les modèles multi-view
- Les modèles VLM
- Synthèse des résultats

6 Conclusion et perspectives

CLIP en *zero-shot learning*

Principe

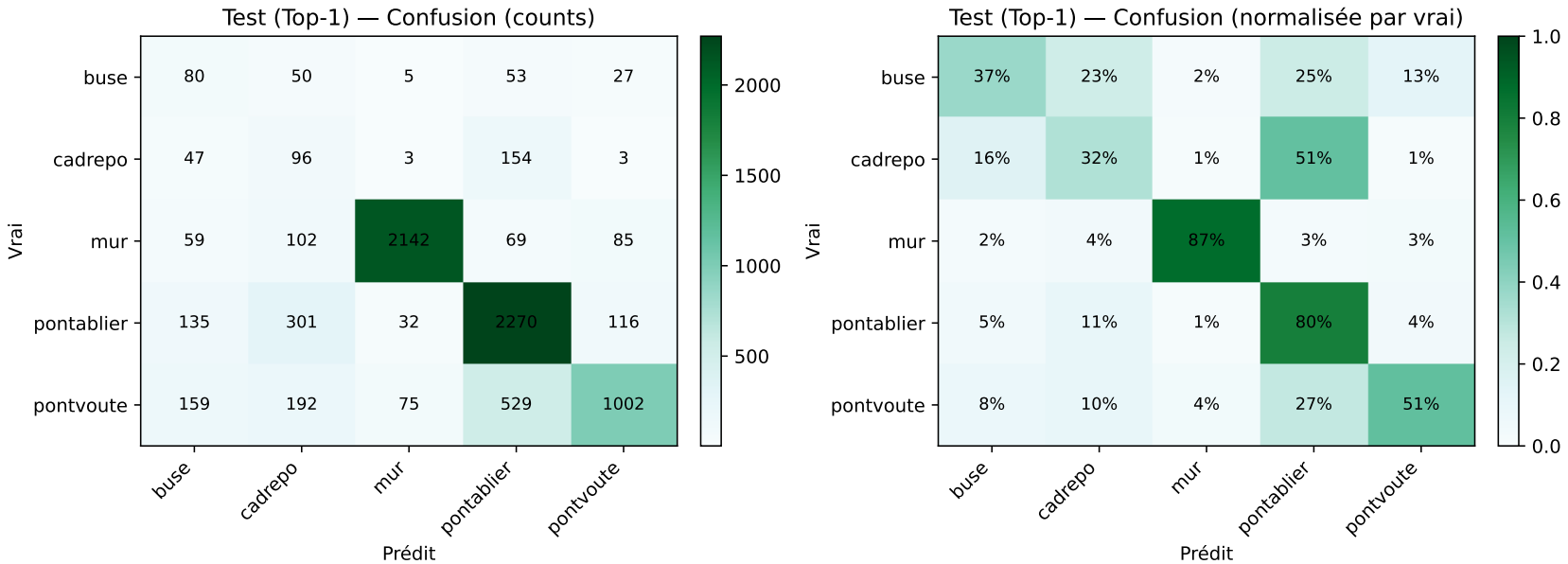
- **CLIP** [13] associe une image à la **description textuelle la plus cohérente** parmi une liste de phrases candidates.
- Aucun ré-entraînement.

```
"photo d'un {}", "photo de {}", "a photo of a {}",
"vue d'élévation d'un {}",  "an elevation view of a {}",
"une image d'un {}", "an image of a {}"
```

Résultats

	Accuracy	F1-macro
Top-1	0.7180	0.5434
Top-2	0.8422	0.6960

Table – Résultats globaux — CLIP en *zero-shot learning*.



