HETEROGENEOUS CHANGE DETECTION WITH SELF-SUPERVISED DEEP CANONICALLY CORRELATED AUTOENCODERS

F. Figari Tomenotti[†], *L.T. Luppino*[‡], *M.A. Hansen*[‡], *G. Moser*[†], *S.N. Anfinsen*[‡]

[†]University of Genoa, DITEN Department, Genoa, Italy
[‡]UiT The Arctic University of Norway, Department of Physics and Technology, Tromsø, Norway

ABSTRACT

This paper proposes a new method for bitemporal change detection in heterogeneous remote sensing images. A modified canonical correlation analysis is used to align the code layers of two deep convolutional autoencoders, one for each image domain. It weights the input with a new affinity-based prior, which measures changes in pixel relations across the image domains and is used to reduce the influence of data points prone to change. By this procedure of self-supervision, we adapt the intrinsically supervised architecture to the unsupervised case, noting that the censoring of change pixels is key to efficiently learning the required data transformations. The result is an unsupervised algorithm which allows change detection in either of the image domains, or a combination of those, since efficient domain translation is obtained by coupling cross-domain encoders and decoders. We demonstrate state-of-the-art performance on real test datasets.

1. INTRODUCTION

Heterogeneous change detection (HCD) is an emerging topic in earth observation. It answers the increasing availability of remote sensing data by offering methods that can combine radically different images and still extract reliable information about changes on the surface. The images could be acquired by multimodal sensors, such as optical instruments and synthetic aperture radar (SAR), or recorded with different sensor parameters or under distinct environmental conditions, cases that would otherwise not be comparable unless possibly through meticulous preprocessing and co-calibration. We collectively label change detection under these circumstances as heterogeneous. In the bitemporal setting, HCD is particularly useful to obtain situational awareness after sudden change events such as natural disasters, when we may want to use the first image of opportunity to map change, instead of waiting for an acquisition that permits a comparison of homogeneous images. For monitoring long-term trends, the joint analysis of heterogeneous sources allows us to extend the time frame of the analysis or to increase the temporal resolution.

This work was partially funded by the Research Council of Norway under research grant no. 251327.

Regardless of the motivation, HCD relies on the fundamental assumption that the changed areas have a distinct signature for all the sensors involved, even though the physical origin of this signal may be different. Moreover, since an absolute reference is lacking when we contrast heterogeneous data, the problem is inherently ill-posed and the labelling of pixels or segments as changed and unchanged is generally ambiguous. We have to assume some additional prior information in order to discern the change class. This could be that the change concerns small regions or a minority of the pixels in an image, or knowledge about characteristic signatures of one of the classes involved in the transition. The mentioned minority assumption is common in generic methods, and we adopt it also here, while signature assumptions can be advantageous to customise an algorithm for a thematic application.

In recent years, focus in HCD research has turned from the supervised to the unsupervised case. This makes the methods more relevant for practical cases, since ground truth is sparse and costly to collect. Another trend is that deep learning prevails more and more, as in other areas of computer vision and image analysis. Most HCD approaches adopt transformations between the input domains, or from these to a common latent domain, to bring data to a space where they can be efficiently compared. Convolutional neural network (CNN) architectures, such as autoencoders and generative adversarial networks, are flexible and powerful tools that can accomplish these image translation tasks, as reviewed in [1, 2].

This paper is inspired by a recently proposed architecture for supervised HCD [3]. It uses canonical correlation analysis (CCA) [4] to align the code spaces of two autoencoders, each processing data from one input domain. CCA is a linear method, but the encoders are deep CNNs whose nonlinear transformations can improve the alignment significantly. The decoders are also deep CNNs, and ensure that the codes used as input to the CCA remain meaningful, since they must retain the information required for successful reconstruction. Without the decoders, the scheme is equivalent to deep CCA [5], which has also been used for supervised HCD [6], but did not provide as good performance as the deep canonically correlated autoencoder (DCCAE) [7] we adopt.

