Assimilation de données variationnelle de séries temporelles d'images Sentinel-2 avec un modèle dynamique auto-supervisé

Anthony FRION[1] Lucas DRUMETZ[1] Mauro DALLA MURA[2] Guillaume TOCHON[3] Abdeldjalil AÏSSA EL BEY[1]

[1]IMT Atlantique, UMR CNRS 6285, Lab-STICC, F-29238 Brest, France

[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
Institut Universitaire de France

[3]LRE EPITA, Le Kremlin-Bicêtre, France

Résumé – Au cours des dernières années, l'apprentissage profond a acquis une importance croissante dans de nombreux domaines scientifiques, notamment en ce qui concerne le traitement d'images, et en particulier pour le traitement des données issues de satellites. Le paradigme le plus courant en ce qui concerne l'apprentissage profond est l'apprentissage supervisé, qui requiert une grande quantité de données annotées représentant la vérité terrain pour la tâche d'intérêt. Or, obtenir des données correctement annotées pose souvent des difficultés financières ou techniques importantes. Pour cette raison, nous nous plaçons ici dans le cadre de l'apprentissage auto-supervisé. Nous proposons un modèle d'apprentissage profond inspiré de la théorie de l'opérateur de Koopman qui apprend, à partir de séries temporelles d'images multispectrales Sentinel-2, à modéliser les dynamiques de long terme de réflectance des pixels. Après son entraînement, notre modèle peut être utilisé dans divers problèmes inverses faisant intervenir la dynamique temporelle pour résoudre différentes tâches telles que l'interpolation ou le débruitage de données.

Abstract – During the last few years, deep learning has gained an increasing importance in numerous scientific fields, notably related to image processing, and in particular for the processing of satellite data. The most common paradigm used for deep learning is supervised learning, which requires a great amount of labelled data representing the groundtruth for the task of interest. Yet, obtaining correctly labelled data often poses notable financial and technical challenges. Thus, we place ourselves here in a self-supervised context. We learn a deep learning model inspired from the Koopman operator theory which learns from multispectral Sentinel-2 images time series to model the long-term dynamics of pixel reflectances. After training, our model can be used in various inverse problems involving the temporal dynamics for solving different tasks such as interpolation or data denoising.

1 Introduction

L'imagerie multispectrale en télédétection bénéfice aujourd'hui de données libres d'accès en quantité croissante, et est associée à différents problèmes de longue date tels que la détection de changement [1], la segmentation/classification sémantiques [14] ou encore le démélange spectral [13]. Ces problèmes inverses, bien qu'ayant chacun leurs spécificités, peuvent, lorsqu'ils s'intéressent à des aspects dynamiques, bénéficier de la connaissance de la dynamique de la réflectance obtenue via les observations successives des satellites. Nous proposons d'obtenir des représentations de cette dynamique via l'apprentissage auto-supervisé [9]. L'auto-supervision est un paradigme d'apprentissage consistant à apprendre, à partir de données non annotées, une représentation utile pour résoudre différentes tâches aval. Cette représentation est souvent apprise via une tâche prétexte : par exemple, pour des images, il peut s'agir de prédire les positions relatives de 2 patchs de pixels [5], de prédire les angles de rotation appliqués à des images [8] ou d'obtenir des représentations similaires lorsqu'on applique 2 transformations différentes à une même image source [3]. Ici, nous proposons d'apprendre un modèle de manière auto-supervisée pour des données séquentielles de réflectances issues des satellites Sentinel-2 afin de résoudre *in fine* des problèmes d'assimilation de données. Notre tâche prétexte est simplement la prédiction à long terme des réflectances à partir d'une condition initiale.

Le modèle neuronal utilisé ici a été introduit dans [6]. Un autre travail [7] introduit des séries temporelles d'images Sentinel-2 sur lequelles ce modèle est entraîné de façon auto-supervisée, mais montre essentiellement des applications pour la prédiction de long terme et l'interpolation. Ici, nous réutilisons les données introduites (en libre accès sur `https://github.com/anthony-frion/Sentinel2TS`) pour démontrer la capacité de notre modèle à débruiter des séries temporelles d'images satellites.

Ce travail a été financé par le projet ANR-21-CE48-000 LEMONADE.

