# Réseau antagoniste génératif pour la fusion spatio-temporelle d'images satellitaires multi-spectrales

Amine HADIR[1,2] Ahed ALBOODY[1,3] Matthieu PUIGT[1] Gilles ROUSSEL[1]
Vincent VANTREPOTTE[4] Cédric JAMET[4] Trung-Kien TRAN[4]

[1]Univ. Littoral Côte d'Opale, LISIC – UR 4491, F-62219 Longuenesse, France

[2]L@bISEN, Yncrea Ouest, F-29228 Brest, France

[3]LINEACT, CESI, F-06200 Nice, France

[4]Univ. Littoral Côte d'Opale, CNRS, LOG – UMR 8187, F-62930 Wimereux, France

**Résumé –** Dans cet article, nous étudions la fusion spatio-temporelle d'une série temporelle d'images multi-spectrales avec une série temporelle d'images hyper-spectrales. Nous proposons pour cela une nouvelle approche fondée sur un réseau antagoniste génératif (GAN). Notre contribution principale réside dans le fait que le GAN prend en entrée des images satellitaires plutôt que du bruit. Nous montrons que notre approche fournit de meilleures performances en termes de PSNR et SAM que les approches de la littérature sur des images Sentinel-2 et Sentinel-3.

**Abstract –** In this article, we study the spatio-temporal fusion of a time series of multi-spectral images with a time series of hyper-spectral images. We propose for this a new approach based on a generative adversarial network (GAN). Our main contribution lies in the fact that the GAN takes as input satellite images rather than noise. We show that our approach provides a better PSNR et SAM performance than SotA approaches on Sentinel-2 and Sentinel-3 images.

## 1 Introduction

En observation marine côtière de la couleur de l'océan où les phénomènes qui s'y produisent sont complexes, il est nécessaire que les données disponibles offrent à la fois une bonne résolution spatiale, spectrale et temporelle. Cependant, aucun satellite dédié à la couleur de l'océan – par exemple Sentinel-2 (S2) ou Sentinel-3 (S3) – ne répond à de telles contraintes. En effet, l'augmentation du nombre de bandes spectrales entraîne généralement une diminution de la résolution spatiale. Par exemple, S2 offre une résolution spatiale élevée, allant de 10 à 60 m, mais seulement 13 bandes spectrales alors que S3 fournit 21 bandes spectrales pour une résolution fixe de 300 m [1]. De plus, la période d'échantillonnage de ces instrument peut sensiblement varier. Ainsi S2 et S3 collectent respectivement une image d'une zone donnée tous les 5 et 1,4 jours [2].

En conséquence, plusieurs méthodes de fusion ont été proposées pour combiner les différents types d'images satellitaires. Alors que de nombreuses approches de *sharpening* – ou fusion spatio-spectrale – ont été proposées [8], la prise en compte de séries temporelles est une problématique plus récemment étudiée [6]. Ces approches cherchent à compléter une série temporelle de données multispectrales – acquises à faible fréquence d'échantillonnage – à partir d'une série d'images hyperspectrales acquise à plus haute fréquence. Les approches de fusion spatio-temporelle qui ont été proposées pour résoudre ce problème sont basées sur des approches pondérées [13], du démélange [15], des approches d'apprentissage profond [2, 7] ou des stratégies hybrides [9]. Enfin et plus récemment, des méthodes d'apprentissage basées sur les réseaux antagonistes génératifs (GAN pour *Generative Adversarial Networks* en anglais) ont été proposées [10, 11].

Dans cet article, nous proposons une nouvelle approche basée sur les GAN pour la complétion de série temporelle de données S2 à partir de données S3. L'originalité de notre approche réside dans le fait que notre réseau prédit de nouvelles données S2 non-observées à partir de cinq images S2 et S3 acquises (et non à partir d'un vecteur de bruit comme cela est classiquement réalisé). La suite de l'article est organisée de la manière suivante. Nous introduisons le problème et la méthode proposée dans la section 2. Nous étudions ses performances dans la section 3 et concluons dans la section 4.

