Mise en correspondance d'images satellites dans un espace latent variationnel contraint par information mutuelle

Cassandra MUSSARD[1] Alexandre CONSTANTIN[1] Emmanuelle SARRAZIN[1]

[1]Centre National d'Études Spatiales, 18 av. Edouard Belin, 31401 Toulouse, France

Résumé – La mise en correspondance entre deux images nécessite une mesure de similarité adaptée. Une mesure adaptée donne un poids plus fort aux images similaires. Nous proposons une nouvelle approche basée sur un auto-encodeur variationnel pour mesurer la distance entre deux lois à densité, avec une contrainte d'information mutuelle entre le vecteur latent et la loi en entrée. Nous montrons de bons résultats pour la correspondance entre images multi-modales, appuyé par l'ajout de la contrainte.

Abstract – Matching two images requires a suitable similarity measure. A suitable measure gives a higher weight to similar images. We propose a new approach based on a variational auto-encoder to measure the distance between two density laws, with a mutual information constraint between the latent vector and the law in input. We show good results for multimodal image matching, supported by the addition of the constraint.

1 Introduction

La mise en correspondance d'images est une étape clé d'une chaîne de traitement d'images afin de fournir un ensemble de données cohérentes. En imagerie satellitaire, elle permet également d'améliorer des performances de localisation absolue au sol. Pour ce faire, la similarité est mesurée entre deux extraits d'une paire d'images, centrés autour d'un pixel d'intérêt. Définir une mesure de similarité adaptée est critique pour la précision du résultat. Une bonne mesure de similarité retourne un extremum marqué lorsque deux extraits sont semblables. Elle permet alors de choisir de manière non ambiguë le déplacement entre les deux extraits.

Deux familles de mesures existent [3], des mesures de dépendance linéaire, comme l'inter-corrélation de deux variables aléatoires nommée ZNCC (Zero-mean Normalized Cross-Correlation) et une dépendance statistique non-linéaire, comme l'information mutuelle [8] (IM), à l'état de l'art. Dans les deux cas, on considère l'ensemble des valeurs des pixels (autour du pixel d'intérêt) comme des réalisations de variables aléatoires, alors inconnues et estimées de manière empirique. Les méthodes basées sur l'apprentissage contrastif [9, 7] permettent de réduire le temps de calcul en inférant directement les représentations latentes là où les métriques statistiques requièrent d'estimer des distributions. Cependant, elles nécessitent la mise en place d'une base d'apprentissage supervisée, basée sur des données de référence produites par ces mêmes algorithmes. Cette étude se concentre sur des modèles à apprentissage auto-supervisé.

L'objectif est de définir une représentation à densité connue et contrainte, afin de mesurer une distance ou divergence entre ces lois. Nous étudions ici une représentation de lois à partir d'images en utilisant des modèles variationnels [3, Section 6].

Notations. Par la suite, on note $x \sim p(x)$ la loi sur les réalisations en l'entrée. $\mathbf{z} \in \mathbb{R}^p$ le vecteur latent dont la loi est, conditionnellement à x, $p(\mathbf{z}|x) \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma = \text{diag}(\boldsymbol{\sigma}^2))$, et $\boldsymbol{\sigma} \in \mathbb{R}^p$ ($\text{diag}(\boldsymbol{\sigma}) = \{\Sigma \in \mathbb{R}^{p\times p} \mid \Sigma_{i,j} = \boldsymbol{\sigma}_i \text{ si } i = j, 0 \text{ sinon}\}$).

2 VAE contraint et mesure de similarité entre lois à densité

2.1 Concept de la mesure

Définissons x et y deux lois uni-variées à valeur réelle sur lesquelles une distance est calculée. La plupart des mesures de similarité, par exemple définies dans [3], sont faites sur un ensemble de réalisations $x_i \sim p(x)$ et $y_j \sim p(y)$ avec $(i, j) \in \{1, \dots, N\}^2$, et N l'ensemble des valeurs. En pratique, il s'agit d'extraits d'images à comparer, appelés patchs.

Dans cette optique, nous souhaitons encoder un patch centré autour d'un pixel d'intérêt, dit source, pour la référence et encoder plusieurs patchs dans l'image cible, centrés autour des pixels dans le voisinage du pixel source (appelée fenêtre de recherche). Il s'agit ensuite de calculer une distance sur les lois encodées. Ce principe est illustré Figure 1.

