Semi-supervised Siamese Networks for Change Recognition with VHR Imagery

Matthew J. Gibson, Arcot Sowmya

University of New South Wales

Abstract. Change detection in very high resolution imagery of urban areas is a challenging problem due to the varied environments and high intra-class heterogeneity as well as the expense and paucity of labels. Semi-supervised learning is a family of techniques for leveraging small amounts of labelled data with large amounts of unlabelled data to improve model performance. Recent semi-supervised deep learning approaches have shown that performance can be consistently improved with increasing quantities of data. By extending existing methodologies, this work shows that three specific approaches to semi-supervised learning, namely pseudo-labelling, MixMatch and Virtual Adversarial Training (VAT), are powerful tools to solve change detection problems in remote sensing. In particular, it is shown that convolutional Siamese networks can be fruitfully combined with these semi-supervised methods to achieve better performance. The studied methods are benchmarked on a new aerial urban change detection dataset derived from Sydney suburbia. The proposed method consistently achieves performance comparable to transfer learning with labels available, and directions for future work are also discussed.

Keywords: Change detection · semi-supervised learning · Siamese networks

1 Introduction

In this work, changes to building and urban structure are studied with an aim to distinguish relevant from non-relevant changes. The focus is on two problems: can these changes (adding, removing, or modifying a building or urban infrastructure) be detected, and can the types of these changes be learned (particularly relevant from non-relevant changes). This work proposes a Siamese-network based architecture to learn changes, and the proposed method is applied to two datasets. Siamese networks were originally developed for handwriting recognition [3]. Siamese networks take two inputs that share a common representation and produce a task-specific output which in this case is a classification label. These methods are also natural for change detection.

In this work, the focus is on the problem of change recognition, that is, the classification of pairs of image tiles, rather than pixel level classification. This is a simpler model of change detection, where instead of providing labels at

the pixel level, it is sufficient to provide image or tile level labels. This can be helpful as a precursor to full change detection or may even be sufficient for many object-based change detection tasks.

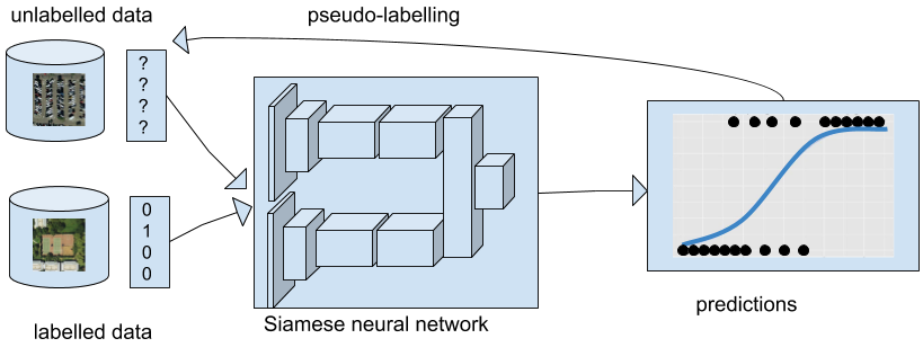


Fig. 1: Overview of the semi-supervised learning system for change detection using pseudo-labelling with Siamese networks. The model works with a mixture of labelled and unlabelled data.

Semi-supervised learning allows a model to learn from a small set of labelled samples by using a larger set of unlabelled samples and assumptions about the distribution of the data. Label information in remote sensing is difficult and expensive to procure and even in the best of times may be unreliable. In the space of learning methods, semi-supervised learning lies between supervised and unsupervised learning, in that limited label information is available at training time [4].

This paper proposes a novel Siamese-network based semi-supervised change recognition method for VHR remotely sensed imagery of urban areas that is trained on a small labelled dataset and a larger unlabelled dataset. The system overview is in Figure 1. The proposed method is tested on two different datasets: aerial imagery over suburban development in Sydney, Australia and a synthetic change detection dataset derived from the EuroSAT Sentinel-2 dataset [9]. It is shown that in limited circumstances, unlabelled data can be utilised effectively for change recognition problems using semi-supervised learning with Siamese networks via the method of Virtual Adversarial Training. This is compared and contrasted with several other state-of-the-art semi-supervised algorithms, and it is concluded that Virtual Adversarial Training performs the best.

1.1 Contribution

The main contributions and novelty of this work are as follows:

1. the application of deep semi-supervised methods to approach the change detection problem is novel within the prior context;

2. we evaluate three contemporary semi-supervised techniques and our experiments use a variety of larger optical datasets at varying levels of resolution;
3. the use of Siamese networks requires the adaptation of semi-supervised learning methods studied in the literature.