## 2 Problèmatique et méthode proposée

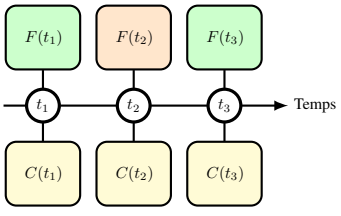


FIGURE 1 : Périodes d'échantillonnage considérées.

Dans cette section, nous introduisons le problème considéré et la méthode développée. Plus précisément, nous considérons deux séries temporelles d'images à haute et faible résolution spatiales. Nous les notons par la suite images $F$ (pour *Fine spatial resolution*) et $C$ (pour *Coarse spatial resolution*). Nous supposons par ailleurs que (i) la période d'échantillonnage des images $C$ est plus petite que celle des images $F$ et que (ii) certaines images $F$ et $C$ sont acquises les mêmes jours. De plus, nous supposons que pour une image $F$ acquise à l'instant $t_2$ et notée $F(t_2)$, il existe respectivement deux paires $F(t_i)$ et

$C(t_i)$ pour $i = 1$ ou 3 et une image $C(t_2)$. Ainsi, comme dans [2,7] et comme on peut le voir sur la figure 1, nous cherchons à estimer $F(t_2)$ à partir de $F(t_1)$, $F(t_3)$, $C(t_1)$, $C(t_2)$ et $C(t_3)$. Nous proposons pour cela une approche fondée sur les GAN.

Les GAN sont des modèles génératifs proposés en 2014 par Goodfellow *et al.* [4] et ont été extrêmement appliqués depuis. Les GAN fonctionnent en mettant en compétition deux réseaux neuronaux, un réseau générateur et un réseau discriminateur. Le générateur doit créer de nouvelles données synthétiques qui doivent ressembler à de vraies données. Le discriminateur lui cherche à discriminer les données générées de vraies données. Lorsque le discriminateur réussit sa tâche, le générateur est ré-entraîné et vice-versa, lorsque le discriminateur ne peut plus distinguer données réelles et simulées, il est ré-entraîné. L'entraînement du GAN est terminé lorsque ni le générateur ni le discriminateur ne peuvent être améliorés.

Dans cet article, nous proposons d'utiliser nous-aussi une architecture à deux réseaux. Cependant, notre approche – appelée S2S3-STFGAN (pour *Sentinel-2 Sentinel-3 Spatial-Temporal Fusion using GANs*) – fait appel à des données existantes. La structure générale du générateur – fourni dans la figure 2 – consiste en des blocs d'extraction de caractéristiques, des blocs résiduels et un auto-encodeur suivi par des blocs de compression. Cette architecture est inspirée de [5] mais le simplifie en n'utilisant qu'un seul générateur et un seul discriminateur (au lieu de deux de chaque dans [5]).