CCA is a supervised method [4] and does not immediately lend itself to unsupervised HCD. By applying CCA di-

recly to paired image patches without labels, we risk contaminating the learning of efficient image transformations by the data from changed areas, where the sought correspondence between the domains does not hold. Our solution is a novel method for extracting prior information about the probability of "changedness" directly from the unlabelled data. This pixel-level prior is proposed and used in [2] to develop other approaches to unsupervised HCD. It is based on spectral clustering concepts and formalized in terms of a pixel-wise distance measure between domain-specific local affinity matrices. We exploit that affinities are normalised and can be compared across image domains, to formulate a cross-domain pixel distance that is given a probabilistic interpretation and used to weight input data to the CCA. We study the performance of the prior-weighted DCCAEs by performing change detection in both the code and the input spaces.

2. METHOD

Let us assume to have acquired co-located images in two different domains $\mathcal{X}$ and $\mathcal{Y}$ at times t_1 and t_2, respectively, that are also co-registered and have the same spatial resolution. The images, $\mathcal{I}_\mathcal{X} \in \mathbb{R}^{H\times W\times C_\mathcal{X}}$ and $\mathcal{I}_\mathcal{Y} \in \mathbb{R}^{H\times W\times C_\mathcal{Y}}$, have height H, width W, and respectively $C_\mathcal{X}$ and $C_\mathcal{Y}$ channels. Lastly assume that the changed pixels are a minority.

2.1. Affinity-based Change Prior

Our prior information is an affinity-based cross-domain pixel distance proposed in [2], which is interpreted as a probability of change of that pixel. To obtain this distance measure we first compute the domain-specific affinity matrices $\mathbf{A}^\mathcal{X}$ and $\mathbf{A}^\mathcal{Y}$, whose elements $A^\mathcal{X}_{ij}$ and $A^\mathcal{Y}_{ij}$ are pairwise affinities between pixels i and j. These are computed from pairwise distance measures $d^\mathcal{X}_{ij}$ and $d^\mathcal{Y}_{ij}$ as $\mathbf{A}^\mathcal{X}_{ij} = \exp(-(d^\mathcal{X}_{ij})^2/h_\mathcal{X})$ and $\mathbf{A}^\mathcal{Y}_{ij} = \exp(-(d^\mathcal{Y}_{ij})^2/h_\mathcal{Y})$ by use of the common Gaussian kernel function with kernel widths $h_\mathcal{X}$ and $h_\mathcal{Y}$.

The cross-domain pixel distance for pixel i is obtained as

$$\alpha_i = \frac{1}{n-1}\sum_{j=1}^{n} |\mathbf{A}^\mathcal{X}_{ij} - \mathbf{A}^\mathcal{Y}_{ij}|\,, \tag{1}$$

which is the average absolute affinity difference between pixel i and n other pixels. This assures that $\alpha_i \in [0,1]$, providing small values when pixel relations within the size n image patch or neighbourhood remains similar across image domains, and large values otherwise.

We will utilize α_i to suppress the influence of pixels with a high probability of change, and must therefore define a weighting function $\Pi(\alpha) : [0,1] \to [0,1]$ that is monotonically decreasing. Hence, the higher is $\Pi(\alpha)$, the lower is the probability of that pixel to be changed from one acquisition to the other, and the higher is the confidence to use it as a learning sample. We use the simple function $\Pi(\alpha) = 1 - \alpha$.

