Apprentissage autosupervisé de représentations spatio-temporelles de séries temporelles d'images satellites

Iris DUMEUR[1] 2, Silvia VALERO[1] , Jordi INGLADA[1]

[1]CESBIO, Université de Toulouse, CNES/CNRS/INRAe/IRD/UPS 31000 Toulouse, France

Résumé – Nous présentons une nouvelle stratégie d'apprentissage autosupervisée de représentations de séries temporelles d'images satellites (STIS), dénommée U-BARN. L'architecture conçue, qui mêle Unet et Transformer, exploite la synergie entre les dimensions spatio-spectrales et temporelles. Afin de pré-entrainer U-BARN sur des données non étiquetées, une tâche prétexte de reconstruction de séries temporelles inspirée du modèle BERT est proposée. Pour démontrer sa capacité d'extraction de primitives pertinentes, les représentations issues de U-BARN pré-entrainé sont ensuite utilisées pour générer des cartes de segmentation sémantique. Les performances de classification démontrent l'intérêt de notre stratégie de pré-entrainement.

Abstract – We present a new self-supervised strategy for learning meaningful representations of complex optical Satellite Image Time Series (SITS), named U-BARN. The designed architecture mixing Unet and Transformer enhances the synergy between spatio-spectral and temporal dimensions. To pre-train U-BARN on unlabelled data, a time series reconstruction pretext task inspired by the BERT strategy is proposed. To demonstrate its feature learning capability, representations of SITS encoded by U-BARN, are then used to generate semantic segmentation maps. Experimental results, on a labelled dataset, corroborate that the self-supervised training strategy is consistent.

1 Introduction

L'apprentissage profond (*Deep Learning, DL*) est récemment devenu un outil essentiel pour l'analyse de séries temporelles d'images satellites (STIS). Ces données de haute résolution temporelle, spatiale et spectrale sont essentielles pour la surveillance de notre planète. Elles permettent, entre autre, de cartographier l'occupation des sols sur des grandes étendues. Néanmoins, il reste des défis importants à relever pour les architectures de DL afin d'exploiter les spécificités des STIS. En particulier, l'exploitation de la synergie entre les dimensions spatiale, spectrale et temporelle est complexe.

Tout d'abord, la dynamique spectro-temporelle étant cruciale pour l'analyse des sols, les premières architecture DL exploitant des STIS étaient exclusivement temporelles [4], [8]. De récentes études [1], [5],[7] ont démontré que de meilleurs résultats pouvaient être obtenus en prenant également en compte la dimension spatiale. Ces premières approches spatio-temporelles peuvent être classées en deux catégories. D'une part, certaines combinent des réseaux de neurones convolutifs (CNN) avec des réseaux de neurones récurrents (RNN) [1], [5] [7]. D'autre part, d'autres méthodes sont totalement convolutionelles [6]. Néanmoins, l'utilisation de RNN ou de convolutions temporelles pour exploiter la dimension temporelle n'est pas parfaitement adaptée au STIS. En effet, ces architectures ne peuvent pas prendre en compte des séries temporelles irrégulières et une interpolation est ainsi nécessaire en amont. Par ailleurs, elles ne peuvent que capter des relations temporelles de courtes portées, alors que pour les STIS le début et la fin d'une année peuvent être très fortement corrélés. Afin de contourner ces limitations, les approches les plus récentes utilisent un mécanisme d'attention temporel défini dans le Transformer [10], une architecture initialement développée pour le traitement du langage naturel. L'architecture spatio-temporelle la plus récente, le U-TAE [3] propose en particulier la fusion d'un Unet avec le Transformer. Le mécanisme d'attention temporelle y est placé dans le goulot du Unet. Bien que cela permette de réduire la complexité calculatoire du réseau, le mécanisme d'attention est calculé à la plus basse résolution spatiale, pouvant probablement dégrader les performances de classification.