2 Related Work

Siamese networks are a well-studied idea developed to allow a neural network to learn similarity metrics between two images. Siamese networks were originally designed for the purpose of comparing hand-written signatures [3]. Since the revival of deep neural networks, Siamese networks have been used in many applications in computer vision including: one-shot learning [10], stereo-vision [19], and patch similarity [18]. Siamese networks are a neural network architecture that has two branches for inputs, in which the weights on the two branches are shared and then the joint embeddings are later fused into a single representation. This capacity for multiple inputs for the network means that Siamese networks can address different types of learning tasks from similarity learning to classification.

Of main interest here is the use of Siamese networks in change detection for high resolution optical imagery, and there have been several studies in this area. An improved triplet loss has been described [21] and the model trialled on very small datasets. In other work [6] and [5], a novel dataset and fully convolutional change Siamese architecture have been presented. Siamese networks have also been proposed to be used in other areas of remote sensing. Siamese networks were used to find the location of ground images in large aerial photographs [13]. The potential of this avenue of research for deep learning applied to multiview image understanding in remote sensing has been described [12].

A large portion of semi-supervised learning for change detection has focused on kernel methods. The earliest example of investigation of this problem [2] using semi-supervised learning is based on Support Vector Machines (SVM) as a component in a completely unsupervised pipeline.

The semi-supervised problem can be defined as follows: given a dataset $\mathcal{X}$ which is composed of a part X with labels Y and the part U which does not possess labels, construct a model F which can use both (X, Y) and U to improve performance. Overfitting is particularly easy when the model in question is a deep neural network [16]. Several methods have been recently proposed to alleviate this problem in the case of semi-supervised learning, including pseudo-labelling [11], MixMatch [1], and VAT [15].

The work presented in this paper may be distinguished from the state of the art surveyed so far. The application of deep semi-supervised methods to the change detection problem is novel within the parameters previously defined. The use of Siamese networks here requires some modest adaptations of semi-supervised learning methods developed in the computer vision literature.

3 Methodology

In order for semi-supervised learning to work, certain assumptions need to be made about the data. These assumptions allow us to to develop a loss function $\mathcal{L}(X, Y, U)$ in which we can incorporate a training signal from the unlabelled data. All the loss functions which we will see in this work are of the form $\mathcal{L}(X, Y, U) = L_X(X, Y) + \alpha L_U(U)$ where $L_X(X, Y)$ is the supervised loss function for the labelled data and $L_U(U)$ is a loss function on the unlabelled data which varies with the semi-supervised learning method. Finally α is a generic weight function that controls the relative weight of the supervised and unsupervised loss.

A Siamese network is a type of deep neural network where learning occurs from two sets of distinct inputs. Formally, two representations $h_A(x_1) = f_1$ and $h_B(x_2) = f_2$ are computed and then fused into a further joint network $h_J(f_1, f_2)$. The model is trained by backpropagating the error once through $h_A(x_1)$ and once through $h_B(x_2)$, so it can also be viewed as a model with two arms that share weights. The Siamese network used in this work is shown in Figure 1. Our network uses a Resnet-18 architecture as a backbone [8].

A key challenge in the study of semi-supervised learning is avoiding overfitting hence much work has focused on novel methods for regularisation. The three methods explored in this work: pseudo-labelling[11], MixMatch [1] and virtual adversarial training (VAT) [15], are now briefly outlined.

3.1 Pseudo-labelling

In pseudo-labelling [11], the ith entry of the pseudo-label $\overline{y}$ is defined to be:

$$\overline{y}_i = \begin{cases} 1, \text{ if } i \text{ is the argmax of } \overline{f_i}(u) \\ 0, \text{else} \end{cases}$$

Given a pseudo-label, a pseudo-label loss function can be defined by:

$$\mathcal{L}_{\text{pseudo}}(X, Y, U, \overline{Y}) = \frac{1}{|X|} \sum_{n=1}^{|X|} \sum_{j=1}^{C} L_X(f_j, y_j) + \alpha \frac{1}{|U|} \sum_{n=1}^{|U|} \sum_{j=1}^{C} L_U(\overline{f_j}, \overline{y}_j)$$

where y_j and $\overline{y}_j$ are the label and pseudo-label respectively, f_j and $\overline{f_j}$ are the output on the labelled and unlabelled data, C is number of classes, α is a balancing coefficient and L_X and L_U are the same cross-entropy loss function. Pseudo-labelling can be used with Siamese networks without modification.

