

# Deep Image Translation with an Affinity-Based Change Prior for Unsupervised Multimodal Change Detection

Luigi T. Luppino\*, Michael Kampffmeyer, Filippo M. Bianchi, Gabriele Moser, Sebastiano B. Serpico, Robert Jenssen, and Stian N. Anfinsen

**Abstract**—Image translation with convolutional neural networks has recently been used as an approach to multimodal change detection. Existing approaches train the networks by exploiting supervised information of the change areas, which, however, is not always available. A main challenge in the unsupervised problem setting is to avoid that change pixels affect the learning of the translation function. We propose two new network architectures trained with loss functions weighted by priors that reduce the impact of change pixels on the learning objective. The change prior is derived in an unsupervised fashion from relational pixel information captured by domain-specific affinity matrices. Specifically, we use the vertex degrees associated with an absolute affinity difference matrix and demonstrate their utility in combination with cycle consistency and adversarial training. The proposed neural networks are compared with state-of-the-art algorithms. Experiments conducted on two real datasets show the effectiveness of our methodology.

## I. INTRODUCTION

### A. Background

THE goal of change detection (CD) methods based on earth observation data is to recognise changes on Earth by comparing two or more satellite or aerial images covering the same area at different times [1]. Multi-temporal applications include the monitoring of long term trends, such as deforestation, urban planning, and earth resources surveys, whereas bi-temporal applications mainly regard the assessment of natural disasters, for example earthquakes, oil spills, floods, and forest fires [2]. This paper will focus on the latter case, and more specifically on the scenario where the changes must be detected from two satellite images with high to medium spatial resolution (10 to 30 meters). These resolutions allow to detect changes in ground coverage (forest, grass, bare soil, water etc.) on a very large area of several square kilometers, but are not suitable to deal with changes affecting small objects (buildings, trees, cars etc.). At these resolutions it is common to assume that co-registration can be achieved by applying simple image transformations such as translation, rotation, and re-sampling [3], [4], [5], [6]. This means that each pixel in the first image and its corresponding one in the second image represent the exact same point on the Earth. Consequently, even a

\*L.T. Luppino, M. Kampffmeyer, R. Jenssen and S.N. Anfinsen are with the Machine Learning Group, Department of Physics and Technology, UiT The Arctic University of Norway, e-mail: luigi.t.luppino@uit.no.

F.M. Bianchi is with NORCE Norwegian Research Center, Norway.

G. Moser and S.B. Serpico are with DITEN Department, University of Genoa, Italy.

simple pixel-wise operation (e.g. a difference or a ratio) would highlight changes when working with homogeneous data [4], [7], [8], i.e. data collected by the same kind of sensors, under the same geometries and seasonal or weather conditions, and using the same configurations and settings. More robust and efficient approaches consider complex algorithms rather than simple mathematical operations to detect changes, and many examples of homogeneous CD methods can be found in the literature [8], [9], [10], [11].

### B. Motivation

To rely on only one modality of data represents a limitation, both in terms of response time to sudden events and in terms of temporal resolution when monitoring long-term trends. The alternative is to combine heterogeneous data, which on one hand allows to exploit the capabilities of all the available sensors, but on the other hand raises additional challenges. Heterogeneous sensors usually measure different physical quantities, meaning that one terrain type might be represented by dissimilar statistical models from sensor to sensor, while surface signatures and their internal relations may change completely across different instruments [4], [7], [12]. In other words, it is not guaranteed that the data acquired by heterogeneous sources lie in a common domain, and a direct comparison is meaningless without processing and co-calibrating the data first [2].

Heterogeneous CD methods are meant to cope with these issues, and as discussed in [13], [14], there is not a unique way to categorize them. However, a principal taxonomy is the following: 1) unsupervised methods or supervised methods; 2) deep learning methods or traditional signal processing methods. The analysis in this paper will exclusively cover unsupervised frameworks: even though they must rely on some sort of self-supervision given their lack of supervised information about the change, they are more appealing than the supervised counterparts. Indeed, collecting labelled data is often costly and non-trivial, both in terms of the time and competence required [3], [15]. Concerning the second distinction, deep learning has become the state-of-the-art in many image analysis tasks, including in the field of remote sensing [4], [6]. Deep learning methods can achieve high performance thanks to the flexibility of neural networks, which are able to apply highly non-linear transformations to any kind of input data. For these reasons, the analysis of the literature will mainly focus

on deep learning, although many important methods, based on minimum energy [16], nonlinear regression [14], dictionary learning [13], manifold learning [17], or copula theory [18] can be mentioned.

### C. Proposed method

We propose a deep image translation approach to perform unsupervised CD based on heterogeneous remote sensing data. Most importantly, a comparison of domain-specific affinity matrices allows us to retrieve in a self-supervised manner the a priori change indicator driving our training process, referred to as the prior. This prior is reliable, informative and representative of the whole feature space, which is an advancement when compared to other priors previously used for heterogeneous CD, such as randomly initialised change maps, clustering/post-classification-comparison outputs, or supervised sample selection.

Two architectures are proposed: The X-Net is composed of two fully convolutional networks, each dedicated to map the data from one domain to the other; The ACE-Net consists of two autoencoders whose code spaces are aligned by adversarial training. Their performance and consistency are tested against two recent state-of-the-art methods. Summing up, the main contributions of this work are:

- A novel procedure to obtain a priori information on structural changes between the images based on a comparison of intramodal information on pixel relations.
- Two neural network architectures designed to perform unsupervised change detection, which explicitly incorporate this prior.
- Experimental results on two benchmark datasets that illustrate how the proposed networks perform favorably as compared to the state-of-the-art.

The implementations of our architectures are available at this link: [https://github.com/llu025/Heterogeneous\\_CD](https://github.com/llu025/Heterogeneous_CD), together with the re-implementation of the two reference methods and the two datasets used in this paper.

The remainder of this article is structured as follows: Section II describes the theoretical background and the related work. Section III introduces the reader to the notation, the proposed procedure and the architectures. Results on two datasets are presented in Section IV. Section V includes a discussion of the main features and drawbacks of each method used in this work. Section VI concludes the paper.

## II. RELATED WORK

The most common solution to compare heterogeneous data is to transform them and make them compatible. This is the main reason why many of the heterogeneous CD methods are related to the topics of domain adaptation and feature learning. In the following we list the main deep learning architectures that are found in the heterogeneous CD literature, along with some examples of methods implementing them.

### A. Stacked Denoising Autoencoders

1) *Background*: The autoencoder (AE) is a powerful deep learning architecture which has proven capable of solving

problems like feature extraction, dimensionality reduction, and clustering [19]. A denoising AE (DAE) is a particular type of AE trained to reconstruct an input signal that has been artificially corrupted by noise. The stacked denoising autoencoder (SDAE) is probably the most used model to infer spatial information from data and learn new representations and features. SDAEs are trained following the same procedure as DAEs, but their ability of denoising is learned in a layerwise manner by injecting noise into one layer at the time, starting from the outermost layer and moving on toward the innermost one [20]. In the following, some examples from the heterogeneous change detection literature are presented.

2) *Applications*: Su *et al.* [21] used change vector analysis to distinguish between the classes unchanged areas, positive changes and negative changes, as defined in [22]. They exploit two SDAEs to extract relevant features and transfer the data into a code space, where code differences from co-located patches are clustered to achieve a preliminary distinction between samples from the three classes. These samples are then used to train three distinct mapping networks, each of which learns to take the features extracted from image one as input and transform them into plausible code features related to image two. The goal of the first network is to reproduce the expected code from image two in case of a positive change, the second aims to do the same in case of a negative change, and the last takes care of the *no-change* case. A pixel is eventually assigned to the class corresponding to the reproduced code showing the smallest difference with the original code from image two.

In a very similar fashion, Zhang *et al.* [23] first use a spatial details recovery network trained on a manually selected set to coregister the two images, but then extract relevant features from them with two SDAEs trained in an unsupervised fashion. Starting from these transformed images, manual inspection, post-classification comparison or clustering provides a coarse change map. This is used to select examples of unchanged pairs of pixels, which are used to train a mapping network. Once the data are mapped into a common domain, feature similarity analysis highlights change pixels, which are isolated from the rest by segmentation;

In a paper by Zhan *et al.* [15], SAR data is log-transformed and stacked together with the corresponding optical data. Next, a SDAE is used to extract two relevant feature maps from the stack, one for both of the input modalities. These are then clustered separately and the results are compared to obtain a difference image. The latter is segmented into three clusters: pixels certain to belong to changed areas, pixels certain to belong to unchanged areas, and uncertain pixels. Finally, the pixels labelled with certainty are used to train a classification network which is then able to discriminate the uncertain pixels into the *change* and *no-change* clusters, providing the final binary change map.

Zhan *et al.* [3] proposed to learn new representative features for the two images by the use of two distinct SDAEs. A mapping network is then trained to transform these extracted features into a common domain, where the pixels are forced to be similar (dissimilar) according to their probability to belong to unchanged (changed) areas. The probability map

is initialised randomly and the training alternates between two phases: updating the parameters of the mapping network according to the probabilities, and updating the map according to the output of the network. Once the training reaches its stopping criterion, the difference between the two feature maps is obtained. Instead of producing a binary change map, this method introduces a hierarchical clustering strategy which highlights different types of change as separate clusters.

The symmetric convolutional coupling network (SCCN) was proposed by Liu *et al.* [4]: After two SDAEs are pretrained separately on each image, their decoders are removed, one of the encoders is frozen, and the other is fine-tuned by forcing the codes of the pixels most likely to not represent changes to be similar. The pixel probability of *no-change* is initialised randomly, and is updated iteratively and alternately together with the parameters of the encoders. A stable output of the objective function is eventually reached and the probability map is finally segmented into the usual binary change map. This method was later improved in [24] by modifying slightly the objective function and the probability map update procedure.

## B. Generative Adversarial Networks

1) *Background*: Among the most important methods in the literature of domain adaptation and data transformation are the generative adversarial networks (GANs). Proposed by Goodfellow *et al.* in [25], these architectures consist of two main components competing against each other. Drawing samples from a random distribution, a generator aims at reproducing samples from a specific target distribution as output. On the other hand, a discriminator has the goal to distinguish between *real* data drawn from the target distribution and *fake* data produced by the generator. Through an adversarial training phase, the generator becomes better at producing fake samples and it is rewarded when it fools the discriminator, whereas the latter improves its discerning skills and is rewarded when it is able to detect fake data. Both the two parts try to overcome their opponent and become better, benefiting from this competition.

A drawback of this method is the difficulty in balancing the strength of the two components. Their efforts have to be equal, otherwise one will start to dominate the other, hindering the simultaneous improvement of both. Conditional GANs [26] are a particular case, where the generator samples from a distribution conditioned on the input data. This architecture is suitable for the task of *image-to-image translation*: images from one domain are mapped into another (e.g. drawings or paintings into real pictures, winter landscapes into summer ones, maps of cities into aerial images).