Malgré les résultats prometteurs obtenus par ces architectures existantes, elles ne peuvent pas être appliquées dans de nombreuses applications de télédétection à grande échelle en raison du manque de disponibilité et de qualité des données de référence nécessaires à l'entraînement de modèles de DL. Au cours des dernières années, l'apprentissage autosupervisé est apparu comme une solution potentielle pour réduire, voire éliminer, le besoin de construire des jeux de données étiquetés. En effet, l'apprentissage autosupervisé de représentations consiste à pré-entraîner des modèles profonds sur de grands jeux de données non étiquetées, afin d'être par la suite utilisés pour résoudre des tâches en aval possédant un faible nombre de données étiquetées. En particulier, cela permet l'entraînement sur une tâche en aval de modèles peu profonds qui utilisent les représentations générées par des modèles complexes pré-entrainés. Dans le domaine de la vision par ordinateur, les approches autosupervisées récentes reposent principalement sur l'apprentissage constrastif qui nécessite d'augmenter les données. Toutefois, l'augmentation de données pour les séries temporelles multispectrales n'est pas simple. Ainsi, les stratégies basées sur des tâches prétextes génératives sont actuellement davantage envisagées pour les données temporelles. Par exemple, deux approches autosupervisées, la méthode temporelle SITS-BERT [11] et spatio-temporelle SITS-Former [12] ont été appliquées sur des STIS. Ces deux approches s'inspirent du modèle BERT [2], initialement développé dans le domaine du langage. Leurs stratégies de pré-entraînement tentent d'apprendre la structure des données via une tâche de reconstruction consistant à restaurer

des informations masquées dans les données. Cependant, la version spatio-temporelle, le SITS-Former [12], ne considère qu'un faible voisinage spatial et n'est pas efficace pour la génération de cartes de classification. Outre ces limitations architecturales, la tâche prétexte proposée n'exploite pas suffisamment les différences entre le domaine du langage et les STIS. La stratégie de masquage proposée (taux de masquage et valeur du masque), inspirée de [2], n'est pas adaptée aux STIS. Compte tenu de tout ce qui précède, une nouvelle méthode autosupervisée, appelée **U-BARN**, est présentée ici pour apprendre des représentations pertinentes de séries temporelles d'images satellitaires optiques. Nos contributions peuvent être résumées ainsi : (i) nous avons conçu une nouvelle architecture capturant les informations spatio-temporelles contenues dans des STIS irrégulièrement échantillonnées ; (ii) nous avons défini une stratégie d'entraînement autosupervisée adaptée au STIS, afin d'apprendre des représentations latentes de haut niveau sémantique. Pour évaluer les performances de notre approche, nous avons utilisé la tâche en aval de segmentation sémantique proposée par le jeu de données annotées PASTIS [3].

2 Méthode proposée

Cette section détaille l'architecture du réseau de neurones U-BARN ainsi que la tâche prétexte générative proposée.

2.1 Architecture U-BARN

U-BARN (*Unet-Bert SpAtio-temporal eNcoder*) est divisé en deux blocs : (i) un encodeur d'images (représenté en gris dans Figure 1) qui calcule une représentation spectro-spatiale indépendante pour chaque image de la STIS et (ii) un Transformer qui capture les relations temporelles entre les images ainsi encodées. Les représentations générées par U-BARN sont de même résolution temporelle et spatiale que la STIS d'entrée. Ainsi, étant donné un batch de taille (B,T,C,H,W) avec B la dimension du batch, T la dimension temporelle, C la dimension spectrale ainsi que H et W les dimensions spatiales, la représentation latente est de taille (B,T,d_{model},H,W).

2.1.1 Encodeur d'images

Tout d'abord, la série temporelle d'images est traitée par un encodeur spectro-spatial (ESS), qui encode individuellement chaque image d'entrée en une carte de caractéristiques de taille $H \times W \times d_{model}$. L'architecture proposée de l'ESS est un Unet doté de 4 blocs de sous et sur-échantillonnage, conçue pour capturer des primitives spectrales et spatiales riches avec champ de vue large.

Afin d'incorporer des informations temporelles dans la représentation transmise au Transformer, l'encodage positionnel classique proposé dans [10] est ajouté à chaque image de la série temporelle encodée, tel que décrit dans Figure 1. Comme dans [11], la date d'acquisition du patch, encodée en jour de l'année (*day of year DOY*), est utilisée dans l'encodage positionnel. De plus, conformément aux recommandations de [3], une constante de normalisation de 1000 est utilisée pour définir la pulsation $\omega_j = \frac{DOY}{1000^{2i/d_{model}}}$ de la fonction sinusoidale qui encode la variable indexée par $2i$ ou $2i+1$.