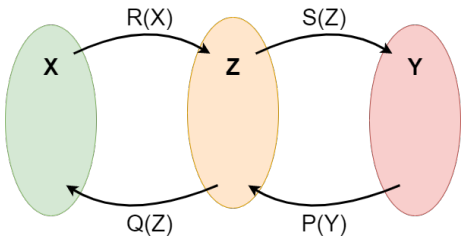


Fig. 1: Domain translation

2.2. Modified Canonical Correlation Analysis

Given two observations of the same object, linear CCA aims to find paired projections of the two views that make them maximally correlated [4]. Suppose that we have a sample of paired data vectors, $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$, where $\boldsymbol{x} \in \boldsymbol{R}^{C_\mathcal{X}}$ and $\boldsymbol{y} \in \boldsymbol{R}^{C_\mathcal{Y}}$. Let the $\boldsymbol{x}_i$ be stored as rows in the $n \times C_\mathcal{X}$ data matrix $\boldsymbol{X}$ and the y_i as rows in the $n \times C_\mathcal{Y}$ data matrix $\boldsymbol{Y}$. The goal is to find the solution of

$$\boldsymbol{u}^*, \boldsymbol{v}^* = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}} \operatorname{corr}(\boldsymbol{X}\boldsymbol{u}, \boldsymbol{Y}\boldsymbol{v})$$

with constraints $\|\boldsymbol{X}\boldsymbol{u}\| = 1$ and $\|\boldsymbol{Y}\boldsymbol{v}\| = 1$ and assuming that the correlation matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$ are non-singular.

The solutions for $\boldsymbol{u}$ and $\boldsymbol{v}$ are given as the left and right singular vectors of the matrix

$$\boldsymbol{R} = (\boldsymbol{X}^T\boldsymbol{X})^{-1/2}(\boldsymbol{X}^T\boldsymbol{Y})(\boldsymbol{Y}^T\boldsymbol{Y})^{-1/2}\,, \tag{2}$$

respectively [4]. We identify $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{Y}^T\boldsymbol{Y}$ as sample correlation matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\boldsymbol{X}^T\boldsymbol{Y}$ as their sample cross-correlation matrix. The novel way the CCA has been exploited in this work involves weighted computation of the sample covariances and cross-covariances. This can be done by replacing the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ by the premultiplied versions $\boldsymbol{W}\boldsymbol{X}$ and $\boldsymbol{W}\boldsymbol{Y}$, where $\boldsymbol{W}$ is a diagonal weight matrix with elements $\boldsymbol{W}_{ii} = \sqrt{\Pi(\alpha_i)}$, $i = 1, \ldots, n$.

As in ordinary CCA, we may select the desired number of projection vectors from the singular vectors associated with the highest singular values, and store them as the columns of projection matrices denoted $\boldsymbol{U}$ and $\boldsymbol{V}$.

2.3. Prior-Weighted Canonically Correlated Autoencoders

Deep convolutional autoencoders are CNNs that learn an encoding of the input, located in a central code layer, from which we can decode or reconstruct the input. The output is the reconstructed input, but we are often just as interested in the code layer representation, which can be forced to be compressed or sparse, depending on the neural network architecture. Inspired by [2] we have developed an architecture composed of two parallel autoencoders, one for each image domain, that share a code layer or latent space denoted $\mathcal{Z}$. In our case, we use CCA on the code layers to align them. When

code alignment is accomplished, we can couple encoders and decoders across domains to perform domain translation.

Figure 1 illustrates the code-aligned autoencoders. Let $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ denote data samples from image patches in $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$. The encoders aim to learn mapping functions $R(\boldsymbol{X}) : \mathcal{X} \to \mathcal{Z}$ and $P(\boldsymbol{Y}) : \mathcal{Y} \to \mathcal{Z}$ that encode information from image patches in $\mathcal{X}$ and $\mathcal{Y}$ into $\mathcal{Z}$. The decoders are represented by functions $S(\boldsymbol{Z}) : \mathcal{Z} \to \mathcal{X}$ and $Q(\boldsymbol{Z}) : \mathcal{Z} \to \mathcal{Y}$ that restore image patches in $\mathcal{X}$ and $\mathcal{Y}$ from codes in $\mathcal{Z}$. To train the network we combine the five loss terms described next (where the vector ϑ collects all parameters of both encoders and decoders).