2) *Applications*: The potential of this method to transform data acquired from one satellite sensor into another is glaring, and it was first explored in [27] to match optical and SAR images. The dataset used consists of pairs of co-located optical and SAR images acquired at the same time. The generator learns during the training to produce a plausible SAR image starting from the optical one, without knowing what the corresponding real SAR data look like. The same optical image and one of the two SAR images, either the generated or the

original, are provided to the discriminator which has to infer whether the data is *real* or *fake*. For testing, the generator takes the optical images as input and provides the synthetic SAR data, whereas the original SAR data becomes the ground truth.

In [7], the same concept is applied to perform heterogeneous CD. The scheme is always the same: a generator tries to reproduce SAR patches starting from the corresponding optical ones, and a discriminator aims at detecting these *fake* patches. In order to facilitate a direct comparison, they introduce an approximation network which learns to transform the original SAR patches into the generated ones. Note that the training of all these networks must be carried out on patches not containing changing pixels, and any other patch must be flagged and excluded from this process. At first all the flags are set to *no-change*, then these steps are iterated: the conditional GAN is updated, the approximation network is tuned accordingly, and finally the generated and approximated patches are compared to flag the ones containing changes. Once the training phase is over, the generated image and the approximated image are pixel-wise subtracted and segmented binarily.

## C. Cyclic Generative Adversarial Networks

1) *Background*: A more complex framework than the conditional GAN is the cycle GAN [28]. The idea is simple: instead of using just one generator-discriminator couple dealing with the transformation from domain X to domain Y, another tandem generator-discriminator is added to do the vice versa. This means that the framework can be tested for so-called *cycle consistency*: It should be possible to perform a composite translation of data from domain X to domain Y, and then further to domain X (denoted  $X \rightarrow Y \rightarrow X$ ), and the full translation cycle should reproduce the original input. Equivalently, the cycle  $Y \rightarrow X \rightarrow Y$  should reproduce the original input in domain Y.

In [29], this framework is applied and extended further: Along with the two input domains X and Y, a latent space Z is introduced in the middle between them. Data from the original domains are transformed to Z, where they should ideally not be discernible. Thus, four generators are used to map data across domains: from X to Z, from Z to Y, from Y to Z, and from Z to X. The accurate reconstruction of the images is the first enforced principle: Data mapped from domain X (Y) to Z must be mapped back correctly to X (Y). The next requirement is cycle-consistency: Starting from X (Y) and going first to Z and then to Y (X), the images must go back to X (Y) passing through Z again and match exactly with the original input. Concerning the discriminators, there are three: one should distinguish whether data mapped into Z come originally from X or Y; another discriminates between original images from X and images which started from Y and performed half a cycle; the third does the same in the domain Y.

2) *Applications*: Inspired by these concepts, Gong *et al.* proposed the coupling translation networks to perform heterogeneous CD [12]. However, their architecture is simpler. Two variational AEs are combined so that their encoders separately take as input optical and SAR patches, respectively,

and the two codes produced are stacked together. The stacked code is then decoded by both decoders and each of which yields two output patches: one is the reconstruction of the input patch from the same domain, the other is the transformation of the input patch from the opposite domain. The latter must be detected by a discriminator which is taught to discern reconstructed data from *fake* transformed data. This framework has only two discriminators, one after each decoder, whereas the code spaces of the two AEs are aligned throughout the training, eventually becoming the common latent domain, namely  $Z$ . Together with the adversarial loss, the reconstruction and the cycle-consistency drive the learning process, which enables the two networks to translate data across domains, such that a direct comparison is feasible.

In the following section we explain how our methodology fits in this picture, framed in-between cycle-consistency and adversarial training.

### III. METHODOLOGY

The same geographical region is scanned by sensor  $\mathcal{X}$  at time  $t_1$ , producing an image  $\mathbf{X} \in \mathbb{R}^{H \times W \times C_1}$ , and by sensor  $\mathcal{Y}$  at time  $t_2$ , yielding the image  $\mathbf{Y} \in \mathbb{R}^{H \times W \times C_2}$ .  $H$  and  $W$  denote the common height and width of the two images, with  $C_1$  and  $C_2$  as the respective number of channels. The common dimensions are obtained through re-sampling and co-registration.

We further assume that a limited part of the image has changed between time  $t_1$  and  $t_2$ . The final goal of the presented method is to transform data consistently from one domain to the other. To do so, it is crucial to learn a one-to-one mapping between the land cover signatures of one domain and the corresponding signatures in the other. Since no prior information is available, the only reasonable option is to learn a mapping from every pixel in  $\mathbf{X}$  to the corresponding pixel in  $\mathbf{Y}$  and vice versa.

A possibility would be to train two regression functions  $\hat{\mathbf{Y}} = F(\mathbf{X})$  and  $\hat{\mathbf{X}} = G(\mathbf{Y})$  by using the entire images  $\mathbf{X}$  and  $\mathbf{Y}$  as they are. However, the presence of areas affected by changes would distort the learning process, because they would promote a transformation from one land cover in one domain to a different land cover in the other domain; For example, forests and fire scars may be erroneously connected, as may land and flooded land. To reduce the effect of these areas on training, we first perform a preliminary analysis to highlight changes. Then, the contribution of each pixel to the learning process is inversely weighted with the probability (or what we interpret as a probability) of it being affected by a change. In this section, we first describe the algorithm providing the preliminary change analysis. Then, two deep learning architectures are proposed, and we explain how they can exploit the prior computed in the change analysis.

#### A. Prior computation

To compute a similarity based on affinity matrices, we present an improved version of the original method proposed in our previous work [14]. A  $k \times k$  sliding window covers an area  $p$  of both  $\mathbf{X}$  and  $\mathbf{Y}$ , from which a pair of corresponding

patches  $p^{\mathbf{X}}$  and  $p^{\mathbf{Y}}$  are extracted.  $p_i^l$  stands for pixel  $i$  of patch  $p$  and the modality  $l$ , with  $i \in \{1, \dots, k^2\}$  and  $l \in \{\mathbf{X}, \mathbf{Y}\}$ . The distance between a pixel pair  $(i, j)$  in patch  $p$  with modality  $l$  is defined as  $d_{i,j}^l$ . The appropriate choice of distance measure depends on the underlying data distribution. The hypothesis of normality for imagery acquired by optical sensors is commonly assumed [22], [30]. Concerning SAR intensity data, a logarithmic transformation is sufficient to bring it to near-Gaussianity [2], [15]. We use the computationally efficient Euclidean distance, as it is suitable for (nearly) Gaussian distributed data.

Once computed, the distances between all pixel pairs can be converted to affinities, for instance by the Gaussian kernel:

$$A_{i,j}^l = \exp \left\{ -\frac{(d_{i,j}^l)^2}{h^2} \right\} \in (0, 1], \quad i, j = 1, \dots, k^2. \quad (1)$$

$A_{i,j}^l$  are the entries of the affinity matrix  $A^l \in \mathbb{R}^{k^2 \times k^2}$  for the given patch and modality.  $h$  is the kernel width, which can be automatically determined. Our choice is to set it equal to the average distance to the  $K^{th}$  nearest neighbour for all data points in  $p^l$ , with  $K = \frac{3}{4}k^2$ . In this way, a characteristic distance within the patch is captured by this heuristic, which is robust with respect to outliers [31]. Silverman's rule of thumb [32] and other common approaches to determine the kernel width have not proven themselves effective in our experimental evaluation, so they were discarded. Once the two affinity matrices are computed, their element-wise absolute difference  $D = |A^{\mathbf{X}} - A^{\mathbf{Y}}|$  can be obtained.

At this point, [14] evaluated the Frobenius norm of  $D$  and assigned this value to all the pixels belonging to  $p$ . Then, the  $k \times k$  window shifts by one pixel, and all these operations must be repeated for the whole set  $\mathcal{P}$  of overlapping patches  $p$  inside  $\mathcal{l}$ . Clearly, the loop over all these patches is computationally heavy. If  $M = H \cdot W$  is the total number of pixels, the cardinality of  $\mathcal{P}$  is

$$|\mathcal{P}| = (H - k + 1) \cdot (W - k + 1) \\ = M + (k - 1)^2 - (k - 1)(H + W) \quad (2)$$

The final result for each pixel is derived by averaging the set  $\mathcal{S}^F$  of Frobenius norms obtained with all the patches covering that pixel. Shifting the sliding window by a factor larger than one would speed up the algorithm, but with a much poorer result: intuitively, the final outcome would exhibit an unnatural tile pattern.

To address this issue, we propose to compute the following mean over the rows of  $D$  (or columns, since  $A^l$  and consequently  $D$  are symmetrical):

$$\alpha_i = \frac{1}{k^2} \sum_{j=1}^{k^2} |A_{i,j}^{\mathbf{X}} - A_{i,j}^{\mathbf{Y}}|, \quad i = 1, \dots, k^2 \quad (3)$$

The main rationale for this operation is that pixels affected by changes are the ones perturbing the structural information captured by the affinity matrices, and so, on average, their corresponding rows in  $D$  should present larger values. Thus,  $\alpha_i$  not only contains a more reliable information, but, most importantly, it relates only to pixel  $i$ , instead of being the same

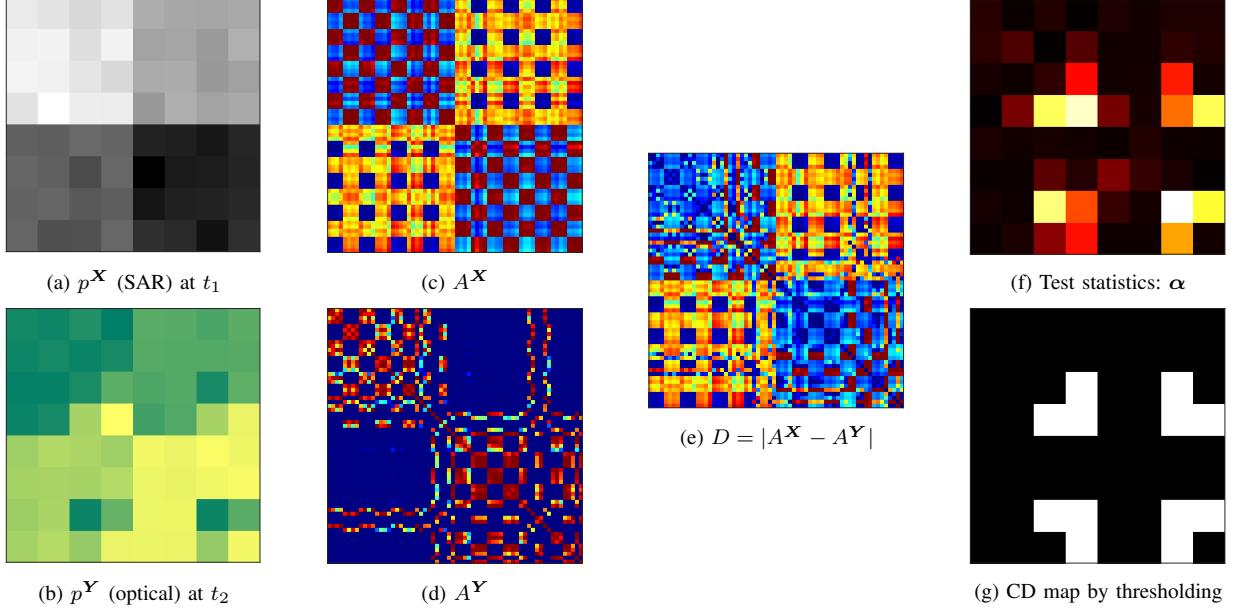


Fig. 1: Toy example: a) Patch from the SAR image at time  $t_1$ ; b) Corresponding patch in the optical image at time  $t_2$ ; c-e) Affinity matrices and their absolute difference; f) Test statistics  $\alpha$  obtained from  $D$  by applying Eq. 3; g) CD map obtained by thresholding  $\alpha$ . Best viewed in colour.

