Context-Aware Semantic Segmentation via Stage-Wise Attention

Antoine Carreaud[1 2] · Elias Naha[1 2] · Arthur Chansel[1 2] · Nina Lahellec[1 2] · Jan Skaloud[1] · Adrien Gressin[2]

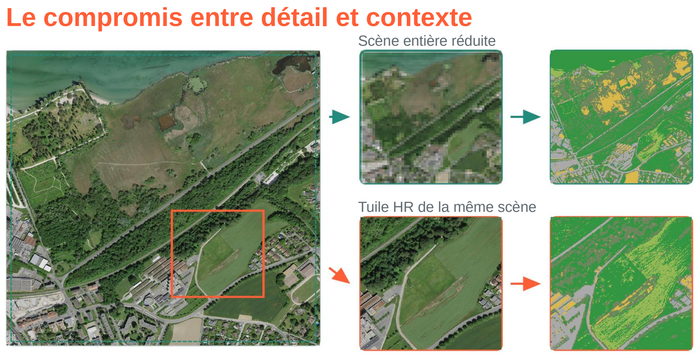
[1] EPFL, Lausanne, Suisse · [2] HEIG-VD / HES-SO, Yverdon-les-Bains, Suisse

Motivation

Le contexte et la très haute résolution

En très haute résolution (UHR), la segmentation doit préserver les contours fins tout en exploitant le contexte global.

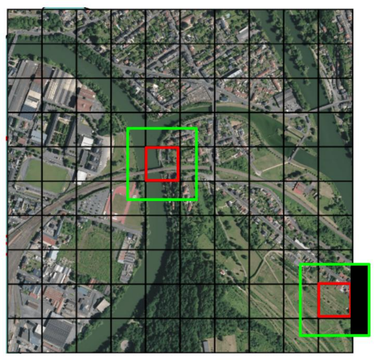
Le compromis entre détail et contexte



Réduire la scène entière
Contexte conservé, détails perdus

Extraire une tuile HR
Détails conservés, contexte perdu

Protocole UHR sur FLAIR-HUB



La tuile HR sert à l'évaluation. La tuile LR associée fournit le contexte de la scène.

Cible HR	Tuile de 512 × 512 à segmenter
Tuile LR	Voisinage 3 × 3 co-enregistré

241 k tuiles	15 classes	0,20 m au sol

Hypothèses de CASWiT

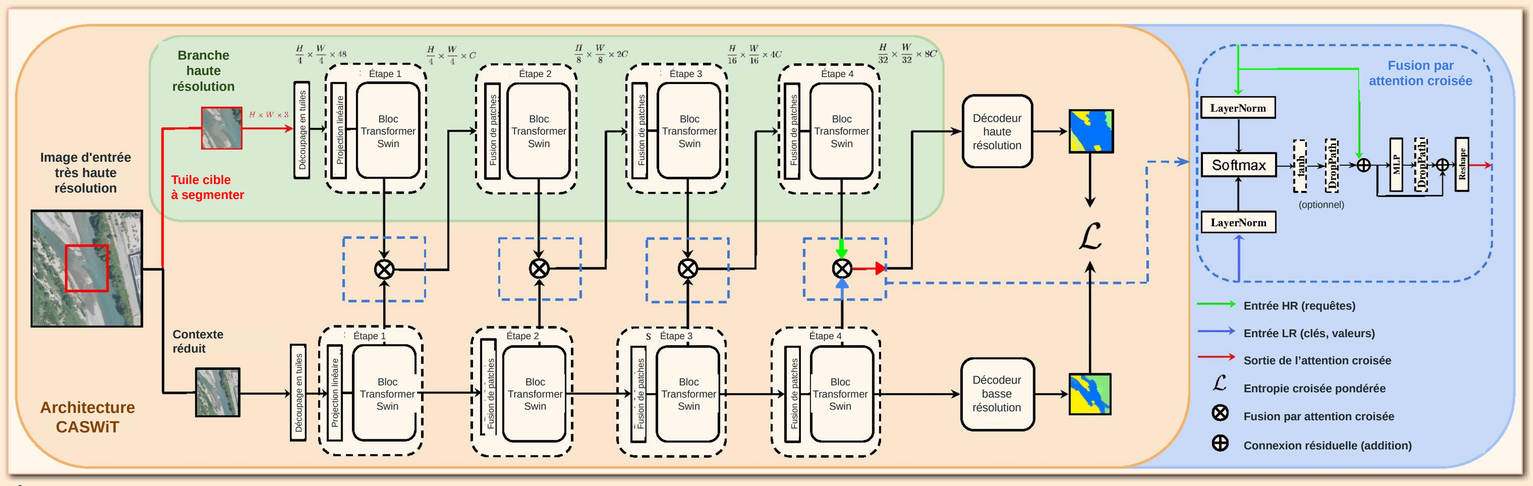
1. La branche HR préserve les détails de la tuile, les contours fins et les petits objets.