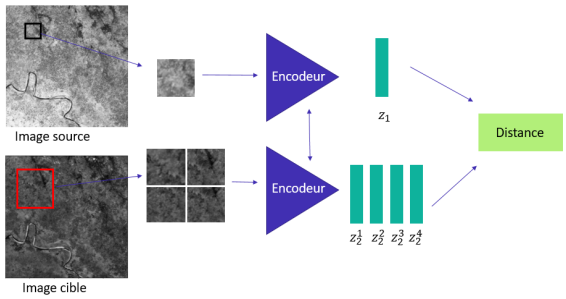

FIGURE 1 : Principe d'encodage et mesure de similarité entre lois à densité entre deux images.

2.2 Modèle VAE Contraint (VAE-EIM)

Pour cela, il est nécessaire de définir une fonction d'encodage de l'image, par auto-encodeur variationnel (VAE). Un VAE [4] permet d'encoder une entrée x par le réseau paramétré par $\boldsymbol{\phi}$ à partir de ses réalisations de manière probabiliste et continue. Puis, la représentation latente $\mathbf{z}$ est décodée pour reconstruire

l'entrée, le résultat est noté x'. L'auto-encodeur apprend de manière auto-supervisé à minimiser l'erreur de reconstruction et est régularisé par une loi normale centrée réduite (a priori) sur $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}))$. La sortie de l'encodeur contient deux vecteurs, la moyenne $\boldsymbol{\mu}$ et la diagonale $\boldsymbol{\sigma}$.

Lors du calcul de similarité, seul l'encodeur du VAE est utilisé pour réduire la dimension des patchs et avoir une représentation probabiliste continue de ceux-ci dans l'espace latent. Afin de maximiser le contenu informationnel partagé entre x et $\mathbf{z}$ [5], le réseau intègre une contrainte supplémentaire d'information mutuelle. Par ailleurs, d'après [5], cette contrainte réduit le risque d'effondrement de la loi a posteriori [2]. La fonction de coût s'écrit :

$$\mathcal{L}(x, x') = -\beta * \underbrace{\mathrm{IM}_{\boldsymbol{\theta}}(x, \mathbf{z})}_{\textbf{contrainte IM}} + \underbrace{\|x - x'\|^2}_{\textbf{reconstruction}} + \alpha * \underbrace{\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma})), \mathcal{N}(0, \mathbb{I}_n))}_{\textbf{régularisation}}, \quad (1)$$

où α pondère la loi a priori et β contrôle la contrainte sur l'espace latent. Cette contrainte est estimée par le réseau de neurones MINE [1] (noté $\mathrm{IM}_{\boldsymbol{\theta}}(x, \mathbf{z})$) et est uniquement utilisée lors de la phase d'entraînement du VAE. Une illustration globale est donnée à la Figure 2.

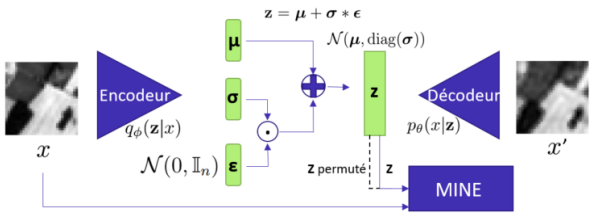

FIGURE 2 : Principe de fonctionnement d'un VAE avec la contrainte d'IM

L'estimateur MINE utilise la borne inférieure de l'information mutuelle [1], et se sert d'une fonction paramétrée par le vecteur $\boldsymbol{\theta}$. Pour deux variables aléatoires continues ou discrètes $\mathcal{I}$ et $\mathcal{J}$, on a :

$$\mathrm{IM}(\mathcal{I}, \mathcal{J}) \geq \mathrm{IM}_{\boldsymbol{\theta}}(\mathcal{I}, \mathcal{J}).$$

Le réseau $\mathrm{IM}_{\boldsymbol{\theta}}$ prend en entrée deux vecteurs dans des espaces différents, par batch de taille K, et retourne l'IM pour chaque élément du batch. D'une part, le couple de vecteurs $(\mathbf{x}_k, \mathbf{z}_k), k \in \{1, \ldots, K\}$ est donné en entrée et modélise la loi jointe, tel que $\mathbf{x}_k$ est le vecteur des valeurs des pixels du patch k et $\mathbf{z}_k \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_k)$. D'autre part, des vecteurs $\mathbf{z}_k$ sont permutés dans le batch pour les lois marginales (afin de supprimer la dépendance à x_k), noté $\hat{\mathbf{z}}$ (ou $\mathbf{z}$ permuté) sur la Figure 2.