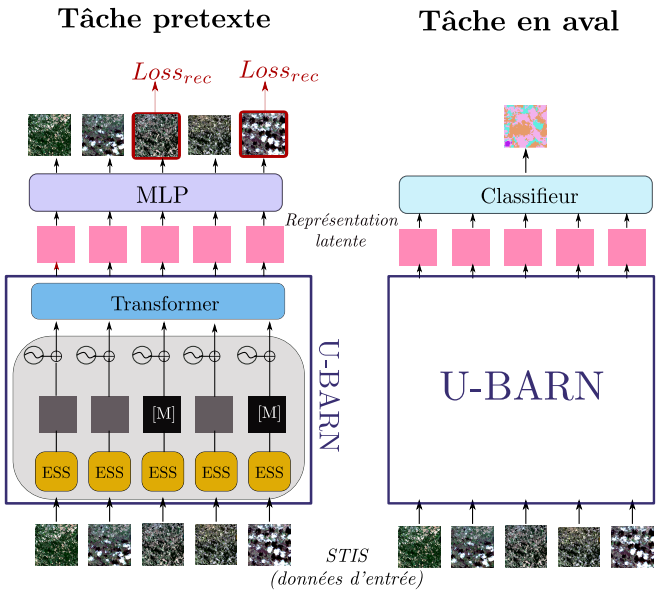

FIGURE 1 : À gauche : représentations du pré-entrainement de U-BARN avec une tache prétexte de reconstruction d'images masquées. L'indication [M] correspond à une image qui a été masquée par la stratégie de masquage décrite dans lu section 2.2. À droite la tâche en aval de segmentation sémantique. Un classifieur peu profond est entrainé à classifier des STIS à l'aide de représentations fournies par U-BARN.

2.1.2 Transformer

Le Transformer, modèle exclusivement temporel, capture les relations temporelles au sein de séries temporelles pixelliques de primitives. Au coeur de l'architecture Transformer, le module d'attention multi-tête, permet de calculer des scores de similarité pour toutes les paires de positions d'une série temporelle d'entrée X. Ces scores résultants pondèrent ensuite une représentation de X, $V = W_v X$ (valeur), et indiquent quels instants sont essentiels pour l'entrainement du réseau. Nous avons implémenté un Transformer à 3 couches et 4 têtes.

2.2 Stratégie d'apprentissage autosupervisée

Notre tâche prétexte consiste à reconstruire des données corrompues au sein d'une série temporelle. Comme détaillé dans Figure 1, le masquage intervient après l'encodage par l'ESS et avant l'encodage positionnel. Afin d'éviter le décalage de distribution entre la tâche prétexte et en aval, nous mélangeons de manière aléatoire les valeurs des pixels parmi les cartes de caractéristiques à masquer. Autrement dit, une valeur d'un pixel à corrompre se voit remplacée par une valeur provenant d'un autre pixel et/ou d'une autre date et/ou d'une autre caractéristique. Des expériences, non détaillées dans ce papier ont permis de calculer le pourcentage optimal de dates à masquer pour obtenir des représentations riches et complexes. Enfin, tel qu'illustré par Figure 1, les séries temporelles ainsi corrompues sont ensuite traitées par le Transformer et les représentations latentes résultantes transformées par un décodeur. Le décodeur est constitué d'une seule couche linéaire, qui opère exclusivement sur la dimension des primitives.

La fonction de coût utilisée lors du pré-entrainement est l'erreur quadratique moyenne (*MSE*). Comme illustré dans Figure 1, et à l'instar du BERT [2], cette dernière est calculée

exclusivement sur les images préalablement corrompues. Bien que les STIS puissent contenir des pixels nuageux, U-BARN doit apprendre à les ignorer et non à capter ce phénomène. Ainsi, dans le cas où le pixel issu d'un patch corrompu est initialement nuageux, il ne sera pas pris en compte pour le calcul de la MSE.