1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses
- Les modèles single-view
- Les modèles multi-view
- Les modèles VLM
- Synthèse des résultats

6 Conclusion et perspectives

Synthèse des résultats

Configuration	Top-1 Acc.	Top-1 F1	Top-2 Acc.	Top-2 F1
Single-view (RN-18, baseline)	0.8138	0.7035	0.9127	0.8051
Single-view (RN-50)	0.8695	0.7714	0.9420	0.8533
Multi-view (Score Fusion, RN-50)	**0.9398**	**0.8650**	**0.9682**	**0.9183**
Multi-view (Late Fusion, RN-50)	0.8689	0.7607	0.9467	0.9015
VLM – CLIP (Zero-shot)	0.7180	0.5434	0.8422	0.6960

1 Contexte et objectifs du stage

2 Présentation des données

3 État de l'art

4 Chaîne de traitement pour l'apprentissage

5 Résultats et analyses

6 Conclusion et perspectives

Conclusion

Bilan du stage

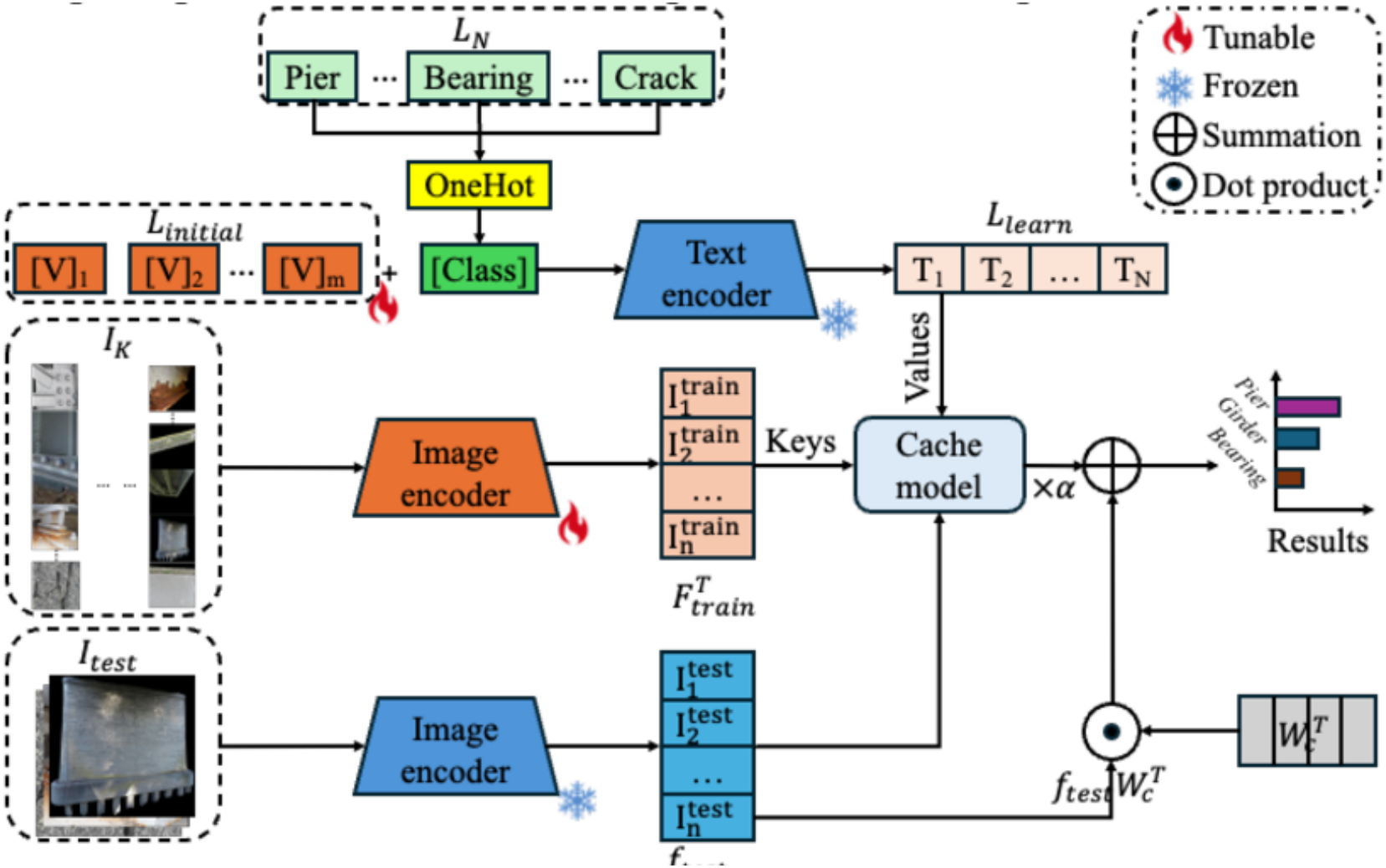
- **Besoin identifié** : automatiser la chaîne d'inspection des ouvrages d'art.
- **Objectif du stage** : évaluer le potentiel des **VLM**.
- **Démarche** :
  - Mise en qualité et préparation de la base **SIRO**.
  - Étude bibliographique.
  - Implémentation et comparaison de modèles *single-view*, *multi-view* et VLM.
- **Résultats** :
  - Meilleurs résultats avec **ResNet-50 multi-vues (score fusion)**.
  - Données PNP exploitables et prometteuses pour l'apprentissage profond.