for all the pixels in  $p$ . Therefore, it is possible to introduce a shift factor  $\Delta > 1$ : on one hand the final result for each pixel would be the average (of yet more informative values) over a smaller set  $\mathcal{S}^\alpha$ , but on the other hand the process is sped up considerably. Potentially, this shift can be as large as the patch size, reducing the amount of patches by a factor of  $k^2$ . However, this is not desirable, since each pixel would be covered only once and the average over its set  $\mathcal{S}^\alpha$  would be evaluated on one element only.

The toy example in Fig. 1 helps to explain the effectiveness of the proposed approach. To make this case easier to explain,  $\Delta$  is set equal to  $k$ : each pixel in the image is covered only once. Fig. 1a simulates a patch  $p^X$  of  $8 \times 8$  pixels extracted from a SAR image captured at  $t_1$ . It consists of blocks representing four different classes, whose pixel intensities are affected by multiplicative speckle noise. The corresponding patch  $p^Y$  extracted from an optical image at  $t_2$  is depicted in Fig. 1b: the same classes are disposed in the same way and the pixel intensities are affected by additive Gaussian noise. Note that the intensities of one class across the two images are not correlated, so they can be similar, as for the classes in the top-right and bottom-left blocks of  $p$ , or completely different, as for the other two classes. Changes are introduced by placing 4 pixels representing each class in the bottom right quadrant of each block of  $p^Y$ . In this way, all the possible transitions between one class and the others occur between  $t_1$  and  $t_2$ . The  $64 \times 64$  affinity matrices  $A^X$  and  $A^Y$  computed from  $p^X$  and  $p^Y$  are depicted in Fig. 1c and 1d. They both show a regular squared pattern with high affinities in red and low affinities in blue, but the latter presents clear irregularities and perturbations due to the changed pixels which are breaking the block pattern in Fig. 1b. Once  $D$  is evaluated (Fig. 1e), Eq. 3 yields the  $8 \times 8$  result of Fig. 1f, where a darker (brighter) pixel

means a smaller (larger)  $\alpha_i$ . Finally, one may even retrieve a CD map by thresholding  $\alpha$ , as shown in Fig. 1g.

---

**Algorithm 1** Evaluation of  $\alpha$ :

---

```

for all patches  $p_\ell, \ell \in \{1, \dots, |\mathcal{P}|\}$  do
    Compute  $d_{i,j}^X \forall i, j \in p_\ell^X$  and  $d_{i,j}^Y \forall i, j \in p_\ell^Y$ 
    Determine  $h_\ell^X$  and  $h_\ell^Y$ 
    Compute  $A_{i,j}^l = \exp \left\{ - \left( \frac{d_{i,j}^l}{h_\ell^l} \right)^2 \right\}, l = X, Y$ 
    Compute  $\alpha_{i,\ell} = \frac{1}{k^2} \sum_j |A_{i,j}^X - A_{i,j}^Y| \forall i \in p_\ell$ 
    Add  $\alpha_{i,\ell}$  to the set  $\mathcal{S}_i^\alpha \forall i \in p_\ell$ 
end for
for all pixels  $i \in [1, \dots, M]$  do
    Compute  $\alpha_i = \frac{1}{|\mathcal{S}_i^\alpha|} \sum_{\alpha_{i,\ell} \in \mathcal{S}_i^\alpha} \alpha$ 
end for

```

---

Given the set  $\mathcal{P}$  of all size  $k \times k$  image patches spaced by a step size  $\Delta$ , Algorithm 1 summarises the procedure to obtain a set of priors  $\{\alpha_i\}_{i=1}^M$  for the whole dataset. For each pixel  $i \in \{1, \dots, M\}$  in the image, the mean over  $\mathcal{S}_i^\alpha$  is computed, where  $\mathcal{S}_i^\alpha$  is the set of the  $\alpha_{i,\ell}$  obtained with all the patches  $p_\ell \in \mathcal{P}$  covering pixel  $i$ . If  $\Delta$  is a factor of  $k$ , this average is calculated over  $(k/\Delta)^2$  values.

The size  $k$  has an important role in the effectiveness of this methodology, because the patches  $p$  could be too small or too big to capture the shapes and the patterns within them. To reduce the sensitivity to this parameter, one may suggest to use different values of  $k$  for Algorithm 1 and combine the results in an ensemble manner. For example, once  $k$  is defined the method can be applied also for  $k_{small} = k/2$  and  $k_{big} = 2 \cdot k$ . However, the size of the matrices containing first

$d_{i,j}^l$  and then  $A_{i,j}^l$  exhibits a quadratic growth with respect to  $k$ , thus becoming quickly unfeasible in terms of memory usage and computational time. Hence, instead of applying the method to the original images with  $k_{big}$ , we suggest to down-sample the images by a factor of 2, apply the algorithm with  $k$ , and re-scale the output to the original size. This procedure might introduce artifacts and distortions, but their effects are mitigated when combined with the results obtained with  $k_{small}$  and  $k$ .

In the following subsections, we explain how to exploit the outcome of Algorithm 1 to train the proposed deep learning architectures in absence of supervision.

### B. X-Net

The main goal of our approach is to map data across two domains. As Fig. 3 illustrates, this means to train a function  $F(\mathbf{X})$  able to transform data from  $\mathbf{X}$  to  $\mathbf{Y}$ , and a second function  $G(\mathbf{Y})$  able to do the opposite. The two maps can be implemented as two convolutional neural networks (CNNs), each one taking data from one domain as input and transferring it into the other domain as output. Hence, the training can be carried out by the minimisation of an objective function with respect to the set  $\vartheta$  of parameters of the two networks. The objective function, commonly referred to as the loss function  $\mathcal{L}$ , is defined *ad hoc* and usually consists of a weighted sum of terms, where each relates to a specific objective or property that we want from the solution. For this particular framework, we introduce three terms. Note that from now on we refer to training patches of size  $N \times N$ , with  $N$  much larger than the

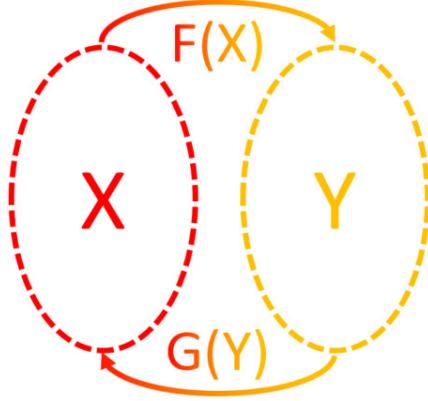


Fig. 3: First proposed framework: two domains and two transformations which can translate data across them.

patch size  $k$  of Sec. III-A, and each of their pixels indexed with  $n \in [1, \dots, N^2]$ .

1) *Prior-weighted translation loss*: For those pixels not affected by changes, we must enforce that

$$\begin{aligned}\hat{\mathbf{y}}_n &= F(\mathbf{x}_n) \simeq \mathbf{y}_n \\ \hat{\mathbf{x}}_n &= G(\mathbf{y}_n) \simeq \mathbf{x}_n\end{aligned}\quad (4)$$

where  $\hat{\mathbf{Y}} = F(\mathbf{X})$  and  $\hat{\mathbf{X}} = G(\mathbf{Y})$  stand for the data transformed from one domain into the other. From the opposite perspective, pixels which are likely to be changed shall not fulfill these same requirements. Ergo, the first term of the loss

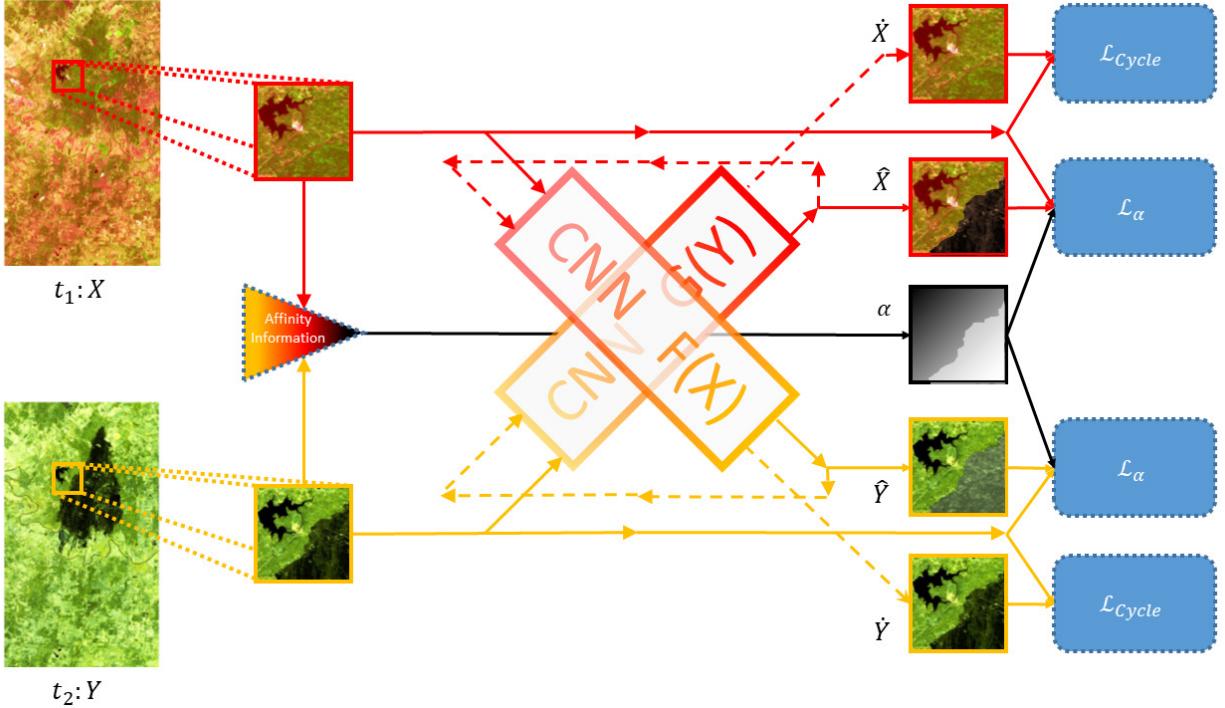


Fig. 2: X-Net: two CNNs transform data from the domain of  $\mathbf{X}$  to the domain of  $\mathbf{Y}$  and vice versa. Solid lines going through them indicate data transferred from one domain to the other, dashed lines indicate data re-transformed back to their original domain.