3 Protocole expérimental

3.1 Jeux de données

Nous utilisons deux jeux de données, sans et avec étiquettes, composés de séries temporelles d'images satellites optiques Sentinel-2 (S2) traitées au niveau L2A par *Theia*[1]. Ces jeux de données sont constitués d'images multispectrales composées de bandes avec une résolution spatiale initiale de 10 m ou 20 m, qui sont interpolées à 10m. Le masque des nuages fourni par *Theia* est également utilisé pour masquer la loss de reconstruction comme décrit précédemment. Le jeu de données de données non étiqueté, utilisé pour le pré-entrainement de U-BARN, est composé de 14 tuiles S2 acquises sur la France et la Catalogne (Espagne). Afin d'assurer une variabilité géographique entre le pré-entraînement et la tâche en aval, il n'y a aucune intersection dans l'ensemble de données non étiquetées et les données issues du jeu de données étiqueté. L'entraînement de la tâche prétexte U-BARN est réalisé en considérant 10 tuiles S2 différentes acquises entre 2018 et 2020. Le jeu de données d'entraînement est composé de 25 600 séries temporelles annuelles d'images dont la taille spatiale est de (64×64) pixels. Le jeu de validation disjoint est composé de 160 STIS issus des 4 tuiles S2 restantes acquises entre 2016 et 2019.

Des séries temporelles de janvier à novembre 2019 issues du jeu de données PASTIS [2] [3] sont utilisées pour la tâche en aval. L'ensemble de données complet contient 2433 séries temporelles de patchs et 18 classes de culture différentes, et il est divisé en 5 *fold* stratifiés pour permettre un entraînement en *k-fold*.

3.2 Tâche en aval : segmentation sémantique

Pour la tâche de segmentation sémantique, telle qu'illustrée dans Figure 1, le décodeur est remplacé par un classifieur peu profond (*shallow classifier, SC*). L'objectif est de classifier les représentations latentes apprises par U-BARN. Plus précisément, le SC est entrainé à générer des cartes de segmentation à partir des séries temporelles d'images encodées. Étant donné que la tâche en aval est une tâche de segmentation, la sortie du classifieur peu profond ne devrait avoir aucune dimension temporelle. Par ailleurs, comme l'encodeur U-BARN fournit des représentations latentes préservant la taille de la série temporelle d'entrée, le mécanisme d'attention *mean query* proposé dans [9] est utilisé dans le classifieur peu profond. L'idée est de calculer une représentation unique $Q^{mean} \in \mathbb{R}^{1*d_{model}}$ pour toutes les dates d'une série temporelle. La sortie de ce mécanisme d'attention altéré a donc une dimension temporelle de taille 1 et est suivie d'une couche linéaire (FC, *fully connected*). Enfin, la *loss* d'entropie croisée est appliquée sur les cartes de segmentation ainsi produites.

[1] https://www.theia-land.fr/

[2] https://github.com/VSainteuf/pastis-benchmark

3.3 Entrainements effectués sur la tâche en aval

Nous considérons deux configurations d'entrainement différentes du U-BARN pré-entrainé suivi du classifieur. Tout d'abord, nous définissons U-BARN^{FR}, cas où les poids de U-BARN pré-entrainé sont gelés. Dans le second cas, nommé U-BARN^{FT}, nous expérimentons le *fine-tuning* : les poids de U-BARN pré-entrainé sont utilisés comme les poids initiaux pour l'entraînement de l'architecture complète (U-BARN et SC). Les deux scénarios autosupervisés précédents sont comparés à trois configurations d'entraînement supervisées de bout en bout sur le jeu de données PASTIS. Le premier scénario, désigné par U-BARN^{e2e}, correspond à un encodeur U-BARN entraîné de bout en bout (*end-to-end*) suivi du SC. L'encodeur U-BARN^{e2e} peut être considéré comme la borne supérieure des performances de U-BARN^{FR} : les performances du modèle gelé ne sont pas censées surpasser celles de U-BARN^{e2e}. En revanche, il est attendu que U-BARN^{FT} surpasse le modèle U-BARN^{e2e}, qui est entraîné à partir de zéro. Ensuite, nous comparons les caractéristiques apprises par U-BARN avec des représentations encodées par une seule couche linéaire (*fully-connected, FC*). Dans cette seconde configuration U-BARN est remplacé par une couche FC, qui opère exclusivement sur la dimension (spectrale) des descripteurs. Cette dernière augmente la dimension spectrale de la série temporelle S2 composée de 10 bandes spectrales à d_{model}. Si l'apprentissage autosupervisé fonctionne, U-BARN^{FR} doit surpasser les performances du réseau FC-SC. Finalement, nous comparons U-BARN avec l'approche supervisé U-TAE [3].