Reconstruction loss: Given a paired data sample $\{\boldsymbol{X}, \boldsymbol{Y}\}$, the standard autoencoder reconstruction loss measures the mean squared errors between the input samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ and their reconstructions, $\hat{\boldsymbol{X}} = Q(R(\boldsymbol{X}))$ and $\hat{\boldsymbol{Y}} = S(P(\boldsymbol{Y}))$:

$$\mathcal{L}_{\text{rec}}(\vartheta) = \mathrm{E}_{\boldsymbol{X}}\left\{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F^2\right\} + \mathrm{E}_{\boldsymbol{Y}}\left\{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_F^2\right\}, \quad (3)$$

where $\|\cdot\|_F$ is the Frobenius matrix norm.

Weighted translation loss: The standard translation loss measures the discrepancies between the real $\boldsymbol{X}$ and $\boldsymbol{Y}$ and the translations $\check{\boldsymbol{X}} = Q(P(\boldsymbol{Y}))$ and $\check{\boldsymbol{Y}} = S(R(\boldsymbol{X}))$. Since in our case, the translation should only match for unchanged pixels, we weight the mean squared translation error by the affinity-based prior, as proposed in [2]:

$$\begin{aligned}\mathcal{L}_{\text{wtr}}(\vartheta) = \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\{\|\boldsymbol{W}(\boldsymbol{X} - \check{\boldsymbol{X}})\|_F^2\} \\ + \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\{\|\boldsymbol{W}(\boldsymbol{Y} - \check{\boldsymbol{Y}})\|_F^2\}\ .\end{aligned} \quad (4)$$

Cycle-consistency loss: A cyclically consistent network should recreate the input when performing a full translation cycle, such as $\tilde{\boldsymbol{X}} = Q(P(S(R(\boldsymbol{X}))))$ or $\tilde{\boldsymbol{Y}} = S(R(Q(P(\boldsymbol{Y}))))$. The cycle-consistency loss [8] is:

$$\mathcal{L}_{\text{cyc}}(\vartheta) = \mathrm{E}_{\boldsymbol{X}}\left\{\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_F^2\right\} + \mathrm{E}_{\boldsymbol{Y}}\left\{\|\boldsymbol{Y} - \tilde{\boldsymbol{Y}}\|_F^2\right\}. \quad (5)$$

Weighted CCA loss: Assume that $R(\boldsymbol{X})$ and $P(\boldsymbol{Y})$ have dimension $n \times C_{\mathcal{Z}}$. To assure that $R(\boldsymbol{X})$ and $P(\boldsymbol{Y})$ converge to a common space, while suppressing change pixels from training, we introduce the weighted CCA loss:

$$\begin{aligned}&\mathcal{L}_{wcca}(\vartheta) = \\ &- \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\left\{\|(\boldsymbol{W}R(\boldsymbol{X})\boldsymbol{U})^T(\boldsymbol{W}P(\boldsymbol{Y})\boldsymbol{V})\|_F^2\right\}.\end{aligned} \quad (6)$$

We use all available CCA components, thus the projection matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ have dimension $C_{\mathcal{Z}} \times C_{\mathcal{Z}}$. Note the minus in (6), as we want to maximize the crosscorrelation. We use ridge penalty regularization in the CCA, and the backpropagation scheme of [5].

L_2 loss: We use standard L_2 norm regularisation on all neural network weights to counteract overtraining.

Total loss: The overall loss function is defined as: $\mathcal{L}_{tot}(\vartheta) = \boldsymbol{\lambda}^T\boldsymbol{\mathcal{L}}$, where $\boldsymbol{\mathcal{L}}$ holds the five loss terms and $\boldsymbol{\lambda}$ contains weights that balance them.