2 Méthodes

On travaille sur une série temporelle de T images contenant chacune N pixels : $(\mathbf{x}_{i,t})_{1\leq i\leq N, 1\leq t\leq T}$ où $\mathbf{x}_{i,t} \in \mathbb{R}^L$, avec L le nombre de bandes spectrales. On traite les pixels indépendamment les uns des autres, et on omettra souvent l'indice spatial i pour désigner la série temporelle correspondant à un pixel quelconque. On dispose d'observations bruitées et/ou incomplètes $(\tilde{\mathbf{x}}_t)_{t\in S}$ avec $S \subset \{1, 2, ..., T\}$.

2.1 Entraînement du modèle

Notre modèle est basé sur la théorie de l'opérateur de Koopman [10, 2], qui agit sur les fonctions d'observation des systèmes dynamiques, en général non linéaires. Dans tous les cas, l'opérateur de Koopman a pour effet de propager les fonctions

d'observation dans le temps de façon linéaire, mais au prix d'une dimension infinie dans le cas général. Néanmoins, il est possible d'en calculer des représentations finies, par exemple via la décomposition en modes dynamiques [11]. Pour notre part, nous tâcherons d'identifier des fonctions d'observation formant un sous-espace stable par l'opérateur de Koopman, dont la restriction à cet espace peut alors s'exprimer par une matrice finie [2, 12].

Notre architecture est basée sur un auto-encodeur classique (ϕ, ψ), dont l'espace latent est supposé contenir les informations utiles sur les données traitées. Par ailleurs, notre modèle comporte également une matrice $\mathbf{K} \in \mathbb{R}^{k \times k}$ (ou, pour utiliser le vocabulaire des réseaux de neurones, une couche linéaire complètement connectée et sans biais), avec k la dimension de l'espace latent défini par l'auto-encodeur. $\mathbf{K}$ permet de faire avancer les états encodés dans le temps via l'équation :

$$\psi(\mathbf{K}^\tau \phi(\mathbf{x}_t)) = \hat{\mathbf{x}}_{t+\tau} \approx \mathbf{x}_{t+\tau}. \quad (1)$$

Ainsi, (ϕ, ψ) et $\mathbf{K}$ étant entraînés conjointement, l'auto-encodeur apprend à passer de l'espace des observations à un sous-espace stable par l'opérateur de Koopman, dont la restriction à cet espace est la matrice $\mathbf{K}$. En outre, on peut avoir certains a priori sur cette matrice. Par exemple, si la dynamique est relativement lente vis-à-vis du pas de temps fixé, alors on s'attend à ce que $\mathbf{K}$ soit proche de l'identité. De manière générale, comme expliqué dans [6], favoriser l'orthogonalité de $\mathbf{K}$ permet de conserver les normes des états latents, et en particulier d'en éviter une croissance exponentielle qui conduirait à une instabilité sur la prédiction de long terme. Par ailleurs, l'orthogonalité de $\mathbf{K}$ est particulièrement adaptée à la modélisation de systèmes ayant une forte composante périodique. Par exemple, un système dynamique de la forme $d\mathbf{x}/dt = \mathbf{A}\mathbf{x}$ avec $\mathbf{A}$ antisymétrique possède des orbites périodiques, et la solution s'écrit $\mathbf{x}(t) = \exp(\mathbf{A}t)\mathbf{x}(0)$. Ici, $\exp(\mathbf{A}t)$ est une matrice orthogonale, ce qui motive la recherche d'une matrice $\mathbf{K}$ orthogonale. Ainsi donc, la fonction de coût pour l'entraînement des composantes (ϕ, ψ) et $\mathbf{K}$ est composée des termes suivants :

- Le terme de prediction $L_{pred,\tau}$ mesure directement l'erreur quadratique entre l'estimation via (1) de l'état à τ pas de temps et la vérité terrain :

$$L_{pred,\tau}(\Theta) = \sum_{\substack{1 \le i \le N \\ 1 \le t \le T-\tau-1}} ||\mathbf{x}_{i,t+\tau} - \psi(\mathbf{K}^\tau \phi(\mathbf{x}_{i,t}))||^2. \quad (2)$$

- Le terme de linéarité $L_{lin,\tau}$ mesure l'erreur quadratique entre l'état latent prédit à τ pas de temps et l'encodage de la vérité terrain correspondante :

$$L_{lin,\tau}(\Theta) = \sum_{\substack{1 \le i \le N \\ 1 \le t \le T-\tau-1}} ||\phi(\mathbf{x}_{i,t+\tau}) - \mathbf{K}^\tau \phi(\mathbf{x}_{i,t})||^2. \quad (3)$$