3.2 MixMatch

The method MixMatch [1] is a data augmentation based method which combines the regularisation technique of MixUp [20] with several other semi-supervised methods. The aim of MixUp is to smooth decision boundaries between different classes. By avoiding sharp decision boundaries, the risk of overfitting is expected to decrease. In MixUp two minibatches $B_1 = (X_1, y_1)$ and

$B_2 = (X_2, y_2)$ are combined using λ distributed according to the beta distribution $\beta(\alpha, \alpha)$ as follows:

$X_*, y_* = \textsc{MixUp}(B_1, B_2)$ where:

$$X_* = \lambda X_1 + (1 - \lambda) X_2 \tag{1}$$

$$y_* = \lambda y_1 + (1 - \lambda) y_2 \tag{2}$$

MixMatch then extends MixUp to the semi-supervised setting as follows. Given a minibatch $B = (X_B, y_B, U_B)$, $\textsc{MixMatch}(B)$ returns a batch consisting of augmented labelled data (X_B^*, y_B^*) and augmented unlabelled data with pseudo-labels $(U_B^*, \overline{y}_B^*)$ constructed as follows:

1. Construct i copies of B,
2. Apply randomly chosen data augmentations σ to B_i,
3. Average the predictions B_i,
4. Sharpen predictions,
5. Apply MixUp to blend predicted labels.

It has been observed (by e.g. [1]) that in semi-supervised learning, the model learns better if it is forced to make high confidence predictions so that they are pushed closer to 0 or 1 by using an appropriate sharpening function. One such function can be defined as $\text{sharpen}(p, T) = p_i^{1/T} / \sum_{j=1}^{L} p_j^{1/T}$ where p_k is the prediction and T is another hyperparameter.

The augmentation method used here in steps 1) and 2) is chosen with a view to what is called consistency regularisation. That is given an image I, we produce copies $I_1, I_2, \ldots, I_n$ of I by data augmentation. The predictions for the copies $I_1 I_2, \ldots I_n$ of I should be consistent with that of I, and we can enforce this constraint by averaging the predictions e.g. $p_{\text{avg}} = 1/n \cdot \sum_{i=1}^{n} p_i$. Augmentation is performed on both labelled and unlabelled data.

The semi-supervised loss is then given as

$$\textsc{MixMatch}(X, y, U) = X^*, y^*, U^*, \overline{y}^* \tag{3}$$

$$\mathcal{L}_{\textsc{MixMatch}}(X^*, y^*, U^*, \overline{y}^*) = L_X(X^*, y^*) + \alpha L_U(U^*, \overline{y}^*) \tag{4}$$

where L_X is the cross-entropy loss and L_U is the L_2 loss.

As there is some choice in the manner in which these methods may be applied in the Siamese Network setting, the choices made are outlined. Given a batch $B = (X_1, X_2, y, U_1, U_2)$ where X_1 and U_1 are from time t_1 and X_2 and U_2 from time t_2, we present a simplified version of $\textsc{MixMatch}_{\text{siamese}}$. This adaptation has the following steps:

1. Apply the same randomly chosen data augmentations σ to B,
2. Sharpen predictions,
3. Apply MixUp to blend predicted labels.

The crucial part (according to both our experience and other ablative testing [1]) is step 3. We describe step 3 as follows: a single $\lambda \sim \beta(\alpha, \alpha)$ is drawn, and if X_{a_1}, X_{b_1} are from t_1 and X_{a_2}, X_{b_2} are from t_2, then $X_i^* = \lambda X_{a_i} + (1 - \lambda) X_{b_i}$ where $i \in \{1, 2\}$. That is we blend images across different times.

3.3 VAT

VAT is a form of perturbation-based regularisation [15] and can also be viewed as a form of consistency regularisation. The core idea of VAT is to perturb the model input towards the direction of the worst prediction r_{adv}. This is an adaptation of the idea of Goodfellow et al. [7], to approximate the generation of adversarial examples for semi-supervised learning and adversarial examples can be generated for both labelled and unlabelled data points. The VAT loss for a particular data point can then be calculated as follows:

$$\begin{aligned} \ell_{\text{VAT}}(x_*) &= D(q(y \mid x_*), p(y \mid x_* + r_{\text{adv}})] \\ \text{where } r_{\text{adv}} &= \underset{r;||r||_2<\epsilon}{\arg\max}\, D(q(y \mid x_*), p(y \mid x_* + r)) \end{aligned} \tag{5}$$

and p is the conditional distribution of y given the data x and model parameters, q is the true distribution of the labels, D is the Kullback-Leibler divergence and x_* is either $x \in X$ or $u \in U$