Perspectives

Pistes de poursuite

- **Court terme** : mise en œuvre d'un **VLM spécialisé ouvrages d'art** via *fine-tuning* ([11]).
  ↪ Slide suivante...
- **Long terme** : suite du projet ApPLAuDI.

Le prochain objectif (dernier mois de stage)



Source : figure de [11]

Bibliographie I

[1] Kaiming He et al. « Deep Residual Learning for Image Recognition ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 770-778. URL : https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

[2] Simon Kornblith, Jonathon Shlens et Quoc V. Le. « Do Better ImageNet Models Transfer Better ? » In : *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, p. 2661-2671. URL : https://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html.

Bibliographie II

[3] Guillaume Charpiat. *Deep Learning in Practice*. https://www.lri.fr/~gcharpia/deeppractice. Cours du Master MVA, Université Paris-Saclay / ENS Paris-Saclay. 2025.

[4] Marco Seeland et Patrick Mäder. « Multi-view classification with convolutional neural networks ». In : *PLOS ONE* (jan. 2021). ISSN : 1932-6203. DOI : 10.1371/journal.pone.0245230.

[5] Xiaoqiang Yan et al. « Deep multi-view learning methods : A review ». In : *Neurocomputing* (août 2021). ISSN : 0925-2312. DOI : 10.1016/j.neucom.2021.03.090. (Visité le 09/10/2025).

Bibliographie III

[6] Carmen Andrade von Hillebrandt. *Two-Step Deep Learning Approach for Classifying Bridges using Street Imagery*. Juin 2025. URL : https://cs231n.stanford.edu/papers/Carmen_Final_CV%20-%20Carmen%20Edna%20Andrade%20von%20Hillebrandt.pdf.

[7] Jingyi Zhang et al. « Vision-Language Models for Vision Tasks : A Survey ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (août 2024). ISSN : 1939-3539. DOI : 10.1109/TPAMI.2024.3369699.

[8] Florian Bordes et al. *An Introduction to Vision-Language Modeling*. Mai 2024. DOI : 10.48550/arXiv.2405.17247.

Bibliographie IV

[9] Edward J. Hu et al. *LoRA : Low-Rank Adaptation of Large Language Models*. Oct. 2021. DOI : 10.48550/arXiv.2106.09685.

[10] Kaiyang Zhou et al. « Learning to Prompt for Vision-Language Models ». In : *International Journal of Computer Vision* (sept. 2022). ISSN : 1573-1405. DOI : 10.1007/s11263-022-01653-1. URL : https://doi.org/10.1007/s11263-022-01653-1.