3. EXPERIMENTAL RESULTS

The proposed method is tested on two datasets. The first consists of two multispectral optical images taken before and after a forest fire in Bastrop County, Texas, USA in October 2011. They were acquired by Landsat-5's TM instrument and the EO-1's ALI instrument with 7 and 10 channels, respectively. The second dataset covers a flood event in Sacramento, California, USA in January 2017. The before-image was acquired by Landsat-8 in eight channels from optical to long-wave infrared and the after-image by Sentinel-1A in two polarimetric SAR channels (VV and VH). Data and ground truth is provided by the authors of [9] and [1].

All four CNNs are implemented with fully convolutional layers, 100 filters in the hidden layers, and a coherent number of them in the first and last layer (i.e. the same number of the channel dimension of the input and output space respectively). No bottleneck is used in the autoencoders, hence the image dimension is preserved through all layers. The code space has dimension $H \times W \times 3$. We use a leaky ReLU activation function, a dropout rate of 20%, and an exponentially decaying learning rate initialised to 10^{-4}. Optimization was done with a minibatch gradient descent algorithm and the Adam optimizer run for 100 epochs. A patch size of 100×100 pixels was chosen to have enough information for meaningful covariance estimation in the CCA. The weights balancing the loss terms are: $\lambda_{rec} = \lambda_{\text{wtr}} = \lambda_{\text{cyc}} = 1$, $\lambda_{\text{wcca}} = 10^{-2}$ and $\lambda_{L_2} = 5 \cdot 10^{-5}$, chosen with grid search. The network appears robust to changes in all the parameters but λ_{wcca}.

The chosen evaluation metric is Cohen's kappa coefficient, κ [10]. We have evaluated the proposed method with change detection performed in code space, CCA-projected space and in the input spaces. The latter version performs best and most consistently, and the results of this variant are labelled DCCAE. This method is compared to the ACE-Net and X-Net algorithms developed in [2], in addition to our implementations of the SCCN [11] and cGAN [12] algorithms used as benchmarks in [2]. It was here shown that ACE-Net reaches state-of-the-art performance, whereas X-Net performs slightly worse, but more stably in terms of the κ variance. It is evident from the boxplots in Figure 2 and Figure 3 (boxes contain the 25 to 75 percentiles, whiskers extend to the 5 and 95 percentiles, and remaining data points are plotted as red $+$) that the proposed method combines the best features of the ACE-Net and X-Net and outperforms all considered algorithms, noting that the seemingly good performance of the SCCN algorithm on the California dataset is a side-effect of degenerate behaviour, as explained in [2]. The accurate detection results is also evident in the confusion maps of Fig. 4, where we have colour coded true positives (white), true negatives (black), false positives (green) and false negatives (red).

Reconstructed images, cross-domain translated images and code space images are not shown due to space limitations, but can be examined in [13]. These are vital to the

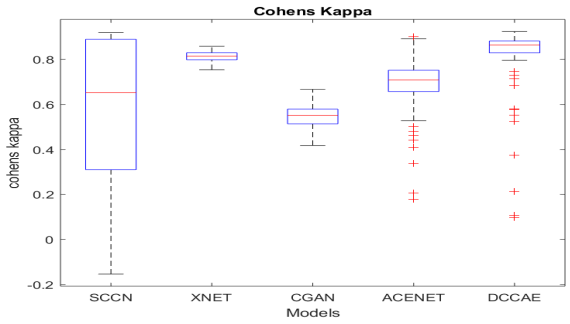


Fig. 2: Comparison of Cohen's κ on Texas dataset.