2. La branche LR encode le contexte du voisinage géoréférencé autour de la tuile.

3. Le contexte LR intervient tôt, avant que les représentations HR ne deviennent trop locales.

Méthode

CASWiT : attention croisée HR→LR à chaque étape



À chaque étape, l'attention croisée HR→LR apporte du contexte aux représentations haute résolution.

Injection du contexte à chaque étape

$$Q = LN(X_s^{HR})W_Q, \quad K,V = LN(X_s^{LR})W_{K,V}$$
$$A_s = MHA(Q,K,V), \quad \tilde{H}_s = X_s^{HR} + \gamma_s \odot A_s + MLP(H'_s)$$

Les requêtes HR utilisent les clés et valeurs LR pour intégrer le contexte à chaque étape.

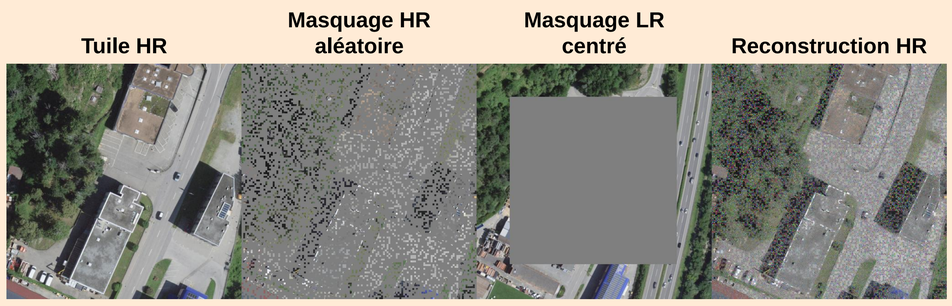
Objectif d'apprentissage

$$\mathcal{L}_{HR} = CE(\hat{Y}^{HR}, Y), \quad \mathcal{L}_{LR} = CE(\hat{Y}^{LR}, Y\downarrow)$$
$$\mathcal{L} = \mathcal{L}_{HR} + \alpha\,\mathcal{L}_{LR}, \quad \alpha = 0.5$$

CE est l'entropie croisée par pixel. La tête LR auxiliaire régularise l'apprentissage et est retirée à l'inférence.

Apprentissage auto-supervisé

Apprendre les relations entre échelles par masquage



SimMIM à deux branches : masquage HR aléatoire et LR centré.

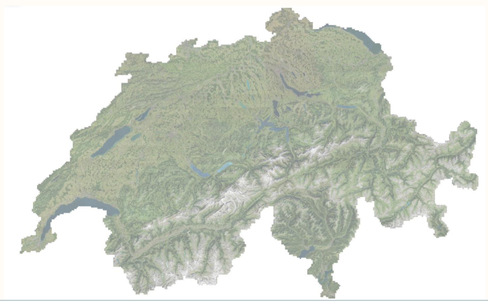
Perte de reconstruction des pixels HR masqués

$$\mathcal{L}_{SSL} = \frac{1}{3\,|M_{HR}^{pix}|} \sum_{p \in M_{HR}^{pix}} \left\| \hat{I}_p^{HR} - I_p^{HR} \right\|_1$$

M_HR est le masque HR ramené à la résolution des pixels. Seuls les pixels HR masqués sont reconstruits.

Corpus de préentraînement : SWISSIMAGE [10]

Orthophotos suisses non annotées



1 067 Gpx	44 000 km²	0,20 m

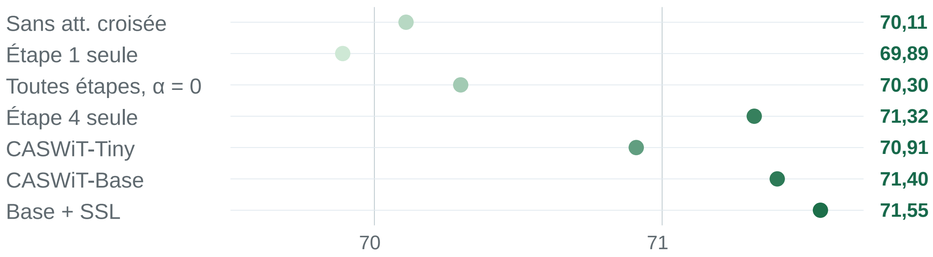
Résultats

Précision, qualité des contours et transfert

FLAIR-HUB	URUR	ISIC	CRAG
66,37 mIoU	49,2 mIoU	86,5 mIoU	90,7 mIoU
+2,32 pts [1]	+1,0 pt [3]	+3,1 pts [3]	+0,4 pt [3]