function is defined as follows:

$$\begin{aligned}\mathcal{L}_\alpha(\vartheta) = & \mathbb{E}_{\mathbf{X}, \mathbf{Y}, \alpha} \left[ \frac{1}{N^2} \sum_n \|F(\mathbf{x}_n) - \mathbf{y}_n\|_2^2 \Pi(\alpha_n) \right] + \\ & \mathbb{E}_{\mathbf{X}, \mathbf{Y}, \alpha} \left[ \frac{1}{N^2} \sum_n \|G(\mathbf{y}_n) - \mathbf{x}_n\|_2^2 \Pi(\alpha_n) \right]\end{aligned}\quad (5)$$

where the contribution to the loss of pixel  $n$  is weighted by its prior  $\Pi(\alpha_n)$ , and  $\alpha$  holds the  $\alpha_n$  associated with the training patches  $\mathbf{X}$  and  $\mathbf{Y}$ .  $\Pi(\cdot)$  is a function defined over  $[0, 1]$  and returns values in the same interval. Note that a high  $\alpha$  indicates a high chance of change, so  $\Pi(\alpha)$  must be a monotonic decreasing function. In this way, we use the preliminary information obtained in Sec. III-A to drive the learning procedure and penalise the contribution of pixels most likely to be affected by changes. We selected  $\Pi(\alpha) = 1 - \alpha$ , but other choices can be considered.

2) *Cycle-consistency loss*: While developing the CycleGANs [28], Zhu *et. al* already pointed out that these transformations must respect the principle of cycle-consistency: ideally, if  $F(\mathbf{X})$  and  $G(\mathbf{Y})$  are perfectly tuned, it must hold true that

$$\begin{aligned}\hat{\mathbf{x}}_n &= G(\hat{\mathbf{y}}_n) = G(F(\mathbf{x}_n)) \simeq \mathbf{x}_n \\ \hat{\mathbf{y}}_n &= F(\hat{\mathbf{x}}_n) = F(G(\mathbf{y}_n)) \simeq \mathbf{y}_n\end{aligned}\quad (6)$$

where  $\hat{\mathbf{X}} = G(\hat{\mathbf{Y}})$  and  $\hat{\mathbf{Y}} = F(\hat{\mathbf{X}})$  indicate the data re-transformed back to the original domains. Consequently, we define the second loss term as:

$$\begin{aligned}\mathcal{L}_{\text{Cycle}}(\vartheta) = & \mathbb{E}_{\mathbf{X}} \left[ \frac{1}{N^2} \sum_n \|G(F(\mathbf{x}_n)) - \mathbf{x}_n\|_2^2 \right] + \\ & \mathbb{E}_{\mathbf{Y}} \left[ \frac{1}{N^2} \sum_n \|F(G(\mathbf{y}_n)) - \mathbf{y}_n\|_2^2 \right]\end{aligned}\quad (7)$$

Finally, the optimisation is carried out by seeking the global minimum of  $\mathcal{L}$  with respect to  $\vartheta$ :

$$\min_{\vartheta} \mathcal{L}(\vartheta) = \min_{\vartheta} \left\{ w_c \mathcal{L}_{\text{Cycle}}(\vartheta) + w_\alpha \mathcal{L}_\alpha(\vartheta) + w_\vartheta \|\vartheta\|_2^2 \right\}.\quad (8)$$

The third and last term of  $\mathcal{L}$  is a weights decay regularisation term, which reduce overfitting by controlling the magnitude of the network parameters.

Fig. 2 shows the scheme of the X-Net: one CNN plays the role of  $F(\mathbf{X})$ , the other represents  $G(\mathbf{Y})$ . Solid lines going through them indicate data transferred from one domain to the other, dashed lines indicate data re-transformed back to their original domain. The patches from  $\mathbf{X}$  and  $\mathbf{Y}$  are used both as input and targets for the CNNs. Recall that  $\alpha$  is a  $H \times W$  image computed in advance, as explained in Sec. III-A. Nevertheless, the patch extracted from it is deliberately depicted as computed on the fly for an easier representation in Fig. 2.

### C. ACE-Net: Adversarial Cyclic Encoders Network

Inspired by the idea proposed in [29], we expand the X-Net framework by introducing a latent space  $\mathbf{Z}$  between domain  $\mathbf{X}$  and domain  $\mathbf{Y}$ . Differently from the X-Net, this architecture consists of four networks representing four regression functions (see Fig. 4):  $R(\mathbf{X})$  and  $P(\mathbf{Y})$  transform data from

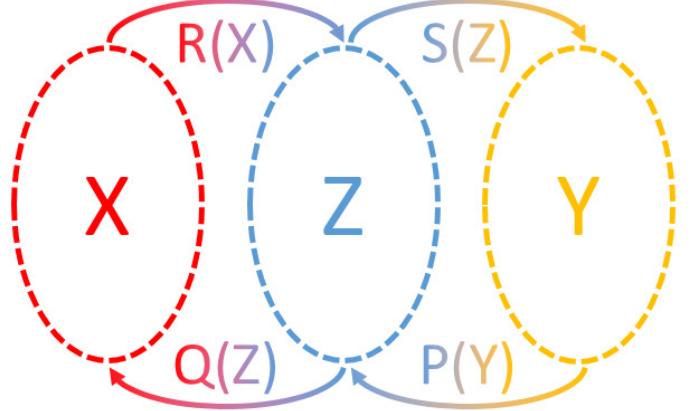


Fig. 4: Second proposed framework: a latent space  $\mathbf{Z}$  is introduced between domains  $\mathbf{X}$  and  $\mathbf{Y}$ , and four regression functions mapping data across them. In this case,  $F(\mathbf{X}) = S(R(\mathbf{X}))$  and  $G(\mathbf{Y}) = Q(P(\mathbf{Y}))$ .

the original domains into the new common space, whereas  $S(\mathbf{Z})$  maps latent space data into  $\mathbf{Y}$  and  $Q(\mathbf{Z})$  does the same into  $\mathbf{X}$ . However, it is simple to notice an analogy between the two schemes. Namely,  $F(\mathbf{X}) = S(R(\mathbf{X}))$  and  $G(\mathbf{Y}) = Q(P(\mathbf{Y}))$ . Therefore, we can include the same loss terms introduced before: cycle-consistency and prior-weighted similarity. In this case,

$$\begin{aligned}\hat{\mathbf{X}} &= G(\mathbf{Y}) = Q(P(\mathbf{Y})) \\ \hat{\mathbf{Y}} &= F(\mathbf{X}) = S(R(\mathbf{X})) \\ \dot{\mathbf{X}} &= G(\hat{\mathbf{Y}}) = Q(P(S(R(\mathbf{X})))) \\ \dot{\mathbf{Y}} &= F(\hat{\mathbf{X}}) = S(R(Q(P(\mathbf{Y}))))\end{aligned}\quad (9)$$

Nonetheless, the new framework allows to define two additional loss terms.

1) *Reconstruction Loss*: Let us consider the two networks representing  $R(\mathbf{X})$  and  $P(\mathbf{Y})$  as two encoders mapping the data into a code space. In the same way, the networks associated with  $Q(\mathbf{Z})$  and  $S(\mathbf{Z})$  can be thought as their respective decoders which transform data back to the original domains. Hence, the cascade of  $R(\mathbf{X})$  with  $Q(\mathbf{Z})$  is equivalent to an AE, and the same can be said the composite of  $P(\mathbf{Y})$  and  $S(\mathbf{Z})$ . AEs have the goal to reproduce their input as faithfully as possible in output, which means that for the reconstructed images  $\tilde{\mathbf{X}}$  and  $\tilde{\mathbf{Y}}$

$$\begin{aligned}\tilde{\mathbf{x}}_n &= Q(R(\mathbf{x}_n)) \simeq \mathbf{x}_n \\ \tilde{\mathbf{y}}_n &= S(P(\mathbf{y}_n)) \simeq \mathbf{y}_n\end{aligned}\quad (10)$$

must hold true. Consequently, we introduce the following reconstruction loss term:

$$\begin{aligned}\mathcal{L}_{\text{Recon}}(\vartheta_{AEs}) = & \mathbb{E}_{\mathbf{X}} \left[ \frac{1}{N^2} \sum_n \|Q(R(\mathbf{x}_n)) - \mathbf{x}_n\|_2^2 \right] + \\ & \mathbb{E}_{\mathbf{Y}} \left[ \frac{1}{N^2} \sum_n \|S(P(\mathbf{y}_n)) - \mathbf{y}_n\|_2^2 \right],\end{aligned}\quad (11)$$