Nous présentons maintenant la structure du réseau générateur. Deux types de blocs résiduels sont intégrés pour les images $C$ et $F$, c.-à-d., un bloc résiduel simple et un auto-encodeur. En outre, nous utilisons deux blocs différents pour l'extraction des caractéristiques, le premier avec 32 filtres dans chaque couche de convolution et le second avec 90 filtres dans chaque couche de convolution. Ensuite, par des opérations de différence et de concaténation, nous construisons deux nouvelles entrées, c.-à-d., une entrée $F$ et une entrée $C$. La première, notée $X_F$, consiste en la concaténation de $F(t_1), F(t_3), C(t_3) - C(t_1)$ et $F(t_3) - F(t_1)$. Les deux différences sont utilisées pour détecter les changements temporels de l'instant $t_1$ à l'instant $t_3$. Ces différencespermettent une première approximation de $F(t_2)$. Le seconde entrée, notée $X_C$, consiste en la concaténation de $C(t_1)$, $C(t_2)$, $C(t_2) - C(t_1)$, $C(t_3) - C(t_2)$, et $C(t_3)$. Comme pour $X_F$, nous utilisons certaines différences entre les images afin d'extraire tous les changements spatio-temporels entre les instants $t_1$ et $t_2$ d'une part, et entre $t_2$ et $t_3$ d'autre part. Ensuite, nous alimentons un bloc résiduel fin par l'entrée fine. Ce réseau est défini par $H(X_F) + X_F$, où $+$ est la fonction d'addition entre $X_F$ et le résultat de deux opérations de convolution, chaque couche de convolution étant suivie d'une fonction d'activation de LeakyRelu appliquée à $X_F$. Nous effectuons un traitement similaire pour le bloc résiduel grossier, sauf que le nombre de blocs est différent. Enfin, nous concaténons les sorties des blocs résiduels fins et grossiers en tant qu'image d'entrée pour l'auto-encodeur qui contient deux blocs. Le réseau encodeur se compose de huit couches de convolution avec la fonction d'activation LeakyReLu pour extraire des informations plus détaillées de la sortie des blocs résiduels fins et grossiers. D'un point de vue mathématique, l'image d'entrée du codeur est la concaténation des blocs résiduels fin et grossier.

La sortie de l'encodeur est ensuite utilisée comme entrée du decodeur. Afin de reconstruire une image $F$ avec une résolution spatiale élevée, le decodeur se compose de sept couches de décodage. Chacune d'entre elles se compose d'une couche UpSampling2D – avec une interpolation bilinéaire – et de trois couches de convolution avec une couche de normalisation par lots. Ensuite, nous appliquons un certain nombre de suppressions (*dropout layer*) et nous appliquons enfin la fonction d'activation LeakyRelu pour obtenir le résultat.

En outre, nous passons notre image décodée dans un bloc de compression pour obtenir une image super-résolue. Pour réaliser une bonne estimation de l'image cible $F(t_2)$, nous concaténons l'entrée fine avec l'image super-résolue pour alimenter les blocs résiduels suivis d'un bloc de compression pour obtenir l'image estimée à partir du générateur.

Une fois que nous avons généré $F(t_2)$, nous utilisons un réseau discriminant comme modèle de classification entre les images générées et les images réelles, comme le montre la Fig. 3. En entrée, nous utilisons les images réelles et générées à l'instant $t_2$. Le réseau discriminant se compose de cinq couches cachées, chacune contenant une couche convolutive suivie d'une couche de normalisation par lots et d'une fonction d'activation LeakyReLu, avec une couche Dropout pour éliminer le surajustement obtenu lors de l'apprentissage.

# 3 Expériences

Dans cette section, nous montrons la performance de notre méthode S2S3-STFGAN proposée pour prédire une image inconnue $F(t_2)$ à partir de $F(t_1)$, $F(t_3)$, $C(t_1)$, $C(t_2)$, et $C(t_3)$. Pour cela, nous considérons une base de données de séries temporelles S2 et S3. Ces données ont subi 3 étapes avant d'être utilisées, c.-à-d. *(i)* sélection de la zone d'intérêt, *(ii)* ré-alignement des images S2 et S3 et *(iii)* gestion des données manquantes dûes à la correction atmosphérique [14]. Nous utilisons la résolution spatiale de 60 m du jeu de données S2 et la résolution spatiale de 300 m du jeu de données S3. Nous comparons notre méthode proposée à plusieurs approches[1], à savoir STARFM [3], DCSTFN [12], DMNET [6] et CSTF [2]. Mis à part CSTF et S2S3-STFGAN, DCSTFN, DMnet et STARFM n'utilisent que trois images observées aux instants $t_1$ et $t_2$ (c.-à-d. $F(t_1)$, $C(t_1)$ et $C(t_2)$) pour compléter la série temporelle S2. Au cours du processus d'apprentissage, chaque image est découpée en patches de taille $256 \times 256$. Nous disposons au total de 360 patches pour réaliser l'entraînement. Pour le test, nous utilisons 3 images – soit 216 patches – pour lesquels nous connaissons la vérité de terrain. Enfin, les données manquantes dûes à la correction atmosphérique sont remplacées par une constante trouvée de manière heuristique.