Bibliographie V

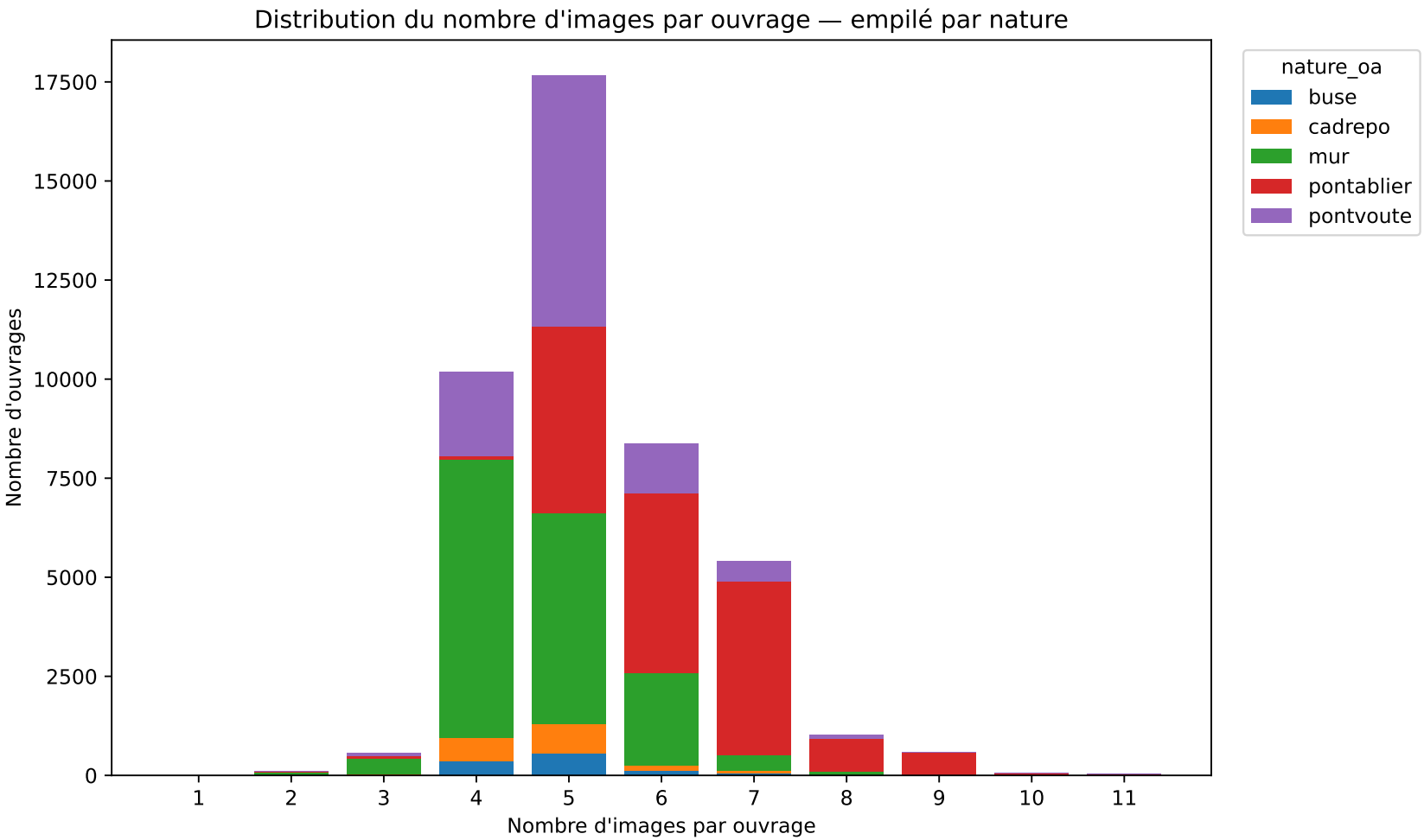
[11] Zhengxing Chen et al. « Bridge Inspection Using A Multi-Modal Vision Language Model ». In : *Proceedings of The Sixth International Conference on Civil and Building Engineering Informatics*. Sous la dir. de Jack Cheng et Yu Yantao. Kalpa Publications in Computing. EasyChair, 2025, p. 578-588. DOI : 10.29007/m7wj.