- Enfin, le terme d'orthogonalité permet d'assurer que les valeurs propres de la matrice $\mathbf{K}$ sont proches du cercle unité :

$$L_{orth}(\mathbf{K}) = ||\mathbf{K}\mathbf{K}^T - \mathbf{I}||_F^2. \quad (4)$$

En pratique, les fonctions de coût utilisées pour entraîner notre modèle sont des combinaisons linéaires de ces termes, avec différents horizons de prédictions τ. L'entraînement du modèle auto-supervisé est réalisé sur des séries temporelles complètes supposées sans bruit. Nous renvoyons à [7] pour les explications plus détaillées.

2.2 Assimilation de données classique

Une fois un modèle pré-entraîné à prédire les dynamiques de long terme comme expliqué dans la section 2.1, on peut l'utiliser comme un a priori dynamique dans des problèmes inverses. Cet a priori a l'avantage d'être entièrement différentiable et donc de pouvoir être intégré facilement dans un problème d'optimisation résolu par descente de gradient via la différenciation automatique. Ainsi, on peut en premier lieu utiliser une formulation générique d'assimilation de données variationnelle pour apprendre une trajectoire satisfaisant :

$$\arg\min_{\hat{\mathbf{x}} \in \mathbb{R}^{L \times T}} \sum_{t \in S} ||\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t||^2 + \alpha \sum_{2 \le t \le T} ||\hat{\mathbf{x}}_t - \mathcal{M}(\hat{\mathbf{x}}_{t-1})||^2 \quad (5)$$

où la première somme correspond à la fidélité aux données disponibles et la deuxième somme correspond à la fidélité au modèle dynamique $\mathcal{M}$. Le paramètre α a pour rôle de déterminer l'importance relative des deux termes. On peut ici utiliser les composantes (ϕ, ψ) et $\mathbf{K}$ d'un modèle pré-entraîné en fixant, conformément à (1), $\mathcal{M}(\hat{\mathbf{x}}_t) = \psi(\mathbf{K}\phi(\hat{\mathbf{x}}_t))$. Il est à noter que, même si nous avons ici choisi d'utiliser uniquement un paramètre α constant, on pourrait envisager de pondérer plus finement les observations, par exemple en fonction du temps. Cela peut accroître les performances de l'assimilation au prix d'une plus grande complexité dans le choix des paramètres de la fonction de coût.

2.3 Assimilation de données contrainte

Une autre façon d'assimiler une série temporelle est d'exploiter la capacité de notre modèle à modéliser les dynamiques de long terme en optimisant non pas une trajectoire complète mais simplement une condition initiale latente, en recherchant :

$$\mathbf{z}_1^* = \arg\min_{z_1 \in \mathbb{R}^k} \sum_{t \in S} ||\tilde{\mathbf{x}}_t - \psi(\mathbf{K}^{t-1}\mathbf{z}_1)||^2, \quad (6)$$

avec k la dimension latente du modèle. La fidélité au modèle dynamique est dans ce cas assurée non pas par un terme dans la fonction de coût mais par une contrainte dure. En effet, les vecteurs de réflectance obtenus sont alors tous issus de la propagation de la condition initiale $\mathbf{z}_1^*$ par le modèle via $\hat{\mathbf{x}}_t = \psi(\mathbf{K}^{t-1}\mathbf{z}_1^*)$. Cette méthode est facile à utiliser en pratique puisqu'elle ne nécessite pas de choisir un hyperparamètre α contrairement à la méthode de la section (2.2). Elle est également très robuste dans des cas difficiles où les données sont lacunaires et très bruitées, puisque toute sortie correspond à une série temporelle plausible selon le modèle.

3 Expériences

Nous présentons ici nos expériences de débruitage sur des séries temporelles d'images Sentinel-2. Il s'agit d'images multispectrales à 10 bandes (domaines visible et proche infrarouge) des forêts de Fontainebleau et d'Orléans, qui comportent des dynamiques saisonnières induisant une pseudo-périodicité d'un an. La période d'acquisition est de 5 jours, mais beaucoup d'images sont inutilisables à cause des nuages. Nous avons donc interpolé les données disponibles pour compléter les séries temporelles, mais également conservé des versions irrégulières contenant seulement les images sans nuages. Ces données sont décrites plus en détail dans [7].