Les paramètres de la méthode proposée sont fournis dans les tableaux 1 et 2. En outre, nous utilisons l'optimiseur ADAM avec une valeur initiale de $10^{-3}$, la fonction de perte utilisée pour l'entraînement des réseaux du générateur et du discriminateur est l'erreur quadratique moyenne (MSE) et l'entropie croisée binaire, respectivement. Le nombre de passages de données du GAN est de 1000 et la taille du lot est égale à 8. Notre modèle est entraîné sur une machine locale de type PC contenant une carte graphique NVIDIA GTX 3060 avec 12 Go de VRAM. L'entraînement dure 12 heures.

Afin d'évaluer les performances des méthodes testées, nous

[1]Nous n'avons pas pu obtenir les codes des méthodes génératives [10, 11] pour pouvoir comparer leurs performances.

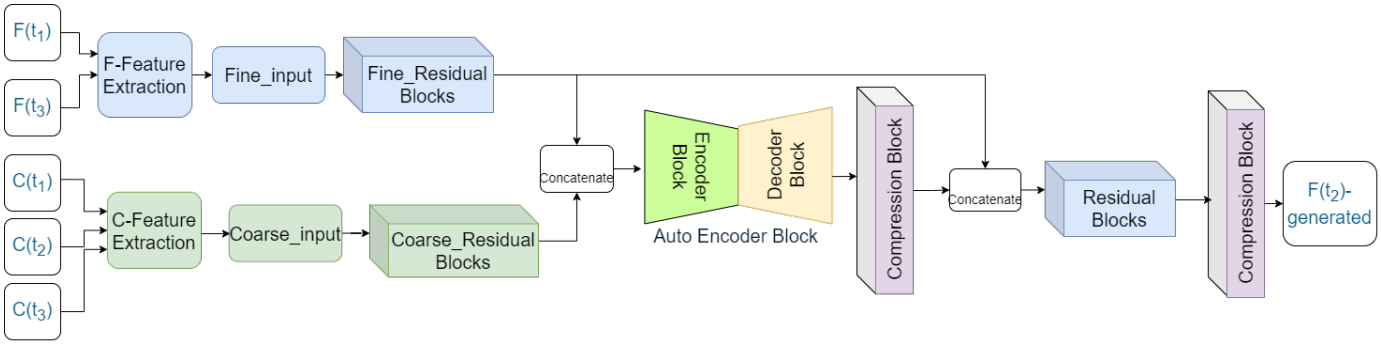


FIGURE 2 : Architecture du générateur S2S3-STFGAN.

| Bloc | Paramètres d'extraction | Noyaux | Filtres |
|---|---|---|---|
| Extraction des features des images $F$ | 2 Conv2D + Leaky ReLu | $3 \times 3$ | 90 |
| Extraction des features des images $C$ | 3 Conv2D + Leaky ReLu | $3 \times 3$ | 32 |
| Bloc résiduel fin | 2 Conv2D + Leaky ReLu | $3 \times 3$ | 64 |
| Bloc résiduel grossier | 2 Conv2D + Leaky ReLu | $3 \times 3$ | 64 |
| Bloc encodeur | 8 Conv2D + Leaky ReLu | $3 \times 3$ | 64, 64, 128, 128, 128, 256, 256, 512 |
| Bloc décodeur | UpSampling2D + 3 Conv2D + 3 Leaky Relu | $4 \times 4$ | 256, 256, 128, 128, 128, 64, 64 |
| Bloc résiduel simple | 2 Conv2D + Leaky ReLu | $1 \times 1$ | 64 |
| Bloc de compression | UpSampling layer + Conv2D + Conv2D | $3 \times 3$ | 256, 1 |

TABLE 1 : Paramètres des blocs utilisés dans le générateur de la méthode proposée.