[6] propose un estimateur non biaisé qui est utilisé dans VAE-EIM, où $\mathrm{NN}_{\boldsymbol{\theta}}$ est un réseau de neurones qui prend en entrée le couple $(\mathbf{x}_k, \mathbf{z}_k)_{k=1,\ldots,K}$ et qui prédit un scalaire en sortie :

$$-\mathrm{IM}_{\boldsymbol{\theta}}(x, \mathbf{z}) = -\frac{1}{K}\sum_{k=1}^{K} \mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{z}_k) + \frac{1}{K}\sum_{k=1}^{K} \exp(\mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{x}_k, \hat{\mathbf{z}}_k))). \quad (2)$$

La fonction de coût définie par (1) optimise les paramètres de l'auto-encodeur (comprenant $\boldsymbol{\phi}$) et ceux de MINE ($\boldsymbol{\theta}$) qui minimisent l'équation (2).

2.3 Architectures VAE et MINE

L'architecture du VAE est constituée de couches convolutives et de perceptrons multi-couches, voir Tableau 1.

TABLE 1 : Architecture VAE-EIM.

Couche	Taille	Nb params
Conv2d	[128,8,15,15]	136
LeakyReLU	[128,8,15,15]	0
Conv2d	[128,16,7,7]	2 064
LeakyReLU	[128,16,7,7]	0
FC ($\boldsymbol{\mu}$)	[128,256]	200 960
FC ($\boldsymbol{\sigma}$)	[128,256]	200 960
FC ($\mathbf{z}$)	[128,784]	201 488
Conv2d-Transpose	[128,8,15,15]	2 056
LeakyReLU	[128,8,15,15]	0
Conv2d-Transpose	[128,1,30,30]	129
LeakyReLU	[128,1,30,30]	0

Cette architecture comporte au total 607.793 paramètres entraînables dont seulement 404.120 sont utilisés avec l'encodeur à l'inférence. La taille de l'espace latent est de 256 et la taille des patchs est 31×31, le pourcentage de compression de l'entrée dans l'espace latent est de 73.34%[1]. En pratique, la dimension varie entre x et $\mathbf{z}$, on utilise l'apprentissage de [5] sur des lois uni-variées où chaque élément de l'entrée ou de l'espace latent suit la même loi. Ce qui, en première approximation, est valable avec le prior de régularisation (1).

2.4 Distances dans l'espace latent à l'inférence

Soient $\mathbf{z}_1 = \mathbf{z}|x \sim \mathcal{N}(\mu_1, \Sigma_1)$ et $\mathbf{z}_2 = \mathbf{z}|y \sim \mathcal{N}(\mu_2, \Sigma_2)$, deux variables latentes obtenues après l'encodage respectif de deux patchs tirés selon $p(x)$ et $p(y)$. La similarité entre ces deux lois normales multivariées $\mathbf{z}_1$ et $\mathbf{z}_2$ est calculée à l'aide de la distance de Wasserstein de l'équation (4) ou de la divergence de Jeffrey définie par $D_J(\mathbf{z}_1, \mathbf{z}_2) = D_{\mathrm{KL}}(\mathbf{z}_1, \mathbf{z}_2) + D_{\mathrm{KL}}(\mathbf{z}_2, \mathbf{z}_1)$, une symétrisation de la divergence de KL définie par (3). Dans notre cas, la distance étant calculée sur des gaussiennes, les équations se simplifient de la manière suivante où l'opérateur $|.|$ correspond au déterminant de la matrice et Tr représente la trace d'une matrice :

$$D_{\mathrm{KL}}(\mathbf{z}_1, \mathbf{z}_2) = \frac{1}{2}\Big((\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \Sigma_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) - \ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) - n\Big), \quad (3)$$

$$W_2(\mathbf{z}_1, \mathbf{z}_2)^2 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2*(\Sigma_2^{\frac{1}{2}}\Sigma_1\Sigma_2^{\frac{1}{2}})^{\frac{1}{2}}). \quad (4)$$

D'autres métriques existent entre densités, cette étude porte sur celles-ci.