Notre protocole expérimental est le suivant : pour un écart type donné σ et un pixel d'indice spatial i de la forêt de Fontainebleau, on ajoute à la série temporelle $(\mathbf{x}_{i,t})_t$ des réflectances (à 10 dimensions) un bruit blanc gaussien d'écart-type σ. Ainsi, on obtient la série temporelle bruitée $(\tilde{\mathbf{x}}_{i,t})_t$ telle que

$$\tilde{\mathbf{x}}_{i,t} = \mathbf{x}_{i,t} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}). \tag{7}$$

Étant données les composantes pré-entraînées $\phi, \psi, \mathbf{K}$ d'un modèle tel que décrit dans la section 2, on peut alors estimer une version débruitée de la dynamique $\tilde{\mathbf{x}}$, en résolvant un problème d'optimisation tel que décrit par l'équation (5) ou par l'équation (6). On mesure ensuite l'erreur quadratique moyenne entre la série temporelle assimilée et la série temporelle originelle $(\mathbf{x}_{i,t})_t$. La moyenne est calculée sur toutes les bandes spectrales et tous les pas de temps.

Pour chaque valeur de bruit σ, nous répétons cette expérience avec 100 pixels différents (sélectionnés aléatoirement) et calculons ainsi une erreur quadratique moyenne globale. À titre de référence, nous comparons cette erreur avec celle réalisée par une interpolation de Cressman avec des poids gaussiens. L'idée est que, pour un temps de référence t, on associe à chaque temps $t' \in S$ où une observation est disponible un poids $w_{t,t'} = \exp \frac{(t-t')^2}{2R^2}$, où R est le rayon d'interpolation, qui détermine la vitesse de décroissance des poids quand on s'éloigne du temps de référence. Une estimation des réflectances débruitées est alors obtenue par

$$\mathbf{x}_t^{Cressman} = \frac{1}{\sum_{t' \in S} w_{t,t'}} \sum_{t' \in S} w_{t,t'} \tilde{\mathbf{x}}_{t'}. \tag{8}$$

Nous réalisons nos expériences dans 2 cadres différents : la figure 1 rend compte de l'assimilation classique via l'équation (5) avec des amplitudes de bruit faibles, et la figure 2 rend compte de l'assimilation contrainte via l'équation (6) avec des amplitudes de bruit fortes. À titre de référence, l'écart-type des données de réflectance issues de Fontainebleau (tous pixels, temps et bandes spectrales confondus) est d'environ 0,20.

L'assimilation contrainte est moins pertinente pour traiter des données faiblement bruitées, dans la mesure où le modèle ne peut pas recréer une série temporelle (même issue des données d'entraînement) de manière exacte, et où l'erreur de reconstruction par le modèle peut alors être plus importante que l'erreur associée au bruit lui-même. L'erreur quadratique moyenne de reconstruction de données de Fontainebleau complètes non bruitées par le modèle est de l'ordre de 2×10^{-4}, ce qui implique qu'une optimisation contrainte par le modèle peut difficilement débruiter des données dont le bruit induit une erreur inférieure ou égale à ce seuil. En revanche, l'optimisation contrainte est très efficace pour traiter des problèmes difficiles où les données sont très bruitées et/ou partielles.

Concernant le débruitage de séries temporelles faiblement bruitées, nous avons comparé l'assimilation de données classique via (5) avec notre modèle au débruitage basé sur l'interpolation de Cressman. Chacune de ces méthodes comprend un paramètre important : le terme α de poids du modèle pour notre méthode et le rayon R des poids gaussiens pour la méthode de Cressman. Pour chaque valeur de bruit, nous avons recherché la valeur optimale de α parmi les valeurs entières et celle de R par tranches de 0,1. On peut constater sur la figure 1 que, si les 2 méthodes ont des performances similaires à faible niveau de bruit, notre méthode d'assimilation devient significativement plus efficace à mesure que le bruit augmente.