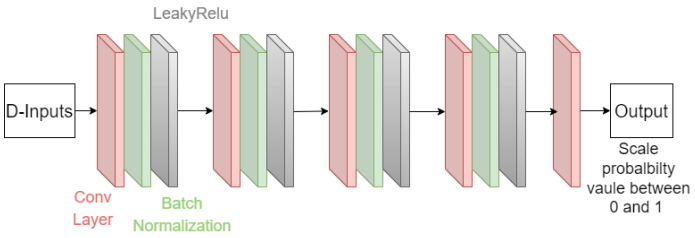


FIGURE 3 : Architecture du discriminateur S2S3-STFGAN.

| Couches cachées | Paramètres d'extraction | Noyaux | Filtres |
|---|---|---|---|
| Couche 1 | Conv2D + Leaky ReLu + BN | $4 \times 4$ | 32 |
| Couche 2 | Conv2D + Leaky ReLu + BN | $4 \times 4$ | 64 |
| Couche 3 | Conv2D + Leaky ReLu + BN | $4 \times 4$ | 128 |
| Couche 4 | Conv2D + Leaky ReLu + BN | $4 \times 4$ | 512 |
| Couche 5 | Conv2D + Leaky ReLu + BN | $4 \times 4$ | 1 |

TABLE 2 : Paramètres des blocs utilisés dans le discriminateur de la méthode proposée.

utilisons les mêmes critères objectifs que dans [2], c.-à-d. le PSNR (*Peak Signal-to-Noise Ratio*), le SAM (*Spectral Angle Mapper*), le SSIM (*Structural SIMilarity*), le SCC (*Spatial Correlation Coefficient*) et l'UIQI (*Universal Image Quality Index*). Mis à part pour le SAM, plus les valeurs de ces indicateurs sont élevées et meilleures sont les performances atteintes. Cependant, la mesure du SAM nous semble la plus importante pour l'application considérée puisqu'elle indiquera l'erreur d'estimation des spectres.

Le tableau 3 présente les performances obtenues par toutes les méthodes testées pour chacune des quatre bandes spectrales S2 à une résolution spatiale de 60 m. La méthode S2S3-STFGAN proposée surpasse toutes les approches en termes de PSNR et de SAM, tandis qu'elle fournit des valeurs SSIM,