This function ℓ_{VAT} can be then extended to apply to all elements via

$$L_{\text{VAT}}(X, U) = \frac{1}{|X| + |U|} \sum_{x_* \in X,U} \ell_{\text{VAT}}(x_*). \tag{6}$$

This term can then be included into the standard semi-supervised loss function $\mathcal{L}_{\text{VAT}}(X, Y, U) = L_X(X, Y) + \alpha L_{\text{VAT}}(X, U)$. The quantity r_{adv} can be calculated for labelled and unlabelled data points. This has the effect of perturbing the input towards the direction of greatest change to the current predictions. The key to effectively computing L_{VAT} is the approximation of $r_{\text{adv}} = \epsilon \frac{g}{||g||}$ by a small vector r drawn from a normal distribution using $g = \nabla_r d(f(x), f(x + r))$.

This is adapted to the Siamese network setting in this work by using the same adversarial vector r for both x_1 and x_2. That is, since the Siamese model f is a function of two arguments $f(x_1, x_2) = y$, set $g = \nabla_r d(f_\theta(x_1, x_2), f_\theta(x_1 + r, x_2. + r))$. While noting that it could also be sensible to use distinct adversarial vectors r_1 for x_1 and r_2 for x_2, from observation this performs worse in practice.

4 Experiments and results

The first dataset consists of orthomosaics at 30cm ground resolution collected by aerial photography acquired by NearMap over Western Sydney at two distinct times 2018-08-30 and 2013-08-16 [14]. This location in Sydney is on the urban periphery and is the site of active residential development, and the region of interest is illustrated in Figure 2 as well as examples of substantive change. The dataset, which we call "Penrith", is class balanced. It consists of 203 images of size 800×800 pixels with a ground resolution of 15cm. The trained model was evaluated on a 25% prediction dataset. There are 104 images of the positive class and 99 of the negative class. Cosmetic changes such as road markings or temporary structures such as sheds or cars are ignored. The presence of urban infill,

(a) The study area consisting of a 7km^2 region over Western Sydney.

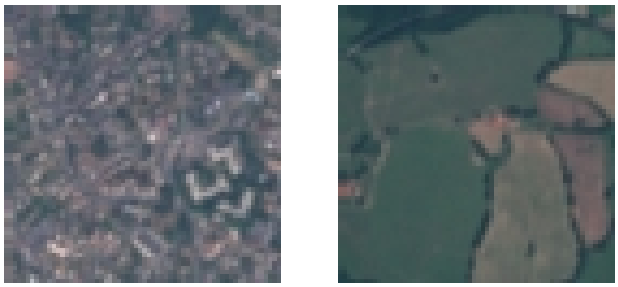
(b) Image tile in 2013.

(c) Image tile in 2018.

Fig. 2: An overview of the study areas as well as example images of a particular location from the Penrith urban expansion dataset.

seasonal variation and lack of photometric calibration makes this a challenging dataset to work with.

The second dataset is a simulation of a change detection dataset based on the EuroSAT dataset. EuroSAT [9] is a dataset comprising 27000 64×64 image tiles. The data is annotated with 10 different land-use types. The model trained in this work is presented with a pair of images and must predict whether they come from the same or different classes. Examples of image pairs whose similarity is to be learned are shown in Figure 3. Although some of these examples are unrealistic (for instance a change from cityscape to river), simulating change detection allows us to compare our results on a change detection task to a comparable land-cover mapping task or image classification task.



(a) Image 1: residential.

(b) Image 2: pasture .

Fig. 3: An example pair of images from the EuroSAT synthetic change detection dataset. This would be labelled 0, since the two land-cover classes differ (residential and pasture).

Transfer learning is seen as an appropriate model for comparison because of the technique's ease of use and its widespread applications. We will explicitly demarcate models which use both transfer learning and SSL. We randomly

partition the data into distinct sets for training and evaluation. In the case of Penrith, we make use of additional data which was part of the dataset but unlabelled. In the case of EuroSAT, we simply omit the labels for the portion of the data which is supposed to be unlabelled.