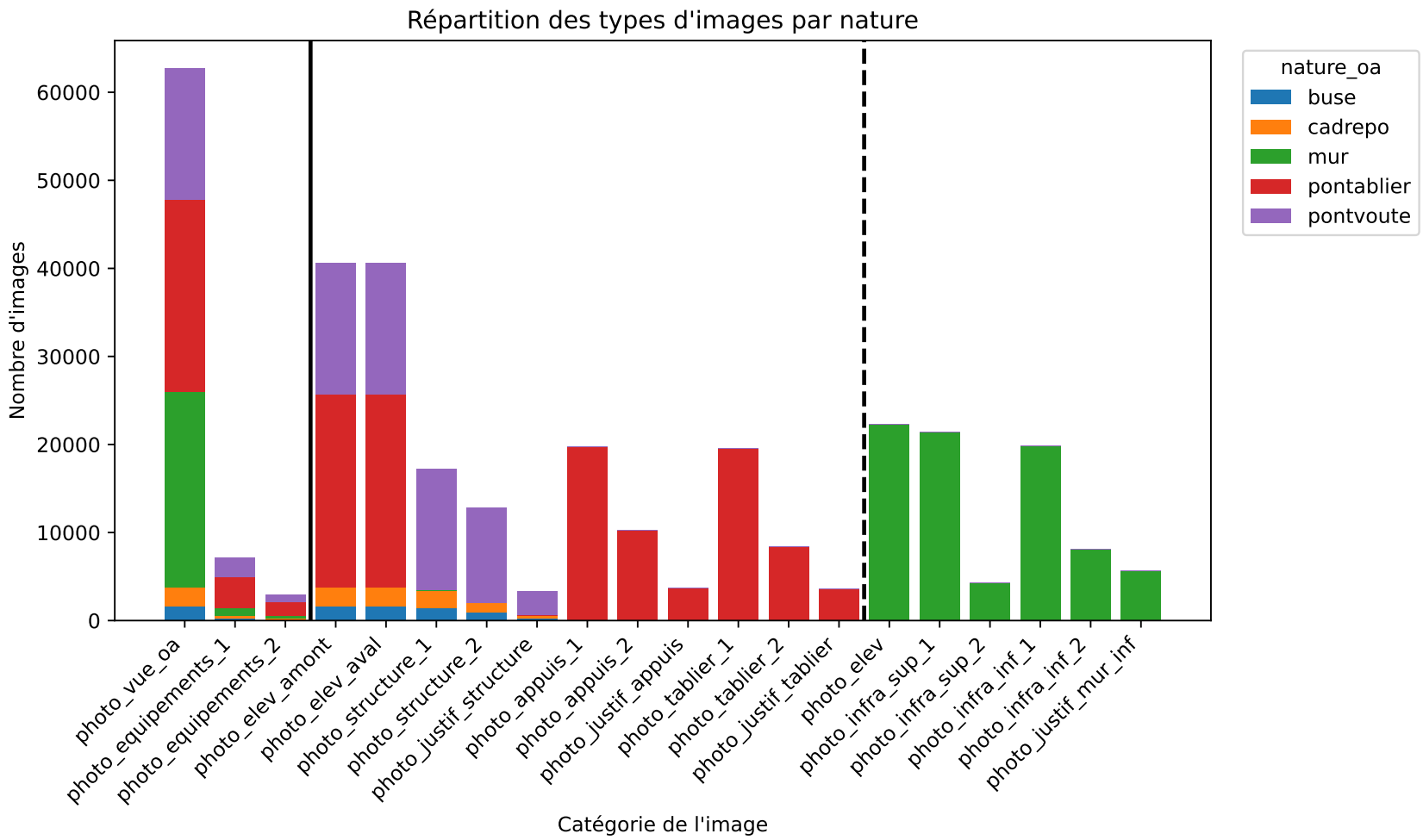
[12] Chenhong Zhang et al. « Automatic bridge inspection database construction through hybrid information extraction and large language models ». In : *Developments in the Built Environment* (déc. 2024). ISSN : 2666-1659. DOI : 10.1016/j.dibe.2024.100549.