[1] Concernant le réseau MINE, l'architecture est reprise de [5], on la retrouve Annexe E de `https://arxiv.org/abs/1912.13361`.

3 Application à des images satellites

Nous appliquons notre méthode au cas du recalage d'images satellitaires multi-modales : une image issue de la bande rouge, notée R et une image issue de la bande infrarouge thermique, notée TIR.

3.1 Dataset

Un dataset a été construit à partir de 20 paires d'images du satellite Landsat-8 de la NASA [2], provenant de diverses régions du globe et présentant des paysages variés (urbains, déserts, montagnes).

Une phase de pré-traitement des données est nécessaire car les images satellites sont codées sur 16 bits ce qui entraîne des variations extrêmes au niveau des intensités des pixels. La première étape consiste à calculer les quantiles à 1% et 99%. Toutes les valeurs en dehors de cette plage sont ramenées aux limites des quantiles correspondants, éliminant ainsi les valeurs aberrantes. Les images sont ensuite standardisées puis divisées en patchs de taille 31×31 pixels, générant 131.537 échantillons. Le dataset est équilibré avec la moitié des échantillons provenant de la bande R et l'autre moitié de la bande TIR. Pour finir, une image est conservée pour la phase de test et les images restantes sont découpées en 80% pour l'entraînement et les 20% restants, non présentés à l'entraînement, sont utilisés pour la validation.

3.2 Phase d'entraînement

Lors de la phase d'entraînement du réseau VAE-EIM, l'objectif est de reconstruire les images issues des deux modalités. Elles sont encodées par le même réseau, donc dans un même espace latent, afin que les variables latentes soient directement comparables.

Le réseau est entraîné pendant 1100 epochs sur une carte graphique A100 de 80G pendant 7 heures en utilisant l'optimiseur Adam et un scheduler Cosine Annealing pour modifier le pas d'apprentissage de manière périodique.

Afin d'évaluer la qualité de reconstruction du VAE-EIM entraîné avec $\beta = \alpha = 1$, le Peak Signal-to-Noise Ratio (PSNR) est calculé sur la paire d'images test, chacune de taille 31×31, issues des bandes R et TIR. Le PSNR de 20.1 dB pour la bande R et 24.1 dB pour la bande TIR indique une bonne qualité de reconstruction des deux modalités. Une visualisation des reconstructions des deux images est donnée à la Figure 3.

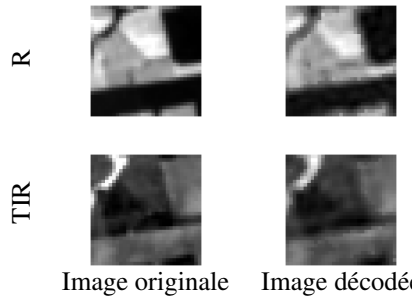

FIGURE 3 : Images de la bande R et TIR à gauche et leur reconstruction à droite

[2] https://www.usgs.gov/landsat-missions/landsat-8

3.3 Résultats

Le VAE-EIM est comparé à deux autres mesures de similarité : ZNCC et l'IM (estimée par la méthode des histogrammes avec le critère Scott) entre $p(x)$ et $p(y)$.

La carte de disparité ligne/colonne regroupe les informations de décalage entre les pixels des deux images, la disparité étant l'extremum des mesures de similarité entre un patch de x et les patchs de la fenêtre de recherche de y associés. Soit $C \in \mathcal{M}_N(\mathbb{R})$ la carte de disparité estimée et $V \in \mathcal{M}_N(\mathbb{R})$ la vérité terrain, i.e. le vrai décalage entre pixels. Dans notre cas, le décalage est forcé à 1 pixel en colonne, car un décalage nul signifie le centre de la fenêtre de recherche. Les décalages autorisés sont de l'ordre de 0.25 pixels, par ré-échantillonage des images en entrée (splines).