Pour les niveaux plus élevés de bruit, le choix du paramètre R étant beaucoup moins critique, nous avons fixé $R = 3$. Notre méthode d'assimilation contrainte ne requiert pour sa part aucun paramètre. Nous avons travaillé avec des données régulières de la forêt de Fontainebleau mais aussi avec des données irrégulières de la forêt d'Orléans, non présentes dans les données d'entraînement de notre modèle, pour lesquelles le cas non bruité correspondrait à l'expérience d'interpolation de [7]. Comme le montre la figure 2, notre méthode permet une bien meilleure reconstruction que la méthode de Cressman dans les 2 cas de figure, d'autant plus que le bruit augmente. Il est cependant à noter qu'il s'agit de valeurs de bruits très importantes, peu courantes pour un bruit blanc. La figure 3 fournit un exemple concret de débruitage pour un pixel aléatoire avec les 2 méthodes. En outre, nous représentons sur la figure 4 le débruitage d'une image de Fontainebleau par les 2 méthodes. Il est à noter qu'il s'agit des bandes du spectre visible et que celles-ci, étant d'amplitude faible, sont très sensibles au bruit. Les données des autres bandes spectrales, de plus forte intensité, contribuent grandement au débruitage par notre méthode. Les 2 débruitages proposés sont réalisés pixel par pixel, dans la dimension temporelle et non spatiale, mais pourraient, comme montré dans [7], utiliser un module de correction spatiale a posteriori sur ces images. On pourrait également intégrer un terme de cohérence spatiale dans l'assimilation de données, ce que nous n'avons pas fait ici pour éviter une comparaison trop injuste avec l'interpolation de Cressman.

4 Conclusion

Nous avons repris le modèle d'apprentissage auto-supervisé présenté dans [7] et avons présenté 2 méthodes d'assimilation de données variationnelle utilisant un tel modèle comme a priori dynamique. L'une des méthodes comporte les 2 termes classiques de fidélité aux données et de fidélité au modèle, tandis que l'autre utilise le modèle comme une contrainte. Les avantages et inconvénients respectifs de ces modélisations ont été discutés et leur efficacité à été démontrée pour le débruitage de séries temporelles complètes ou irrégulières.

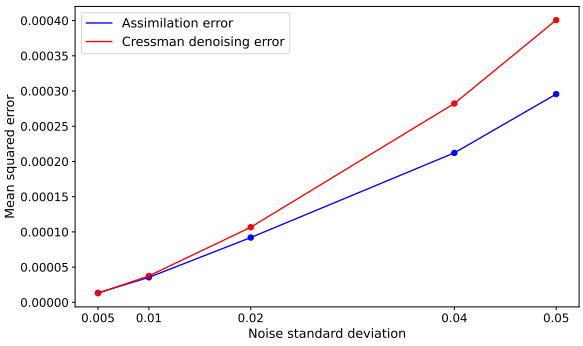

FIGURE 1 : Comparaison des performances de débruitage avec différents niveaux de bruit, pour la méthode de la section 2.2 et pour l'interpolation de Cressman. Les paramètres optimaux des 2 méthodes ont été recalculés pour chaque niveau de bruit.

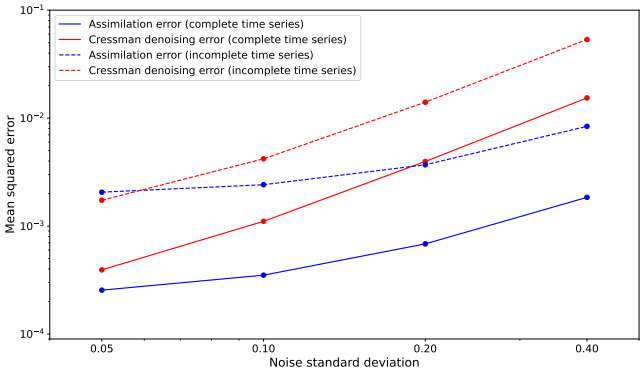

FIGURE 2 : Comparaison des performances de débruitage avec différents niveaux de bruit, pour la méthode de la section 2.3 et pour l'interpolation de Cressman. Les données complètes sont celles de la forêt de Fontainebleau et les données incomplètes proviennent de la forêt d'Orléans.

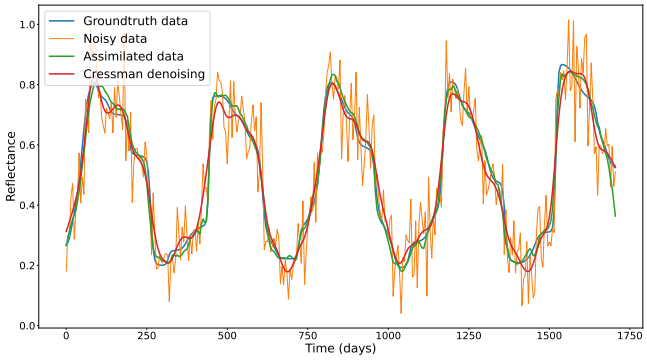

FIGURE 3 : Exemple de débruitage avec notre méthode et une méthode de référence basée sur l'interpolation de Cressman. La bande spectrale examinée est la bande B7, avec un bruit blanc d'écart-type 0,1.