Bibliographie VI

[13] Alec Radford et al. « Learning Transferable Visual Models From Natural Language Supervision ». In : *Proceedings of the 38th International Conference on Machine Learning*. PMLR, juill. 2021.

5 images par ouvrage (en moyenne)




Répartition des types d'images par nature
Nombre d'images
0
10000
20000
30000
40000
50000
60000
nature_oa
buse
cadrepo
mur
pontablier
pontvoute
photo_vue_oa
photo_equipements_1
photo_equipements_2
photo_elev_amont
photo_elev_aval
photo_structure_1
photo_structure_2
photo_justif_structure
photo_appuis_1
photo_appuis_2
photo_justif_appuis
photo_tablier_1
photo_tablier_2
photo_justif_tablier
photo_elev
photo_infra_sup_1
photo_infra_sup_2
photo_infra_inf_1
photo_infra_inf_2
photo_justif_mur_inf
Catégorie de l'image



Élévation du mur



Vue générale



Défauts inférieurs



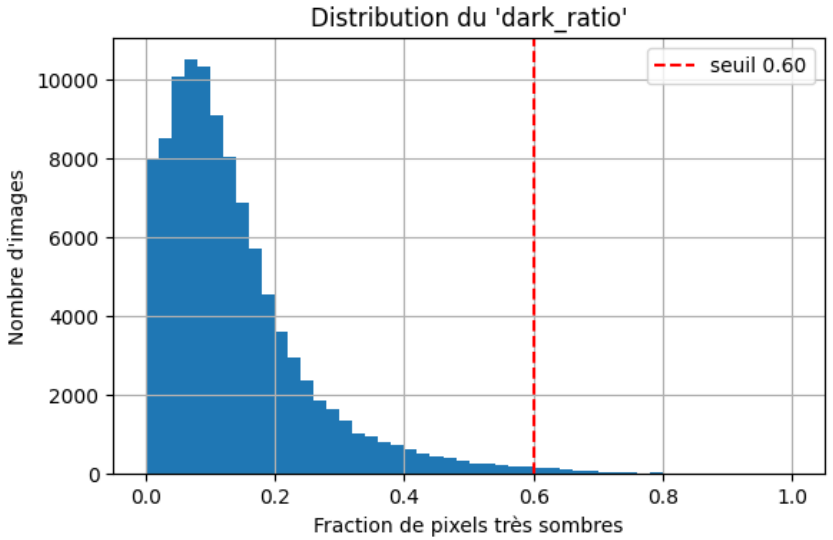
Défauts supérieurs

Figure – Exemple de mur de soutènement — Curemonte (19)

Détection des images sombres

Méthdologie

1. Conversion des images en format HSV.
2. Un pixel est considéré comme trop sombre si $V \geq v_{\text{max}}$, avec $v_{\text{max}} = 0.6$.
3. Suppression des photos avec au moins 60% de pixels sombres par image.