Tout d'abord, nous comparons les résultats des différentes méthodes en calculant les erreurs moyennes ($\mathrm{EM} = \frac{1}{N^2}\sum_{i,j=1}^{N} |\mathrm{C}_{i,j} - \mathrm{V}_{i,j}|$) et en écart-type $\left(\mathrm{EET} = \sqrt{\frac{1}{N^2}\sum_{i,j=1}^{N}(|\mathrm{C}_{i,j} - \mathrm{V}_{i,j}| - \mathrm{EM})^2}\right)$ entre C et V. Nous utilisons la valeur absolue afin de quantifier à quel point l'algorithme prédit correctement le décalage, sans se focaliser sur un éventuel biais directionnel. Les résultats sont présentés dans le Tableau 2.

Ensuite, afin de démontrer l'apport de la contrainte de MINE pour le recalage d'images satellitaires, les EM et EET ont été comparées pour plusieurs VAE-EIM entraînés avec β variant de 0, 10, 25, 50 à 100 en fixant α à 1. Lorsque $\beta = 100$ le réseau est contraint à maximiser la dépendance statistique entre l'espace latent et l'entrée de manière significative, quitte à dégrader les performances de reconstruction.

Enfin, les profils de similarité pour plusieurs mesures de similarité ont été comparés afin de déterminer si une d'entre elles permet de mieux discriminer le pixel homologue. La paire d'images test utilisée pour comparer les mesures de similarité est décalée d'un pixel en colonne et n'est pas décalée en ligne. Le calcul de similarité a été effectué pour toutes les paires de patchs R/TIR issues de cette paire d'images.

TABLE 2 : Erreur moyenne (EM) et en écart type (EET) sur les cartes de disparité colonnes pour chaque mesure de similarité, en vert la mesure de similarité avec l'erreur la plus faible et en gris l'apport du VAE-EIM.

Mesure de similarité	EM	EET
IM-Scott	0.18	0.55
ZNCC	2.08	1.09
VAE + Jeffrey	1.87	1.14
VAE + Wasserstein	1.92	1.12
VAE-EIM $\beta = 50$ + Jeffrey	1.52	1.10
VAE-EIM $\beta = 50$ + Wasserstein	2.03	1.11

L'IM est, comme attendu, la mesure de similarité qui commet le moins d'erreurs en moyenne et en écart type. En revanche, dans ce contexte, l'utilisation d'un VAE est justifié car les EM et EET sont plus faibles que pour ZNCC, utilisée de manière opérationnelle dans le cadre de recalage d'images satellites. L'ajout de la contrainte MINE couplée à la divergence de Jeffrey permet de diminuer les EM et EET par rapport à un VAE classique. En revanche, des résultats inverses sont observés avec la distance de Wasserstein.

La Figure 5 montre l'évolution de l'EM et l'EET selon

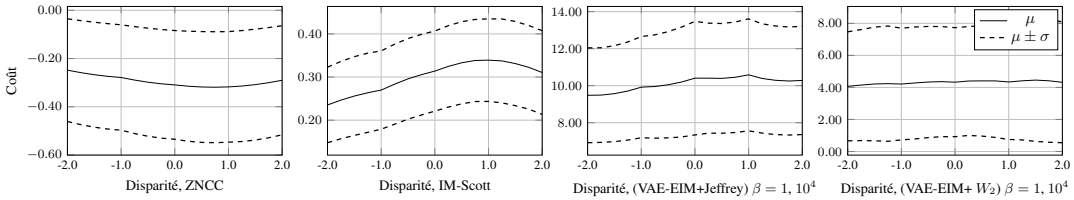

FIGURE 4 : Profils de coût 2D moyennés (μ) sur 100 pixels pour les mesures de similarité : ZNCC, IM, VAE-IM $\beta = \alpha = 1$ + divergence de Jeffrey et distance de Wasserstein avec en pointillé plus ou moins l'écart type autour de la valeur moyenne ($\mu \pm \sigma$)