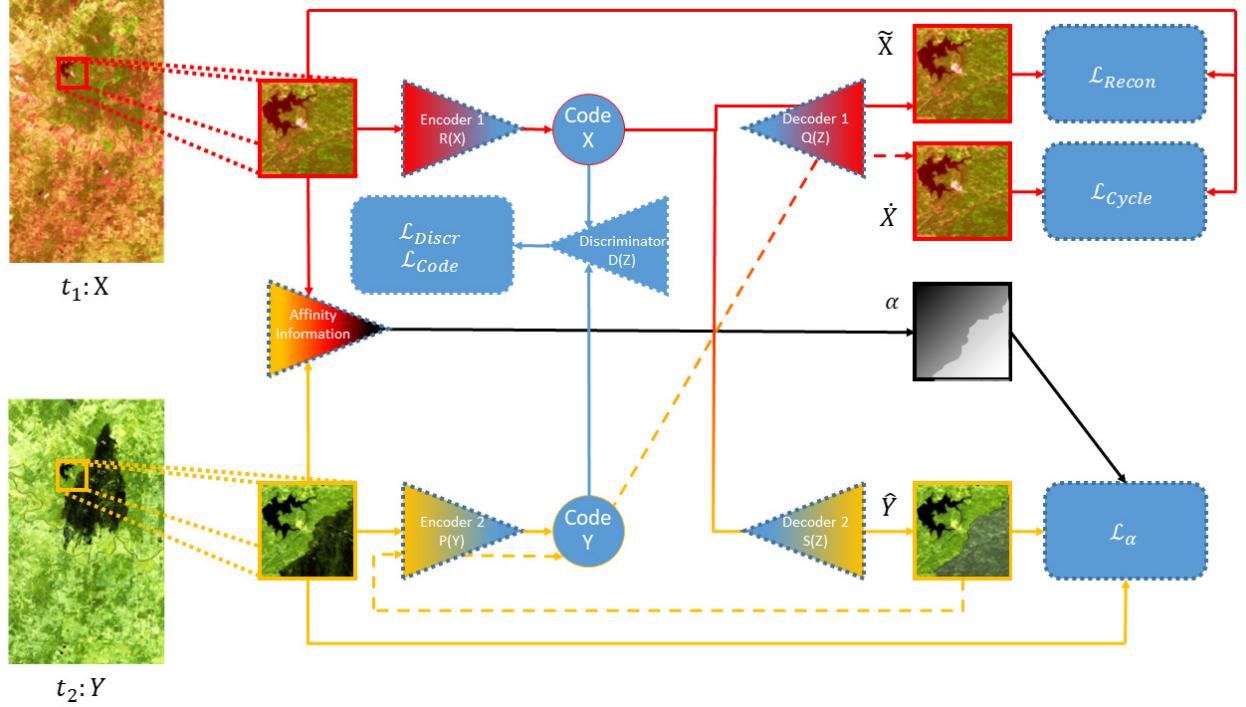


Fig. 5: ACE-Net: the encoders  $R(\mathbf{X})$  and  $P(\mathbf{Y})$  transform incompatible data into two code spaces, which are aligned by adversarial training against the discriminator  $D(\mathbf{Z})$ . The decoders  $Q(\mathbf{Z})$  and  $S(\mathbf{Z})$  are taught to map data from the latent space back into the original spaces. For simplicity, only the loss terms related to  $\mathbf{X}$  and their corresponding data flows are depicted. Dash lines refer to data which have been transformed already once, have gone through the framework again and have been transformed back into their original domain.

where  $\vartheta_{AEs}$  denotes all parameters in the autoencoders, consisting of the CNNs  $R(\mathbf{X})$ ,  $S(\mathbf{Z})$ ,  $P(\mathbf{Y})$  and  $Q(\mathbf{Z})$ .

2) *Adversarial Code Alignment Losses*: Even after considering the cycle-consistency loss and the prior-weighted translation loss, there is still no guarantee that the latent domain is the same for both the AEs. Each AE transforms its input into a code space, but without any additional requirements, these two code spaces might be either very close or diverging completely. In order to enforce their alignment, we feed a discriminator with a stack of the two codes. The discriminator is rewarded if it is able to distinguish the codes, whereas the generators (i.e. the encoders) are penalised when the discriminator succeeds. Let successful discrimination be defined as:  $D(R(\mathbf{x}_n)) = 1$  and  $D(P(\mathbf{y}_n)) = 0$ , thus last two loss terms become:

$$\begin{aligned} \mathcal{L}_{\text{Discr}}(\vartheta_D) = & \mathbb{E}_{\mathbf{X}} \left[ \frac{1}{N^2} \sum_n (D(R(\mathbf{x}_n)) - 1)^2 \right] + \\ & \mathbb{E}_{\mathbf{Y}} \left[ \frac{1}{N^2} \sum_n (D(P(\mathbf{y}_n)))^2 \right], \end{aligned} \quad (12)$$

$$\begin{aligned} \mathcal{L}_{\text{Code}}(\vartheta_{P,R}) = & \mathbb{E}_{\mathbf{X}} \left[ \frac{1}{N^2} \sum_n (D(R(\mathbf{x}_n)))^2 \right] + \\ & \mathbb{E}_{\mathbf{Y}} \left[ \frac{1}{N^2} \sum_n (D(P(\mathbf{y}_n)) - 1)^2 \right], \end{aligned} \quad (13)$$

where the discrimination loss  $\mathcal{L}_{\text{Discr}}$  is used to adjust the parameters  $\vartheta_D$  of the discriminator. The code layer is used as generator, and the code loss  $\mathcal{L}_{\text{Code}}$  is used to train the parameters  $\vartheta_{P,R}$  of the CNNs  $R(\mathbf{X})$  and  $P(\mathbf{Y})$  that generate

the codes. The adversarial scheme is evident from Eq. (12) and (13), the two generators and the discriminator aim at the opposite goal and, therefore, have opposite loss terms. As in [28], we choose an adversarial objective function based on mean squared errors rather than a logarithmic one. Note that two discriminators could also have been placed after the decoders to distinguish transformed *fake* data from the reconstructed ones, as in [12]. However, to train two additional networks and find a good balance between all the involved parties is not trivial and require the correct design of each and every network in the architecture, on top of which fine-tuning of all the involved weights must be carried out. In conclusion, we decided to have a less complex framework with just one discriminator for the code space. Fig. 5 show the schematics of the ACE-Net. For simplicity, the arrows represent the data flow involving only the loss terms related to  $\mathbf{X}$ .  $\mathbf{Y}$  in this image is used only as label for  $\mathcal{L}_\alpha$  and to produce its code. The flow diagram for loss terms related to  $\mathbf{Y}$  would be symmetric. Solid arrows represent images going through the AEs only once (namely  $\hat{\mathbf{X}}$  and  $\hat{\mathbf{Y}}$ ), dashed arrows are the second half of the cycle leading to  $\dot{\mathbf{X}}$ . The discriminator  $D(\mathbf{Z})$  takes as input  $R(\mathbf{X})$  and  $P(\mathbf{Y})$  and tries to tell them apart.

Once the X-Net and the ACE-Net are trained and the transformed images  $\hat{\mathbf{X}}$  and  $\hat{\mathbf{Y}}$  obtained, two distance images can be computed, namely  $d_m^{\mathbf{X}} = \|\hat{\mathbf{x}}_m - \mathbf{x}_m\|_2$  and  $d_m^{\mathbf{Y}} = \|\hat{\mathbf{y}}_m - \mathbf{y}_m\|_2$  for all pixels  $m \in [1, \dots, M]$ . These two are normalised and combined together so that changes are highlighted, whereas false alarms which are present in only

one of the two distance images are de-emphasised. Outliers might affect the two normalisations, so the distances in  $d^X$  and  $d^Y$  beyond three standard deviations of the mean values are clipped.

We combine the normalised distance images with a simple average and obtain the final difference image  $d$ . Then, the latter is filtered first and thresholded afterwards to achieve a binary segmentation, i.e. the final goal of a CD method: the change map.

Concerning the filtering, the method proposed in [33] is used. It exploits spatial context to filter  $d$  with fully connected conditional random field models. It defines pairwise edge potentials between all pairs of pixels in the image by a linear combination of Gaussian kernels in an arbitrary feature space. The main downside of the iterative optimisation of the random field is that it requires the propagation of all the potentials across the image. However, this highly efficient algorithm reduces the computational complexity from quadratic to linear in the number of pixels by approximating the random field with a mean field whose iterative update can be computed using Gaussian filtering in the feature space. The number of iterations and the kernel width of the Gaussian kernels are the only hyperparameters manually set, and we opted to tune them according to [14]: 5 iterations and a kernel width of 0.1.

Finally, it is fundamental to threshold the filtered difference image correctly: a low threshold yields unnecessary false alarms, vice versa a high one leads to possibly dangerous undetected changes. Determining this boundary heuristically is not convenient, because it requires a visual inspection and an arbitrary evaluation. Methods such as [34], [35], [36], [37] are able to set this threshold automatically, therefore they are preferable. Among these, we selected the well known Otsu's method [34].

#### IV. EXPERIMENTAL RESULTS

In Sec. IV-A we first provide the details of our implementation of the various methods. Then, the two data sets used in this work are presented in Sec. IV-B. Moving on, the proposed prior computation is compared against its previous version in Sec. IV-C. For simplicity, we refer to the latter as prior computation (PC) and to the former as improved PC (IPC). The improvements are demonstrated by qualitative comparisons and further reflected in the reduction of the computational time. Finally, in Sec. IV-D the performance of the proposed networks is compared against SCCN [4] and the method in [7], which is from now on referred to as CGAN. Their detailed description can be found in the last paragraph of Sec. II-A2 and Sec. II-B2 respectively. Along with the mean elapsed times, this Section reports the Cohen's Kappa Coefficient  $\kappa$  [38]. The latter is expressed as

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (14)$$

where  $p_o$  is essentially the overall accuracy, i.e. the ratio between the number of data points correctly classified divided

by the total amount of data, and  $p_e$  is the probability of random agreement:

$$p_e = \left( \frac{\text{TP} + \text{FP}}{M} \cdot \frac{\text{FN} + \text{TN}}{M} \right) + \left( \frac{\text{TP} + \text{FN}}{M} \cdot \frac{\text{FP} + \text{TN}}{M} \right) \quad (15)$$

with TP, TN, FP, and FN being respectively the true positives, the true negatives, the false positives, and the false negatives.

The experiments were performed on a machine running Ubuntu 14 with a 8-core CPU @ 2.7 GHz. Moreover, 64 GB of RAM and an NVIDIA GeForce GTX TITAN X (Maxwell) allow to reduce considerably the training times throughout parallel computation. The methods were all implemented in Python using TensorFlow 1.4.0.

##### A. Network settings

1) *X-Net and ACE-Net*: For the design of our frameworks, we opted for CNNs with fully convolutional layers. One of the advantages is their flexibility with respect to the input size. At first, one can use batches of small patches extracted from the original images for the training, but once the latter is over, the banks of filters can be applied directly to the whole dataset at once.

Since the goal is to transform each pixel from one domain to another, there is no interest in having a bottleneck, that is, reducing the size of the input height and width to compress the data. Hence,  $3 \times 3$  filters were applied without stride on the input patches, whose borders were padded with zeros. In the X-Net, both networks have four layers. The first three consist of 100, 50, and 20 filters, the last ones have  $C_2$  and  $C_1$  filters for  $F(\mathbf{X})$  and  $G(\mathbf{Y})$  respectively. Instead, the encoders of the ACE-Net present three layers of 100, 50, and 20 filters, vice versa for the decoders. Finally, the discriminator is the only network in our framework which, after three convolutional layers with 64, 32, and 16 filters, deploys a layer of fully-connected layers with 1 node.

Concerning the activation functions, *Leaky-ReLU* [39] was chosen:

$$\text{Leaky-ReLU}_\beta(x) \begin{cases} x & \text{if } x \geq 0 \\ \beta \cdot x & \text{if } x < 0 \end{cases} \quad (16)$$

where  $\beta$  was set equal to 0.3. The last layer of each network represents an exception: the sigmoid was selected for the discriminator, which must provide outputs between 0 and 1, whereas for every other network the hyperbolic tangent was chosen because our data was normalised between -1 and 1. With this range of data values the training was sped up as expected [40]. Batch normalisation [41] turned out to be unnecessary and was discarded, as it did not improve the optimisation and it actually slowed down our experiments.