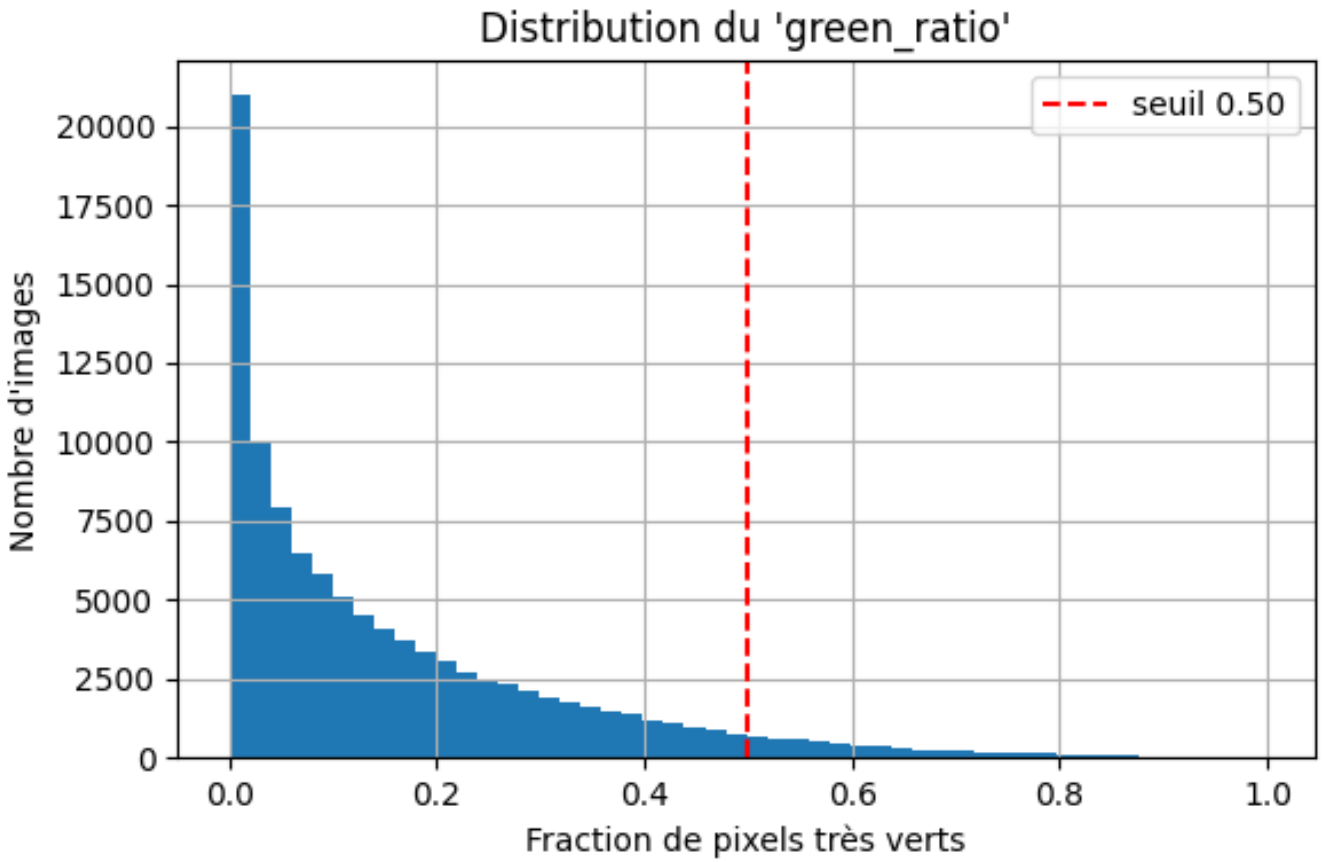


Résultats

878 images trop sombres ou noires supprimées.



Histogramme des images « vertes »



Sous-échantillonner la base ?

La problématique

Les classes **buse** et **cadrepo** sont extrêmement minoritaires.
Comment pallier à cela ?

Sous-échantillonner la base ?

La problématique

Les classes **buse** et **cadrepo** sont extrêmement minoritaires.
Comment pallier à cela ?

Stratégie

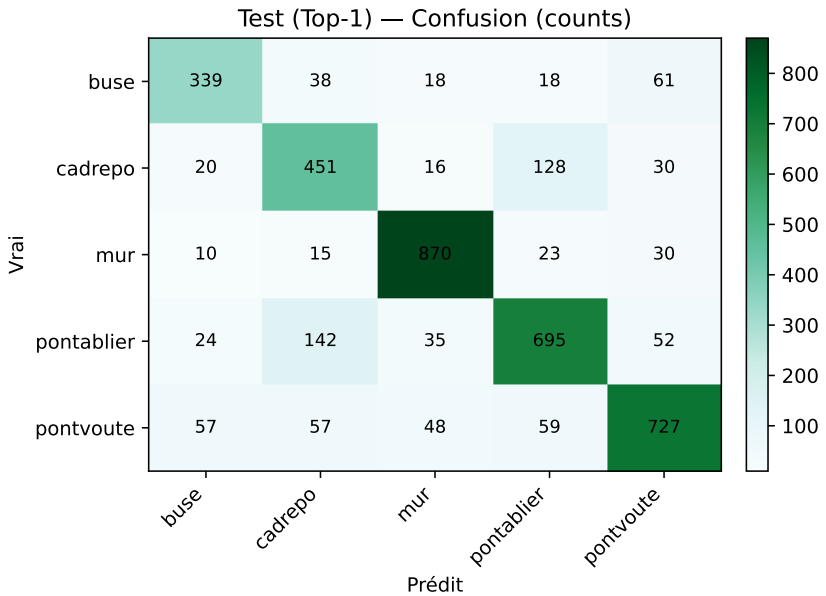
Pour toute les classes c :