4 Résultats et discussion

Les différents réseaux ont été entraînés pendant au minimum 100 époques, avec une taille de *batch* de 2 et un *learning rate* initial égal à 1e-3. Les performances de classification obtenues à partir des scénarios d'entraînement décrits ci-dessus (Section 3.3) sont comparées dans Table 1. U-BARN est pré-entrainé avec un taux de masquage 60%. Quatre métriques de classification différentes sont utilisées pour évaluer la qualité des résultats obtenusă: le coefficient kappa de Cohen, l'*overall accuracy* (OA), le F1 score et la moyenne de l'intersection sur l'Union (*mIoU*). Les deux dernières métriques sont moyennées par classe et non par pixel comme pour l'*OA*. Étant donné que nous procédons à un entraînement en 5-fold avec PASTIS, la moyenne et l'écart-type des métriques de classification sont indiqués.

Comme attendu, les performances du modèle U-BARN^{FR} sont entre les performances de FC-SC et de U-BARN^{e2e}. U-BARN^{FR} surpasse les performances de FC-SC sur toutes les métriques de classification. Ainsi, les représentations générées par le modèle pré-entrainé sont pertinentes. Ensuite, Table 1 montre que les performances U-BARN^{e2e} et le U-TAE sont proches. Bien que U-BARN^{e2e} ne surpasse que légèrement le U-TAE, U-BARN est lui pré-entrainable avec une tâche prétexte inspirée du BERT. Par ailleurs, contrairement à ce qui était attendu, nous remarquons également que les performances entre U-BARN^{FT} ne surpassent pas significativement U-BARN^{e2e}. Ainsi lorsque entrainé sur tout le jeu de données, il n'y a pas de réel gain de l'approche *fine-tuning* comparée à l'approche supervisée de bout en bout. Nous supposons que le

TABLE 1 : Métriques de classification (moyenne et écart-type) sur le jeu de données PASTIS avec validation en K-Folds pour différents encodeurs

	Kappa	OA	F1	mIoU
FC-SC	0.631 ± 0.015	0.770 ± 0.013	0.670 ± 0.011	0.284 ± 0.007
U-BARN^{FR}	0.689 ± 0.006	0.812 ± 0.009	0.715 ± 0.006	0.332 ± 0.002
U-BARN^{FT}	**0.831** ± 0.008	**0.901** ± 0.008	**0.841** ± 0.009	0.536 ± 0.008
U-BARN^{e2e}	**0.831** ± 0.008	**0.902** ± 0.007	**0.848** ± 0.010	0.539 ± 0.011
U-TAE	0.815 ± 0.010	0.893 ± 0.008	0.831 ± 0.008	**0.548** ± 0.007

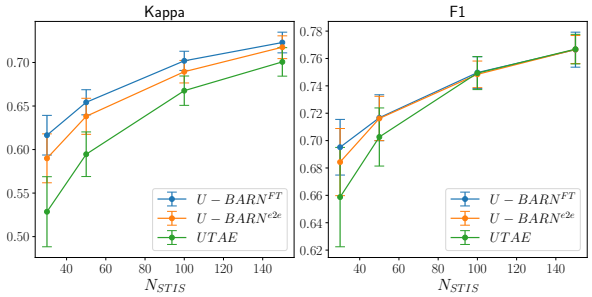

FIGURE 2 : : Performances en fonction de la taille du jeu de données étiquetté. N_{STIS} désigne le nombre de STIS issu du jeu de données PASTIS