After each layer, dropout is applied with a dropout rate of 20% during the training phase to enhance the robustness of the framework against overfitting and input noise [42]. Also, data augmentation helps increasing the size of the training sample by introducing some more variety in the data: before feeding the patches to the network, these were randomly flipped and rotated.

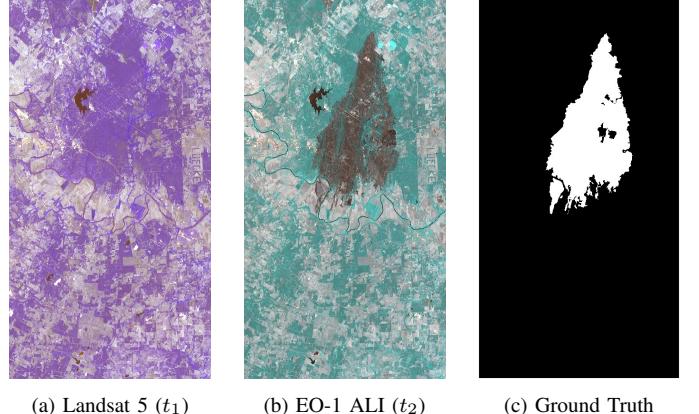
The weights in  $\vartheta$  were initialised with a truncated normal distribution according to [43] and the biases were initialised as zeros. For every epoch of the training 10 batches were used, each containing 10 patches of size  $N = 100$ . The Adam Optimizer [44] minimised the loss function for 240 epochs at a learning rate of  $10^{-5}$ . The weights of the loss function of the ACE-Net are five:  $w_d = 1$  multiplies the losses of Eqs. (12) and (13);  $w_r = 0.2$  scales the effects of the reconstruction errors in Eq. (11);  $w_\vartheta = 0.001$  tunes the regularisation term;  $w_c = 2$  and  $w_\alpha = 3$  emphasise the importance of the cycle-consistency and the prior-weighted similarity. The X-Net presents only three of these, namely  $w_c$ ,  $w_\alpha$  and  $w_\vartheta$ , and they were set in the same way.

2) *SCCN and CGAN*: The most important aspect of the compared architectures is their ability to transform the data and, consequently, the quality of the obtained difference image  $d$ , whereas the image processing applied on the latter should not be relevant for the analysis. Therefore, although [4] and [7] deploy different filtering and thresholding techniques, the methods selected in this work are used on all the difference images for a fair comparison of the final change maps. The implementations of the SCCN and the CGAN were as faithful as possible based on the details shared in [4] and [7]. However, to make the SCCN work we had to replace a fixed parameter described in the paper with the output of Otsu's method to find an optimal threshold for the difference image in the iterative refinement of the change map. We also had to interpret Liu *et al.*, and chose to implement their pre-training phase to avoid trivial solutions with decoders having one coupling layer (convolutional layer with filters of  $1 \times 1$ ) and 250 epochs. This was empirically found to be the minimum amount of epochs needed to consistently obtain a meaningful representation of the data in the code space to be used as starting point for the training procedure. Also, Liu *et al.* selected a rigorous stopping criterion for the latter, but it was hardly reached during our experiments, so a maximum number of epochs was set to 500.

## B. Data sets

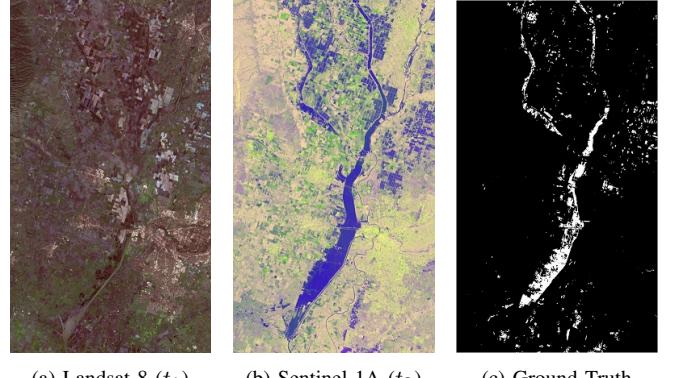
1) *Forest fire in Texas*: Bastrop County in Texas was struck by a forest fire during September–October, 2011. The Landsat 5 TM and the Earth Observing-1 Advanced Land Imager (EO-1 ALI) acquired two multispectral images before and after the event. The resulting co-registered and cropped images of size  $1520 \times 800$  are displayed in false colour in Fig. 6a and Fig. 6b<sup>1</sup>. Both images are optical with 7 and 10 channels, respectively, some of which cover the same spectral bands, so the signatures of the land covers involved are partly similar. Volpi *et al.* [45] provide the ground truth shown in Fig. 6c.

2) *Flood in California*: Fig. 7a represents the RGB channels of a Landsat 8 acquisition<sup>1</sup> covering Sacramento County, Yuba County and Sutter County, California, on 5 January 2017. The OLI and TIRS sensors on Landsat 8 provide a total of 11 channels from optical deep blue to long-wave infrared. The same area was affected by a flood, as it can be noticed in Fig. 7b. This is a Sentinel-1A<sup>2</sup> acquisition,



(a) Landsat 5 ( $t_1$ )      (b) EO-1 ALI ( $t_2$ )      (c) Ground Truth

Fig. 6: Forest fire in Texas: (a) Landsat 5 ( $t_1$ ), (b) EO-1 ALI ( $t_2$ ), (c) ground truth.



(a) Landsat 8 ( $t_1$ )      (b) Sentinel-1A ( $t_2$ )      (c) Ground Truth

Fig. 7: Flood in California: (a) Landsat 8 ( $t_1$ ), (b) Sentinel-1A ( $t_2$ ), (c) ground truth.

recorded in polarisations VV and VH on 18 February 2017 and augmented with the ratio between the two intensities as the third channel. The ground truth in Fig. 7c is provided by Luppino *et al.* [14]. Originally of  $3500 \times 2000$  pixels, these images were re-sampled to  $860 \times 500$  pixels to reduce the computation time.

## C. PC vs IPC

The proposed modifications to the affinity matrix comparison are evaluated by a visual comparison of the results obtained by both the PC and the IPC. Based on [14],  $k = 20$  was selected for all the experiments. Fig. 8 show the outcomes for the two datasets in the two most extreme cases, namely when  $\Delta = 1$  and when  $\Delta = k$ . In the first column, one can notice how the PC provides smoother results with highlighted areas with soft edges. On the contrary, the images in the second column were obtained with the IPC and they arguably represent a more precise result with sharp edges and smaller segments of highlighted pixels. The third column shows the strong impact that a large  $\Delta$  has on the outcomes of [14]. The PC method's assignment of one value to an entire patch leads to the tiled pattern mentioned in Sec. III-A. Instead, the IPC is not as affected by the smaller amount of considered patches, as shown in the fourth column of Fig. 8.

<sup>1</sup>Distributed by LP DAAC, <http://lpdaac.usgs.gov>

<sup>2</sup>Data processed by ESA, <http://www.copernicus.eu/>

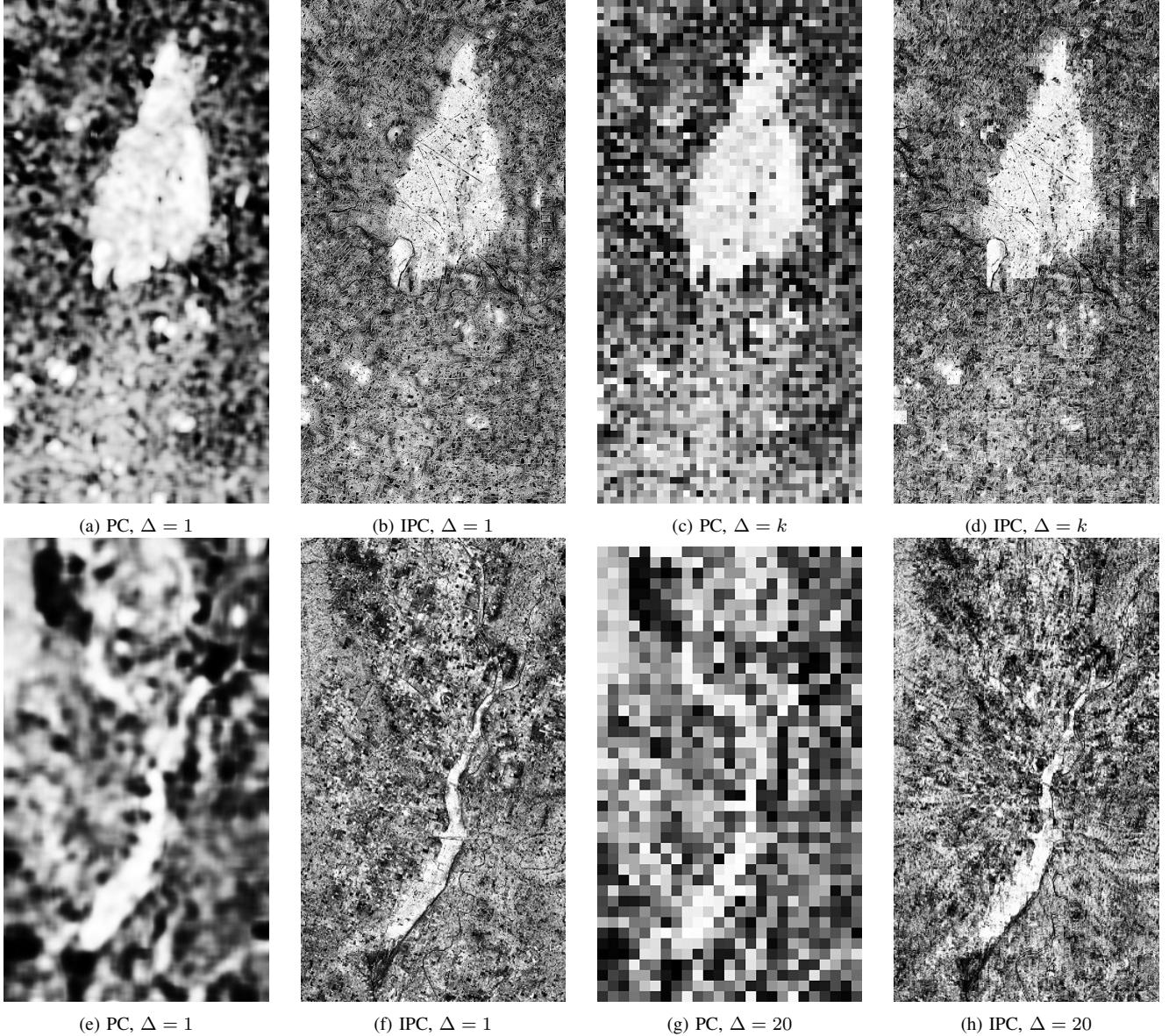


Fig. 8: Results on the two datasets for the PC and the IPC, for  $\Delta = 1$  and for  $\Delta = 20$ .