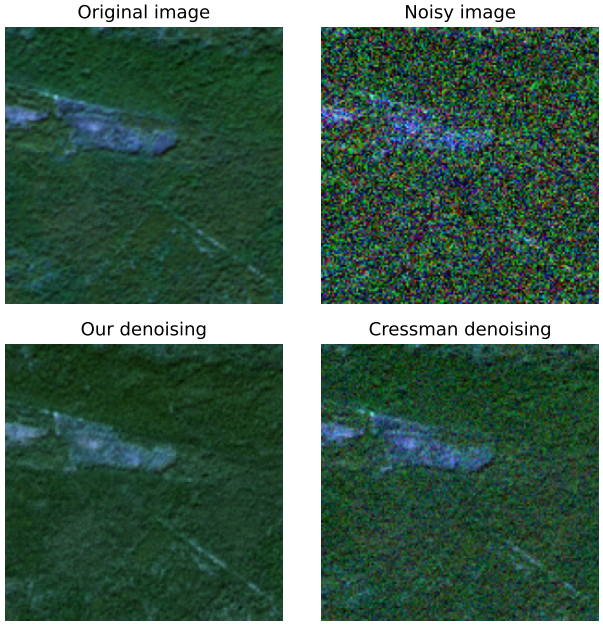

FIGURE 4 : Représentation des bandes RGB pour une image originelle, sa version bruitée (avec un bruit gaussien d'écart-type 0,05) et les débruitages proposés par 2 méthodes.

Références

[1] Anju ASOKAN et JJESI ANITHA : Change detection techniques for remote sensing applications : A survey. *Earth Science Informatics*, 12:143–160, 2019.

[2] Steven L BRUNTON, Bingni W BRUNTON, Joshua L PROCTOR et J Nathan KUTZ : Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. *PloS one*, 11(2):e0150171, 2016.

[3] Ting CHEN, Simon KORNBLITH, Mohammad NOROUZI et Geoffrey HINTON : A simple framework for contrastive learning of visual representations. *In ICML*, pages 1597–1607. PMLR, 2020.

[4] George P CRESSMAN : An operational objective analysis system. *Monthly Weather Review*, 87(10):367–374, 1959.

[5] Carl DOERSCH, Abhinav GUPTA et Alexei A EFROS : Unsupervised visual representation learning by context prediction. *In IEEE ICCV*, pages 1422–1430, 2015.

[6] Anthony FRION, Lucas DRUMETZ, Mauro DALLA MURA, Guillaume TOCHON et Abdeldjalil Aïssa EL BEY : Leveraging neural koopman operators to learn continuous representations of dynamical systems from scarce data. *IEEE ICASSP*, 2023.

[7] Anthony FRION, Lucas DRUMETZ, Guillaume TOCHON, Mauro DALLA MURA et Abdeldjalil AISSA EL BEY : Learning Sentinel-2 reflectance dynamics for data-driven assimilation and forecasting. preprint, février 2023.

[8] Spyros GIDARIS, Praveer SINGH et Nikos KOMODAKIS : Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv :1803.07728*, 2018.

[9] Jie GUI, Tuo CHEN, Qiong CAO, Zhenan SUN, Hao LUO et Dacheng TAO : A survey of self-supervised learning from multiple perspectives : Algorithms, theory, applications and future trends. *arXiv preprint arXiv :2301.05712*, 2023.

[10] Bernard O KOOPMAN : Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

[11] Peter J SCHMID : Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

[12] Naoya TAKEISHI, Yoshinobu KAWAHARA et Takehisa YAIRI : Learning koopman invariant subspaces for dynamic mode decomposition. *NeurIPS*, 30, 2017.

[13] Qunming WANG, Xinyu DING, Xiaohua TONG et Peter M ATKINSON : Spatio-temporal spectral unmixing of time-series images. *Remote Sensing of Environment*, 259:112407, 2021.

[14] Giulio WEIKMANN, Claudia PARIS et Lorenzo BRUZZONE : TimeSen2Crop : A million labeled samples dataset of sentinel 2 image time series for crop-type classification. *IEEE J-STARS*, 14:4699–4708, 2021.