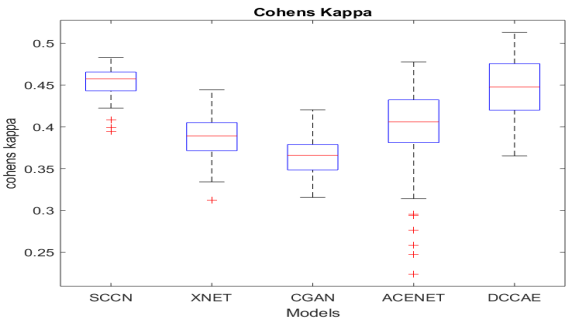
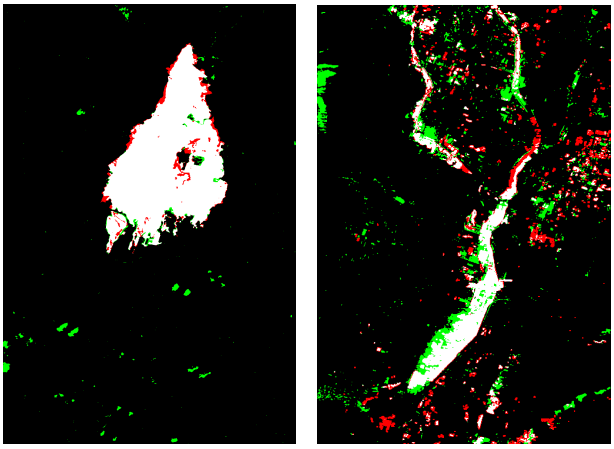


Fig. 3: Comparison of Cohen's κ on California dataset.

interpretation of performance and assessment of whether the algorithm works as intended. We find that input space images (original, translated and cyclically translated images) align consistently well, but that code images do not, and that change detection in code space is therefore not suitable. CCA-projected code images are well aligned, though, but the projection changes too abruptly between images and batches for change detection to be done in CCA-projected space. These observations warrent further investigation.

4. CONCLUSIONS

A new method for unsupervised heterogeneous change detection has been proposed by extending the approach in [2] with CCA weighted by an affinity-based prior. The validation on challenging multisensor datasets suggest the effectiveness of the unsupervised DCCAE method to generate accurate change maps from heterogeneous bitemporal data, also outperforming previous deep learning approaches in terms of median and variance of Cohen's kappa. These results confirm the potential of information extracted from local affinity matrices for this purpose. The proposed weighted CCA loss is shown capable of aligning code spaces without the need for adversarial training, which is often difficult or unstable.



(a) Confusion Map Texas (b) Confusion Map California

Fig. 4: Confusion maps with true positives (white), true negatives (black), false positives (green) and false negatives (red).

5. REFERENCES

[1] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, "Unsupervised image regression for heterogeneous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9960–9975, 2019.

[2] L. T. Luppino, M. C. Kampffmeyer, F. M. Bianchi, R. Jenssen, G. Moser, S. B. Serpico, and S. N. Anfinsen, "Deep image translation with an affinity-based change prior for unsupervised multimodal change detection," Oct 2020, arXiv:2001.04271.

[3] Y. Zhou, H. Liu, D. Li, H. Cao, J. Yang, and Z. Li, "Cross-sensor image change detection based on deep canonically correlated autoencoders," in *Proc. Int. Conf. Artificial Intell. Commun. Netw.*, 2019, pp. 251–257.

[4] T. De Bie, N. Cristianini, and R. Rosipal, "Eigenproblems in pattern recognition," in *Handbook of Geometric Computing*. Springer, 2005, pp. 129–167.

[5] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1247–1255.

[6] J. Yang, Y. Zhou, Y. Cao, and L. Feng, "Heterogeneous image change detection using deep canonical correlation analysis," in *Proc. Int. Conf. Pattern Recognition (ICPR)*, 2018, pp. 2917–2922.

[7] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1083–1092.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *IEEE Int. Conf. Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[9] M. Volpi, G. Camps-Valls, and D. Tuia, "Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis," *ISPRS J. Photogr. Remote Sens.*, vol. 107, pp. 50–63, 2015.

[10] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[11] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, March 2018.

[12] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A conditional adversarial network for change detection in heterogeneous images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 45–49, Jan 2019.

[13] F. Figari Tomenotti, "Heterogeneous change detection on remote sensing data with self-supervised deep canonically correlated autoencoders," Master's thesis, University of Genoa, Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture, 2020.