TABLE I: Approximated  $|\mathcal{P}|$  and computation time of the two methods on the two datasets for  $\Delta = 1$  and  $\Delta = k$ .

	$\Delta = 1$		$\Delta = k$	
	Texas	California	Texas	California
$ \mathcal{P} $	$1.2 \times 10^6$	$4 \times 10^5$	$3 \times 10^3$	$1 \times 10^3$
PC	45 min	15 min	2:37 min	0:37 min
IPC	76 min	24 min	6 min	1:45 min

Table I reports an approximate  $|\mathcal{P}|$  and the computation time spent by the two methods on the two datasets for the two considered cases. As it can be seen, the major drawback of setting  $\Delta = 1$  is the large size of  $\mathcal{P}$ , which increases quadratically with respect to the image dimensions  $H$  and  $W$ . Please recall that we propose to apply the IPC three times:

with  $k_{small} = 10$  and  $k = 20$  to the images at the original sizes, and with  $k = 20$  to the images re-sampled at half the sizes.

Finally, for the training of the ACE-Net and the X-Net we opted for  $k = 20$  and  $\Delta = 5$ , for which the proposed approach took approximately 42min and 13min for the Texas and the California datasets respectively.

#### D. Results

Each of the four architectures was initialised randomly and trained for 100 independent runs, and their metrics are reported in the form of boxplots. These plots represent very clearly the behaviour of  $\kappa$  for the compared methods: a box covers the values from the 25<sup>th</sup> percentile to the 75<sup>th</sup> with an orange line showing the median, while whiskers indicate the span between the 5<sup>th</sup> and the 95<sup>th</sup> percentile. Outliers beyond the whiskers

TABLE II: Average training time of the four methods on the two datasets.

	ACE-Net	X-Net	CGAN	SCCN
Texas	12:52 min	6:46 min	1:09:05 h	15:39 min
California	12:12 min	5:42 min	21:26 min	14:34 min

are marked as dots. As a reference, the  $\kappa$  achieved by directly filtering and thresholding the prior  $\alpha$  is indicated by a red horizontal line.

Table II contains the average times spent to train the four methods on the two datasets. The X-Net is the simplest framework, and this explains its fast training procedure. The ACE-Net and the SCCN have similar complexities, so they require similar times. Instead, the computational load of the CGAN grows quadratically with the sizes of the images, because one epoch of the CGAN is defined as the training of the network using all the  $5 \times 5$  non-overlapping patches. One may suggest to train the networks on a sub-sample randomly picked at every epoch, but there may be a trade-off between speed and performance.

In Fig. 9, the results of the four methods on the Texas dataset are compared. The X-Net and the CGAN show stable and consistent performance. However, only the former achieves better results than the filtered and segmented IPC, which produces  $\kappa = 0.65$ . The ACE-Net and the SCCN sometimes reach higher values of  $\kappa$  than the X-Net, but their median  $\kappa$  is lower and the variation in  $\kappa$  values is high. When compared to the IPC reference, the ACE-Net exceeds its performance in 75% of the test runs, and the SCCN only in 50%.

A different scenario was found for the California dataset, as depicted in Fig. 10. The methods perform similarly and their metrics reach consistently above the reference  $\kappa = 0.2$ , which is the reference value produced by the IPC. The ACE-Net outperforms the X-Net and the CGAN in terms of median  $\kappa$ , but has more variability. The SCCN performs best on this dataset as measured by its  $\kappa$ , which reaches significantly

higher values than the other algorithm, and with a low variability when compared to SCCN behaviour for the Texas dataset. However, upon closer inspection the transformations applied by this method on this dataset are not as intended, and it will be explained in the following section.

Fig. 11 and Fig. 12 show the best examples of output of the four methods on the two datasets. A subset of three channels compose the false colours figures of the original and transformed images. Confusion maps allow to visualise the accuracy of the results: TN are depicted in black, TP in white, FN in red, and FP in green.

## V. DISCUSSION

Stability and consistency are the strong suit of the X-Net and CGAN. They both provide good results on the selected datasets, with the former performing better. The X-Net has other positive aspects, for example the simplicity of its architecture composed of only two CNNs, of few layers each, yielding a total number of  $|\vartheta| \sim 1.3 \times 10^5$  parameters, and the fast convergence of the training thanks to a limited number of terms in the loss function.

The same cannot be said of the CGAN. The framework counts three fully connected networks with  $|\vartheta| \sim 3.1 \times 10^5$ , and taking all the possible  $5 \times 5$  patches as input makes its training epochs time consuming, especially for bigger datasets like the Texas one. In addition, it shows a high tendency to miss some of the changes, due to unwanted alignment of changed areas in the generated and the approximated images. This can be noticed by the high amount of FN in Fig. 11t and Fig. 12t.

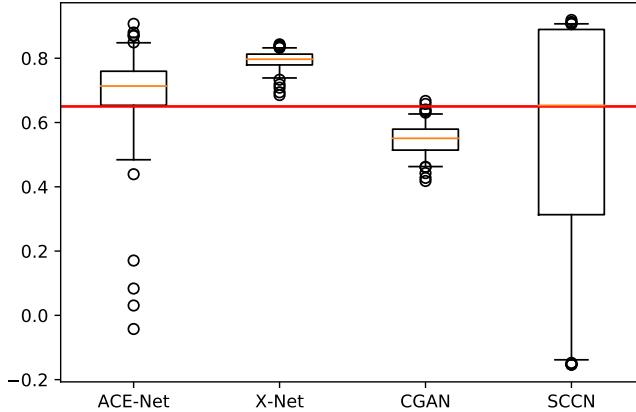


Fig. 9: Boxplots of the  $\kappa$  coefficient for the four methods applied to the Texas dataset. The red horizontal line shows the  $\kappa$  achieved with the affinity matrices comparison.

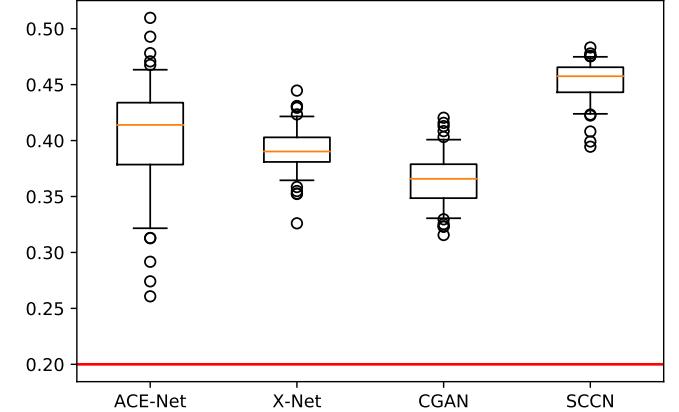


Fig. 10: Boxplots of the  $\kappa$  coefficient for the four methods applied to the California dataset. The red horizontal line shows the  $\kappa$  achieved with the affinity matrices comparison.

The ACE-Net has a large amount of parameters ( $|\vartheta| \sim 2.8 \times 10^5$ ), and together with its complex loss function they guarantee the flexibility that allows to achieve the best performance on the two datasets. However, the complexity is also the main drawback of this architecture, because it implies a difficult and possibly slow convergence, which also results in higher variability in performance. In conclusion, it has the potential to outperform the other methods, but a costly optimisation of its parameters might be necessary.

The SCCN requires a thorough analysis. First of all, this network is very simple: it consists of two symmetric networks with four layers and the total amount of parameters is just  $|\vartheta| \sim 6 \times 10^3$ . Its parameters space is thus limited when compared to its contenders. This may explain why the method very often fails to converge on the first dataset (see Fig. 9). The very good results displayed in Fig. 10 instead are explained by a visual inspection of the image translations it performs on the California dataset. After preliminary training of the two encoders, the one transforming  $\mathbf{Y}$  is frozen, while the other is taught to align the codes of those pixels which are flagged as unchanged. However, it can be seen in Fig. 12e that the encoder is not able to capture more than the background average colour of Fig. 12j, which can be characterized as degenerate behaviour. Basically, the difference image in Fig. 12o is highlighting the water bodies of the SAR image in Fig. 7b, and this coincidentally results in high accuracy when detecting the flood. The same situation was faced when freezing the other encoder. Note that high number of training epochs (500) in our customized implementation of the SCCN was beneficial for the Texas dataset, since it managed to converge more often to a meaningful solution, but it did not make much of a difference on the California dataset, for which the method consistently brings the loss function to a local minimum that corresponds to a degenerate result within the first hundred of epochs, and then not being able to improve it further.

## VI. CONCLUSIONS

In this work we proposed two deep convolutional neural network architectures for heterogeneous change detection: the X-Net and the ACE-Net. In particular, we used an affinity-based change prior learnt from the input data to obtain an unsupervised algorithm. This prior was used to drive the training process of our architectures, and the experimental results proved the effectiveness of our framework. Both outperformed consistently two state-of-the-art methods, and each has its own advantages: the X-Net proved to produce very stable and consistent performance and reliable transformations of the data; the ACE-Net showed to be able to achieve the best results, at the cost of higher complexity and a more diligent training.

## VII. ACKNOWLEDGEMENT

The project and the first author was funded by the Research Council of Norway under research grant no. 251327. We gratefully acknowledge the support of NVIDIA Corporation by the donation of the GPU used for this research.