jeu de données PASTIS dispose de suffisamment d'étiquettes pour entrainer l'architecture U-BARN. Afin d'étudier l'approche de fine-tuning dans un cas réel, avec peu de données étiquetées, une seconde expérience a été réalisé. Des jeux de données issus de PASTIS composés de N_{STIS} STIS ont été générés. Figure 2 présente les différentes métriques de classification en fonction de taille du jeu de données de la tâche de segmentation sémantique décrite par la variable N_{STIS}. Dans le cas avec N_{STIS}=30, on remarque que U-BARN^{FT} surpasse les approches supervisées. Par ailleurs, ces résultats concordent avec [12], où il est montré que le gain de performance issus du pré-entrainement diminue avec l'accroissement du nombre de données d'entrainement. Finalement, il convient de souligner que le U-BARN^{e2e} présente des performances significativement supérieures à celles du U-TAE pour N_{STIS}=30. Nous supposons que le U-TAE, en calculant l'attention temporelle à faible résolution spatiale, traite moins d'exemples de séries temporelles pixelliques que le U-BARN. Par conséquent, dans un cas où il y a peu de données étiquetées pour l'entraînement, le U-TAE pourrait avoir plus de difficultés à classifier.

5 Conclusion

Nous avons établi ici une nouvelle architecture spatio-temporelle ainsi qu'une stratégie d'entrainement autosupervisée adaptée au STIS. Les résultats expérimentaux valident la qualité des représentations générées par U-BARN sur une tache de segmentation sémantique. Les représentations issues de U-BARN pré-entrainé et gelé (U-BARN^{FR}), permettent une meilleure classification que celles issues d'une couche linéaire spécifiquement entrainé sur la tâche en aval. Une deuxième expérience, simulant un manque de données étiquetées, a démontré que le pré-entrainement de U-BARN permet de surpasser les performances de toutes les autres approches supervisées de bout en bout sur la tâche en aval. Enfin, l'architecture U-BARN, entraînée de manière supervisée, présente une légère amélioration par rapport au modèle U-TAE, qui est la référence actuelle pour la classification de STIS.

Références

[1] Paola BENEDETTI, Dino IENCO, Raffaele GAETANO, Kenji OSE, Ruggero G. PENSA et Stephane DUPUY : m^3Fusion : a deep learning architecture for multiscale multimodal multitemporal satellite data fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12) :4939–4949, 2018.

[2] Jacob DEVLIN, Ming Wei CHANG, Kenton LEE et Kristina TOUTANOVA : Bert : Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies - Proceedings of the Conference*, 1 :4171–4186, 10 2018.

[3] Vivien Sainte Fare GARNOT et Loic LANDRIEU : Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *In 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4852–4861, 10 2021.

[4] Dino IENCO, Raffaele GAETANO, Claire DUPAQUIER et Pierre MAUREL : Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(10) :1685–1689, 2017.

[5] Roberto INTERDONATO, Dino IENCO, Raffaele GAETANO et Kenji OSE : Duplo : a dual view point deep learning architecture for time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149 :91–104, 2019.

[6] Sina MOHAMMADI, Mariana BELGIU et Alfred STEIN : 3d fully convolutional neural networks with intersection over union loss for crop mapping from multi-temporal satellite images. *In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5834–5837, 2021.

[7] Lichao MOU, Lorenzo BRUZZONE et Xiao Xiang ZHU : Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2) :924–935, 2019.

[8] Charlotte PELLETIER, Geoffrey WEBB et François PETITJEAN : Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5) :523, 2019.

[9] Vivien SAINTE FARE GARNOT, Loic LANDRIEU, Sebastien GIORDANO et Nesrine CHEHATA : Satellite image time series classification with pixel-set encoders and temporal self-attention. *CVPR*, 2020.

[10] Ashish VASWANI, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N GOMEZ, Ł ukasz KAISER et Illia POLOSUKHIN : Attention is all you need. *In* I. GUYON, U. Von LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN et R. GARNETT, éditeurs : *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[11] Yuan YUAN et Lei LIN : Self-supervised pretraining of transformers for satellite image time series classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14 :474–487, 2021.

[12] Yuan YUAN, Lei LIN, Qingshan LIU, Renlong HANG et Zeng-Guang ZHOU : Sits-former : A pre-trained spatio-spectral-temporal representation model for sentinel-2 time series classification. *International Journal of Applied Earth Observation and Geoinformation*, 106 :102651, 2022.