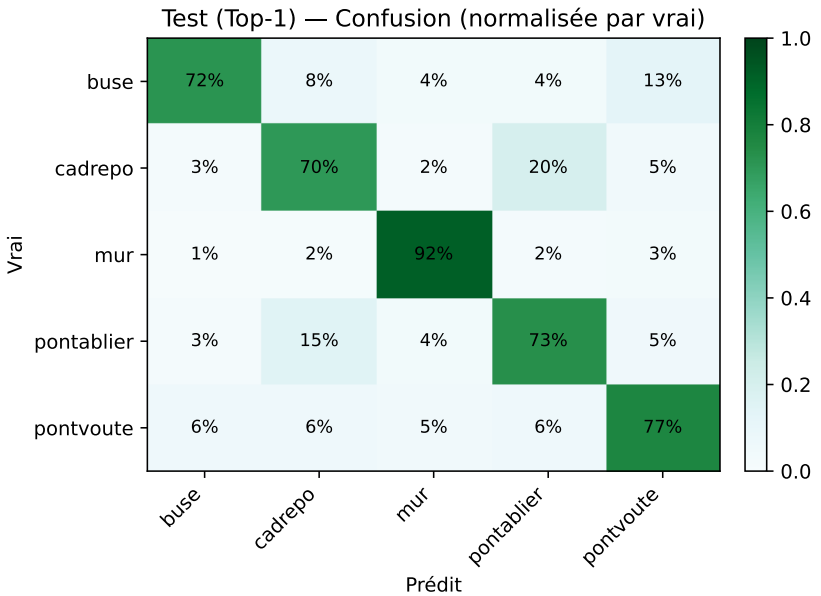
$$N_c = k \times N_{\text{min}}$$

où N_c est le nombre d'échantillons conservés pour la classe c, N_{min} étant l'effectif minimal d'une classe, et $k \in \{1; 2\}$.

Résultat du sous-échantillonnage ($k = 2$)

	Accuracy	F1-macro
Top-1	0.7777	0.7664
Top-2	0.9069	0.9028





Mauvaise détection - Buse



Mauvaise détection - Mur



Mauvaise détection - Pont tablier



Mauvaise détection - Pont voûte

Classe cible: pontvoute
Haut: Faux-positifs | Bas: Faux-négatifs
Ordre des probabilités: ['buse', 'cadrepo', 'mur', 'pontablier', 'pontvoute']

[0.06 0.05 0.03 0.01 0.85]
Vrai: pontablier | Prédit: pontvoute



[0.15 0.07 0.04 0.02 0.72]
Vrai: buse | Prédit: pontvoute



[0.07 0.07 0.03 0.20 0.64]
Vrai: cadrepo | Prédit: pontvoute



[0.09 0.06 0.24 0.07 0.55]
Vrai: pontablier | Prédit: pontvoute



[0.11 0.08 0.03 0.61 0.16]
Vrai: pontvoute | Prédit: pontablier



[0.07 0.07 0.05 0.41 0.39]
Vrai: pontvoute | Prédit: pontablier



[0.09 0.07 0.49 0.03 0.32]
Vrai: pontvoute | Prédit: mur



[0.11 0.07 0.43 0.10 0.29]
Vrai: pontvoute | Prédit: mur