## REFERENCES

- [1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *International journal of remote sensing*, vol. 10, no. 6, pp. 989–1003, 1989.
- [2] L. T. Luppino, S. N. Anfinsen, G. Moser, R. Jenssen, F. M. Bianchi, S. Serpico, and G. Mercier, "A clustering approach to heterogeneous change detection," in *Scandinavian Conference on Image Analysis*. Springer, 2017, pp. 181–192.
- [3] T. Zhan, M. Gong, J. Liu, and P. Zhang, "Iterative feature mapping network for detecting multiple changes in multi-source remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 38–51, 2018.
- [4] 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 transactions on neural networks and learning systems*, vol. 29, no. 3, pp. 545–559, 2016.
- [5] Z. Liu, G. Li, G. Mercier, Y. He, and Q. Pan, "Change detection in heterogeneous remote sensing images via homogeneous pixel transformation," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1822–1834, 2018.
- [6] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [7] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A conditional adversarial network for change detection in heterogeneous images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 45–49, 2018.
- [8] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5407–5423, 2017.
- [9] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 1, pp. 125–138, 2016.
- [10] M. Gong, H. Yang, and P. Zhang, "Feature learning and change feature classification based on deep learning for ternary change detection in sar images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 129, pp. 212–225, 2017.
- [11] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sensing*, vol. 8, no. 6, p. 506, 2016.
- [12] M. Gong, X. Niu, T. Zhan, and M. Zhang, "A coupling translation network for change detection in heterogeneous images," *International Journal of Remote Sensing*, vol. 40, no. 9, pp. 3647–3672, 2019.
- [13] M. Gong, P. Zhang, L. Su, and J. Liu, "Coupled dictionary learning for change detection from multisource data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7077–7091, 2016.
- [14] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, "Unsupervised image regression for heterogeneous change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.
- [15] T. Zhan, M. Gong, X. Jiang, and S. Li, "Log-based transformation feature learning for change detection in heterogeneous images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1352–1356, 2018.
- [16] R. Touati and M. Mignotte, "An energy-based model encoding nonlocal pairwise pixel interactions for multisensor change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1046–1058, 2018.
- [17] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J.-Y. Tourneret, "A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 799–812, 2015.
- [18] G. Mercier, G. Moser, and S. B. Serpico, "Conditional copulas for change detection in heterogeneous remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1428–1441, May 2008.
- [19] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [20] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

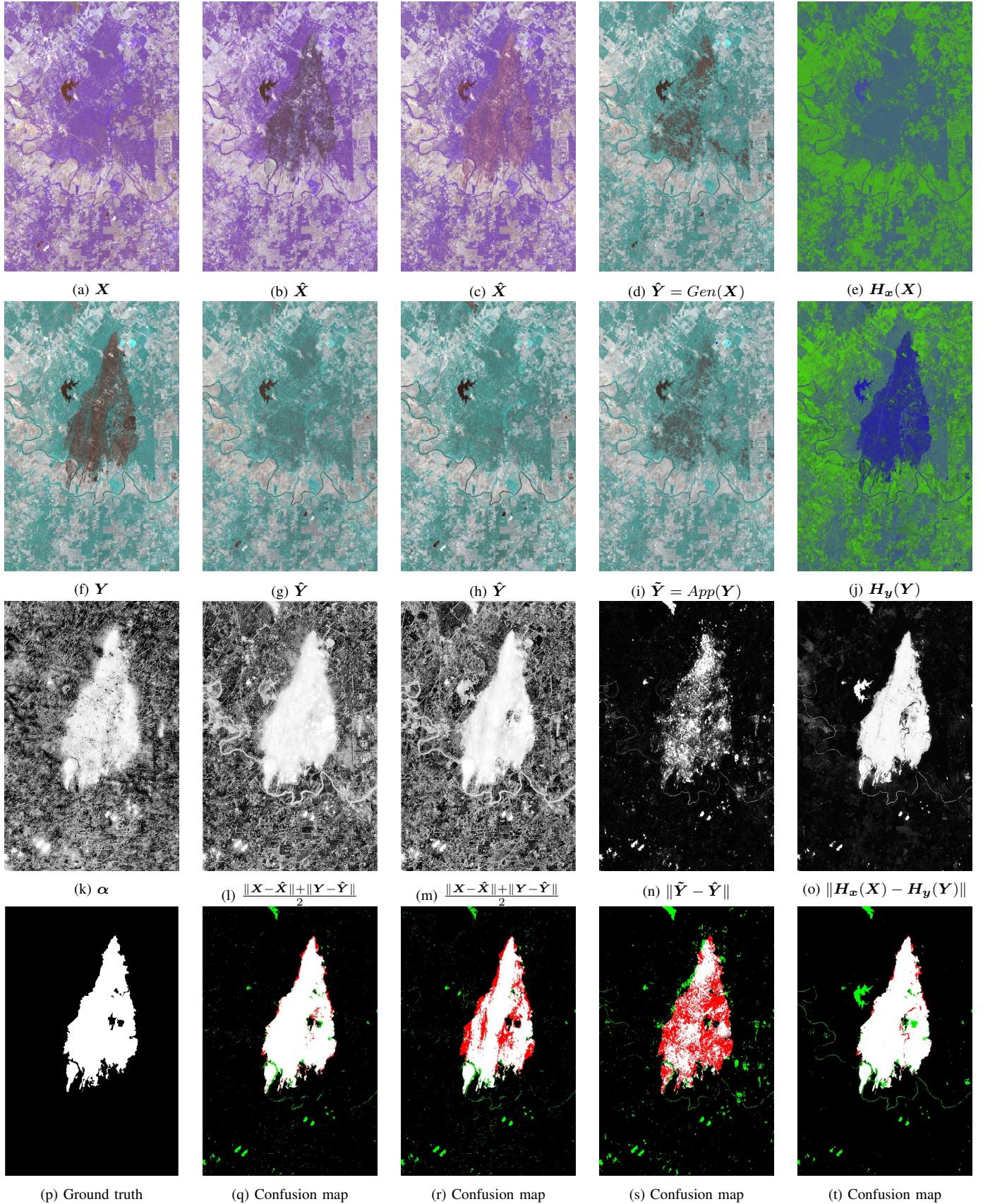


Fig. 11: Texas dataset, columns from left to right: input images  $\mathbf{X}$  (a) and  $\mathbf{Y}$  (f), IPC output  $\alpha$  (k), ground truth (p); ACE-Net; X-Net; CGAN; SCCN.

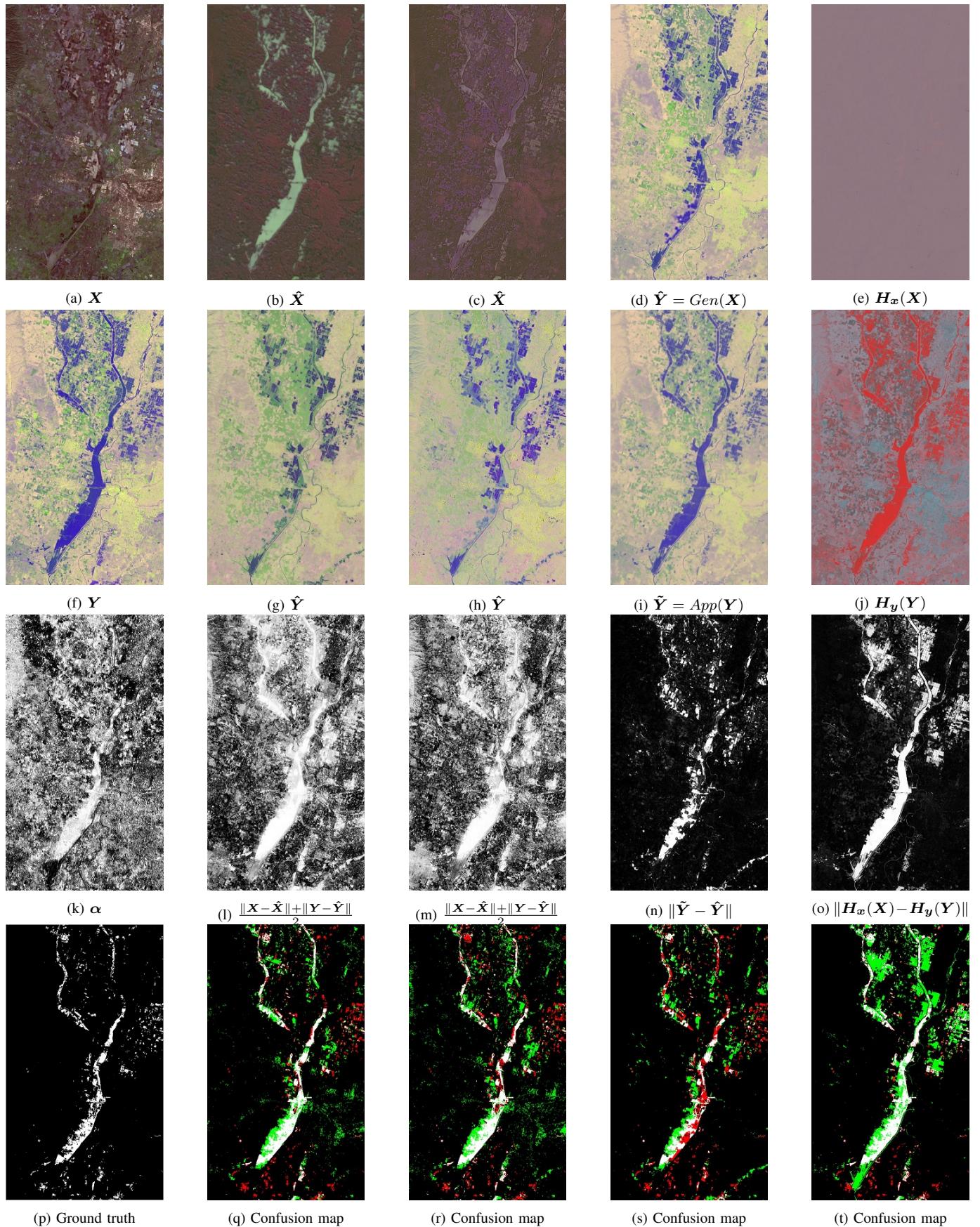


Fig. 12: California dataset, columns from left to right: input images  $\mathbf{X}$  (a) and  $\mathbf{Y}$  (f), IPC output  $\alpha$  (k), ground truth (p); ACE-Net; X-Net; CGAN; SCCN.

- [21] L. Su, M. Gong, P. Zhang, M. Zhang, J. Liu, and H. Yang, "Deep learning and mapping based ternary change detection for information unbalanced images," *Pattern Recognition*, vol. 66, pp. 213–228, 2017.
- [22] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, 2007.
- [23] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 24–41, 06 2016.
- [24] W. Zhao, Z. Wang, M. Gong, and J. Liu, "Discriminative feature learning for unsupervised change detection in heterogeneous images based on a coupled neural network," *IEEE Transactions on Geoscience and Remote Sensing*, 2017.
- [25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [27] N. Merkle, S. Auer, R. Müller, and P. Reinartz, "Exploring the potential of conditional adversarial networks for optical and sar image matching," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1811–1820, 2018.
- [28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [29] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4500–4509.
- [30] F. Bovolo and L. Bruzzone, "The time variable in data fusion: A change detection perspective," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 8–26, 2015.
- [31] J. N. Myhre and R. Jenssen, "Mixture weight influence on kernel entropy component analysis and semi-supervised learning using the lasso," in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*. IEEE, 2012, pp. 1–6.
- [32] M. P. Wand and M. C. Jones, *Kernel Smoothing*, ser. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1995, vol. 60.
- [33] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.
- [34] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [35] J. N. Kapur, P. K. Sahoo, and A. K. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer vision, graphics, and image processing*, vol. 29, no. 3, pp. 273–285, 1985.
- [36] A. G. Shanbhag, "Utilization of information measure as a means of image thresholding," *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 5, pp. 414–419, 1994.
- [37] J.-C. Yen, F.-J. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 370–378, 1995.
- [38] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [39] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.
- [40] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.
- [41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [43] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. PMLR, 13–15 May 2010, pp. 249–256.
- [44] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018.
- [45] M. Volpi, G. Camps-Valls, and D. Tuia, "Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 107, pp. 50–63, 2015.