Étude d'ablation

mIoU en validation sur FLAIR-HUB



Contrôle de capacité

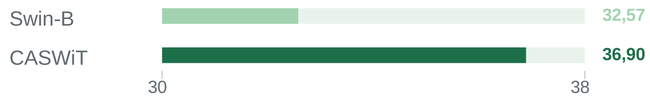
Mêmes entrées HR et LR

Fusion tardive	64,25
CASWiT	66,37

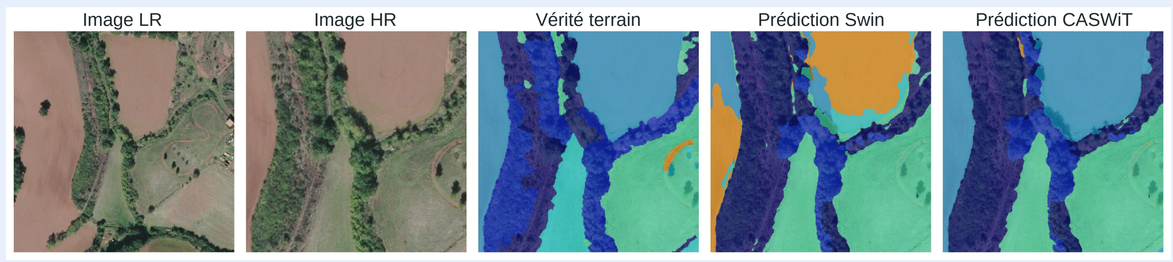
+2,12 vs somme simple

Qualité des contours

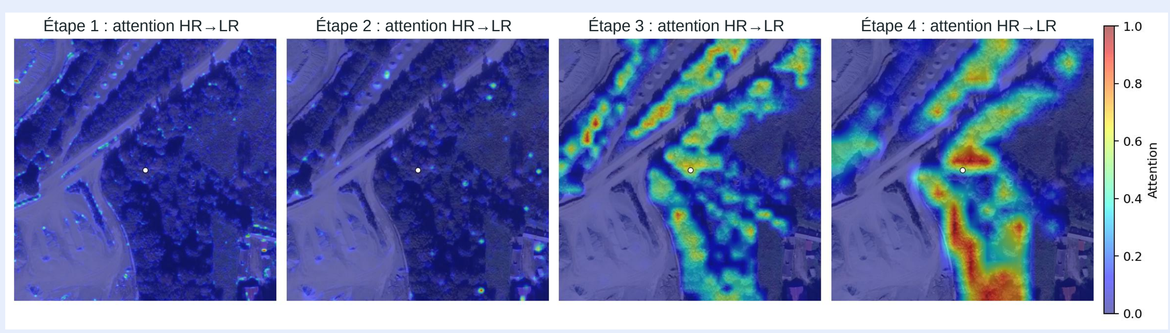
mBIoU sur FLAIR-HUB



+4,33
Gain mBIoU



Exemple qualitatif : CASWiT suit mieux les contours fins de la végétation et réduit le lissage excessif par rapport à Swin-B.



L'attention passe des textures locales aux premières étapes aux structures globales de la scène aux étapes suivantes.

Conclusions

1. L'attention HR→LR à chaque étape apporte du contexte tout en préservant les détails de la tuile.
2. Les ablations à capacité contrôlée montrent que le gain ne vient pas du seul ajout d'une seconde branche.
3. CASWiT améliore la mIoU et les contours, avec un transfert vers l'imagerie aérienne et médicale.

Références

[1] A. Garioud, S. Giordano, N. David, N. Gonthier. FLAIR-HUB: Large-scale Multimodal Dataset for Land Cover and Crop Mapping. arXiv, 2025.
[2] Z. Liu et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV, 2021.
[3] R. Qin, X. Liu, J. Shi, L. Lin, J. Yang. Boosting the Dual-Stream Architecture in UHR Segmentation. CVPR, 2025.
[4] D. Ji, F. Zhao, H. Lu, M. Tao, J. Ye. Ultra-High Resolution Segmentation With Ultra-Rich Context. CVPR, 2023.
[5] E. Xie et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. NeurIPS, 2021.
[6] Z. Xie et al. SimMIM: A Simple Framework for Masked Image Modeling. CVPR, 2022.
[7] B. Cheng, R. Girshick, P. Dollar, A. C. Berg, A. Kirillov. Boundary IoU. arXiv, 2021.
[8] P. Tschandl, C. Rosendahl, H. Kittler. The HAM10000 dataset. Scientific Data, 2018.
[9] S. Graham et al. MILD-Net: Minimal information loss dilated network for gland segmentation. Med. Image Anal., 2019.
[10] Federal Office of Topography swisstopo. SWISSIMAGE: digital color orthophotomosaic of Switzerland, 0.20 m GSD.

Démonstration
Carte de segmentation de la Suisse
rar-indg.heig-vd.ch/map/Suisse.html

Code et contact
huggingface.co/collections/heig-vd-geo/caswit
antoine.carreaud@heig-vd.ch