| | PSNR | SAM | SSIM | SCC | UIQI |
|---|---|---|---|---|---|
| Méthode proposée S2S3-STFGAN | | | | | |
| Bande 1 | **19.1** | **0.4639** | 0.863 | 0.022 | 0.0905 |
| Bande 2 | **19.1** | **0.4639** | 0.863 | 0.021 | 0.0905 |
| Bande 3 | **19.1** | **0.4639** | 0.862 | 0.023 | 0.0905 |
| Bande 4 | **19.1** | **0.4639** | 0.866 | 0.026 | 0.0905 |
| **Moyenne** | **19.1** | **0.4639** | 0.863 | 0.023 | 0.0905 |
| DCSTFN | | | | | |
| Bande 1 | 17.2 | 0.5700 | 0.863 | 0.012 | 0.0615 |
| Bande 2 | 17.2 | 0.5703 | 0.858 | 0.012 | 0.0615 |
| Bande 3 | 17.2 | 0.5706 | 0.859 | 0.012 | 0.0617 |
| Bande 4 | 17.2 | 0.5704 | 0.880 | 0.013 | 0.0613 |
| Moyenne | 17.2 | 0.5703 | 0.865 | 0.012 | 0.0615 |
| DMNet | | | | | |
| Bande 1 | 17.6 | 0.549 | 0.489 | 0.020 | 0.0661 |
| Bande 2 | 17.6 | 0.550 | 0.400 | 0.021 | 0.0661 |
| Bande 3 | 17.6 | 0.551 | 0.335 | 0.021 | 0.0661 |
| Bande 4 | 17.6 | 0.550 | 0.407 | 0.022 | 0.0661 |
| **Moyenne** | 17.6 | 0.550 | 0.400 | 0.021 | 0.0661 |
| STARFM | | | | | |
| Bande 1 | 17.7 | 0.9165 | 0.053 | 0.032 | 0.0692 |
| Bande 2 | 17.7 | 0.9164 | 0.053 | 0.032 | 0.0692 |
| Bande 3 | 17.7 | 0.9165 | 0.053 | 0.032 | 0.0692 |
| Bande 4 | 17.7 | 0.9165 | 0.053 | 0.032 | 0.0692 |
| **Moyenne** | 17.7 | 0.9164 | 0.053 | 0.032 | 0.0692 |
| CSTF | | | | | |
| Bande 1 | 18.2 | 0.5385 | **0.923** | **0.052** | **0.1196** |
| Bande 2 | 18.2 | 0.5388 | **0.923** | **0.053** | **0.1196** |
| Bande 3 | 18.2 | 0.5392 | **0.924** | **0.052** | **0.1195** |
| Bande 4 | 18.2 | 0.5383 | **0.926** | **0.054** | **0.1197** |
| **Moyenne** | 18.2 | 0.5387 | **0.924** | **0.053** | **0.1196** |

TABLE 3 : Performances des méthodes testées.

SCC et UIQI proches de celles atteintes par CSTF, qui surpasse toutes les méthodes testées pour ces critères de performance.

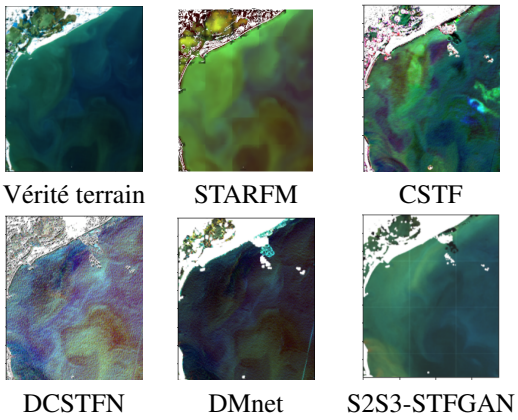


FIGURE 4 : Performance visuelle des méthodes testées.

Enfin, la Fig. 4 montre l'image $F(t_2)$ restaurée obtenue avec toutes les méthodes testées, en couleurs pseudo-RGB. La sortie de la méthode S2S3-STFGAN proposée est clairement la plus proche de l'image de vérité terrain, tandis que des artefacts sont clairement visibles avec DCSTFN, CSTF, et DMnet. Enfin, les couleurs obtenues avec la méthode STARFM sont très éloignées des couleurs réelles. Cela montre la pertinence de la méthode proposée.

# 4 Conclusion

Dans cet article, nous avons proposé une nouvelle méthode basée sur les GAN pour compléter des séries temporelles d'images spatialement fines à l'aide d'une série temporelle d'images spatialement grossières. Nos principales contributions résident dans le fait que nous utilisons (i) 5 images comme entrée de notre générateur, (ii) des blocs résiduels fins et grossiers, et (iii) un auto-encodeur dans notre générateur. La validité expérimentale de la méthode proposée a été démontrée sur des données réelles pour lesquelles nous connaissions l'image S2 à prédire. Nous avons également utilisé un ensemble de données réelles pour entraîner et tester la méthode proposée. Dans les travaux futurs, nous souhaitons étendre notre approche à la fusion spatio-spectro-temporelle qui permette d'améliorer la résolution spectrale des images générées. La prise en compte de modèles physiques pour guider la fusion serait aussi un axe intéressant.