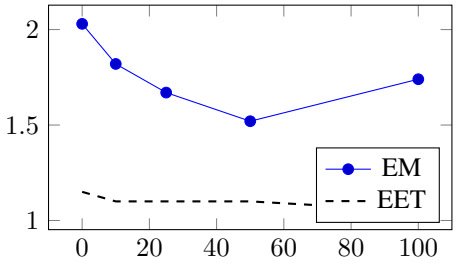

FIGURE 5 : Évolution de l'EM (en trait plein) et de l'EET (en pointillé) en fonction du paramètre β.

l'hyper-paramètre β avec l'architecture VAE-EIM et la divergence de Jeffrey. Ce graphique montre que jusqu'à un certain stade, plus le réseau est contraint à maximiser la dépendance statistique entre l'espace latent $\mathbf{z}$ et la loi sur les données en entrée x, plus l'EM diminue, montrant alors que cette contrainte a un réel apport pour le recalage d'images multi-modales. Cette conclusion n'est plus vérifiée lorsque la contrainte devient trop grande notamment lorsque $\beta = 100$ et que l'on dégrade trop la reconstruction.

Afin d'obtenir une estimation générale du déplacement entre la paire d'images test issue des bandes R et TIR, la moyenne des profils de coûts associée à 100 pixels de l'image source à été calculé en cherchant des décalages sous-pixellique de l'ordre de 0.25 pixels. Les profils de coût ont été comparés en utilisant les mesures de similarité suivantes : ZNCC, IM, VAE-EIM avec les distances de Jeffrey et Wasserstein et $\beta = \alpha = 1$. Les résultats sont donnés à la Figure 4.

Toutes les mesures de similarité, à l'exception de ZNCC ainsi que le VAE-EIM avec distance de Wasserstein, permettent de choisir le bon pixel correspondant. La mesure permettant de discriminer au mieux le pixel correspondant est le VAE-EIM avec divergence de Jeffrey. En effet, on passe d'un coût de 101691 à la disparité 0.75 à un coût de 101809 à la disparité 1.

4 Conclusion

La mesure VAE-EIM proposée démontre la possibilité de recaler deux images au travers de la distance entre deux densités continues, sans atteindre l'état de l'art comme la mesure de similarité par information mutuelle. Ces premiers résultats ont été quantifiés sur un même encodeur pour tout type d'image afin d'avoir un même espace latent. Des tests ont été effectués avec deux encodeurs différents appris sur deux modalités différentes, avec des résultats prometteurs avec la distance de Wasserstein.

Une limitation, qui n'est pas présente dans l'IM, est que la distance entre deux densités est utilisée. Hors l'IM utilise la divergence de la densité jointe au produit des densités marginales. Des premières perspectives portent sur la définition d'un modèle représentant la densité jointe pour une paire d'image.

De plus, l'utilisation de MINE dans le VAE apporte une assurance sur l'information compressée par l'encodeur. Cette approche ouvre la voie à une nouvelle perspective où des densités scalaires seraient utilisées.

Références

[1] M. I. BELGHAZI, A. BARATIN, S. RAJESHWAR, S. OZAIR, Y. BENGIO, A. COURVILLE et D. HJELM : Mutual Information Neural Estimation. *In Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 531–540. PMLR, Jul 2018.

[2] S. BOWMAN, L. VILNIS, O. VINYALS, A. M. DAI, R. JOZEFOWICZ et S. BENGIO : Generating sentences from a continuous space. *In arXiv preprint arXiv :1511.06349.*, 2015.

[3] S. CHAMBON et A. CROUZIL : Similarity measures for image matching despite occlusions in stereo vision. *Pattern Recognition*, 44(9):2063–2075, 2011.

[4] D. KINGMA et M. WELLING : An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[5] A. LOTFI REZAABAD et S. VISHWANATH : Learning representations by maximizing mutual information in variational autoencoders. *In Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 770–774, 2019.

[6] X. NGUYEN, M. WAINWRIGHT et M. JORDAN : Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE transactions on information theory / Professional Technical Group on Information Theory*, 56:5847—-5861, 2010.

[7] D. QUAN, S. WANG, Y. GU, R. LEI, B. YANG, S. WEI, B. HOU et L. JIAO : Deep feature correlation learning for multi-modal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[8] P. VIOLA et W. WELLS : Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[9] J. ZBONTAR et Y. LE CUN : Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 65:1–32, 2016.