**Remerciements** Ce travail est financé par le projet CNES TOSCA « OSYNICO ». Les expériences ont été partiellement réalisées sur la plateforme de calcul CALCULCO de l'ULCO.

# Références

[1] A. ALBOODY, M. PUIGT, G. ROUSSEL, V. VANTREPOTTE, C. JAMET et T.-K. TRAN : Experimental comparison of multi-sharpening methods applied to Sentinel-2 MSI and Sentinel-3 OLCI images. *In Proc. IEEE WHISPERS'21*, 2021.

[2] C. T. CISSÉ, A. ALBOODY, M. PUIGT, G. ROUSSEL, V. VANTREPOTTE, C. JAMET et T. K. TRAN : A new deep learning method for multispectral image time series completion using hyperspectral data. *In Proc. IEEE ICASSP'22*, pages 1546–1550, 2022.

[3] F. GAO, J. MASEK, M. SCHWALLER et F. HALL : On the blending of the landsat and modis surface reflectance : Predicting daily landsat surface reflectance. *IEEE Trans. Geosci. Remote Sens.*, 44(8):2207–2218, 2006.

[4] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE et Y. BENGIO : Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[5] M. JIANG, H. SHEN, J. LI et L.-P. ZHANG : An integrated framework for the heterogeneous spatio-spectral-temporal fusion of remote sensing images. *ArXiv*, abs/2109.00400, 2021.

[6] J. LI, Y. LI, L. HE, J. CHEN et A. PLAZA : Spatio-temporal fusion for remote sensing data : An overview and new benchmark. *Science China Information Sciences*, 63(4):140301, 2020.

[7] X. LIU, D. HONG, J. CHANUSSOT, B. ZHAO et P. GHAMISI : Modality translation in remote sensing time series. *IEEE Trans. Geosci. Remote Sens.*, 60:1–14, 2021.

[8] L. LONCAN, L. B. DE ALMEIDA, J. M. BIOUCAS-DIAS, X. BRIOTTET, J. CHANUSSOT, N. DOBIGEON, S. FABRE, W. LIAO, G. A. LICCIARDI, M. SIMOES *et al.* : Hyperspectral pansharpening : A review. *IEEE Geosci. Remote Sens. Mag.*, 3(3):27–46, 2015.

[9] J. MA, W. ZHANG, A. MARINONI, L. GAO et B. ZHANG : An improved spatial and temporal reflectance unmixing model to synthesize time series of landsat-like images. *Remote Sensing*, 10(9):1388, 2018.

[10] C. SHANG, X. LI, Z. YIN, X. LI, L. WANG, Y. ZHANG, Y. DU et F. LING : Spatiotemporal reflectance fusion using a generative adversarial network. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15, 2021.

[11] Y. SONG, H. ZHANG, H. HUANG et L. ZHANG : Remote sensing image spatiotemporal fusion via a generative adversarial network with one prior image pair. *IEEE Trans. Geosci. Remote Sens.*, 60:1–17, 2022.

[12] Z. TAN, P. YUE, L. DI et J. TANG : Deriving high spatiotemporal remote sensing images using deep convolutional network. *Remote Sensing*, 10(7):1066, 2018.

[13] Q. WANG et P. M. ATKINSON : Spatio-temporal fusion for daily sentinel-2 images. *Remote Sensing of Environment*, 204:31–42, 2018.

[14] M. ZHANG, C. HU et B. BARNES : Performance of POLYMER atmospheric correction of ocean color imagery in the presence of absorbing aerosols. *IEEE Trans. Geosci. Remote Sens.*, PP:1–9, 04 2019.

[15] B. ZHUKOV, D. OERTEL, F. LANZL et G. REINHACKEL : Unmixing-based multisensor multiresolution image fusion. *IEEE Trans. Geosci. Remote Sens.*, 37(3):1212–1226, 1999.