Training was conducted for 100 epochs using ADAM optimizer with learning rate of $\lambda \in \{0.001, 0.0001\}$, with a minibatch size which varies based on the dataset. Two forms of regularisation were used: weight decay and data augmentation. The data augmentations applied were 90° rotations as well as vertical and horizontal reflections in the image midpoint. The image pairs presented to the transfer-learned models were normalised to channel mean and standard deviations. The source code will be released following completion of the project.

The results of these two experiments which directly compare the different methods are presented in Table 1 and Table 2. For evaluating performance of the semi-supervised algorithms on the change recognition task we use standard metrics namely: accuracy (Acc), precision (Prec), recall (Rec) and the f_1 score.

Table 1: Performance of semi-supervised methods on the synthetic EuroSAT change detection dataset, for $n = 250$ labels, $n = 3375$ test cases, and $n = 18250$ unlabelled data.

Variant	Acc	Prec	Rec	f_1
Transfer learning	0.665	0.651	0.682	0.666
VAT	**0.737**	**0.719**	**0.762**	**0.74**
Pseudo-labelling	0.706	0.683	0.747	0.713
MixMatch	0.703	0.693	0.707	0.7

Table 2: Performance of semi-supervised methods on the Penrith change detection dataset for $n = 747$ labels, $n = 124$ test examples and $n = 2252$ for unlabelled data.

Variant	Acc	Prec	Rec	f_1
Transfer learning	0.790	0.828	0.750	0.787
VAT	**0.815**	**0.860**	0.766	**0.810**
Pseudo-labelling	0.742	0.735	0.781	0.758
MixMatch	0.758	0.750	**0.797**	0.773

5 Discussion

Given how data hungry these deep learning methods are, a key question is how much data do we need to successfully develop a model? Our work explores the

answer to this question by examining the model assumptions we can make to limit the amount of supervised data required.

Overall, the performance improvement due to the use of unlabelled data in this study is not comparable to the improvement seen in the computer vision studies where these methods were proposed. We hypothesise several possible explanations for this phenomenon, including the ratio of unlabelled to labelled data. In the EuroSAT synthetic dataset, this ratio is about 70 : 1, whereas in Penrith this ratio is about 3 : 1 Another important difference between EuroSAT and Penrith is that EuroSAT uses 10m Landsat data whereas Penrith uses 30cm resolution data. It is a well-known finding in the remote sensing literature that different resolution imagery requires different algorithms.

As may be observed, performance on the Penrith dataset was generally comparable to that achieved by transfer learning, and slightly better using VAT, as can be seen in Table 2. The best improvement due to the use of unlabelled samples is achieved on the synthetic EuroSAT change detection dataset. On the whole, though, this contrasts with other applications in the computer vision literature where semi-supervised methods more strongly outperform the transfer learning methods. This difference may be attributed partly to the heterogeneous nature of the elements in the changed class. Problems with replication of experiments and overfitting of models to benchmark datasets are not unknown in the computer vision and machine learning literature (as discussed in [17]).

6 Conclusion

The application of semi-supervised deep learning to change detection is a strand of research that merits further investigation from remote sensing researchers. Semi-supervision and self-supervision continues to be heavily studied in computer vision but making these methods achieve the same level of performance with remote sensing data and on remote sensing tasks is a challenging problem.

References

1. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: MixMatch: A Holistic Approach to Semi-Supervised Learning. arXiv:1905.02249 [cs, stat] (May 2019)
2. Bovolo, F., Bruzzone, L., Marconcini, M.: A Novel Approach to Unsupervised Change Detection Based on a Semisupervised SVM and a Similarity Measure. IEEE Transactions on Geoscience and Remote Sensing **46**(7), 2070–2082 (July 2008)
3. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature Verification using a "Siamese" Time Delay Neural Network. In: Advances in neural information processing systems. p. 8 (1994)
4. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-supervised learning. Adaptive computation and machine learning, MIT Press, Cambridge, Mass (2006)
5. Daudt, R.C., Saux, B.L., Boulch, A.: Fully Convolutional Siamese Networks for Change Detection. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 4063–4067 (October 2018)

6. Daudt, R.C., Saux, B.L., Boulch, A., Gousseau, Y.: High Resolution Semantic Change Detection. arXiv:1810.08452 [cs] (October 2018)
7. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. In: International Conference on Learning Representations (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs] (December 2015)
9. Helber, P., Bischke, B., Dengel, A., Borth, D.: EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (July 2019)
10. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese Neural Networks for One-shot Image Recognition. p. 8 (2015)
11. Lee, D.H.: Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. ICML 2013 Workshop : Challenges in Representation Learning (WREPL) (July 2013)
12. Lefèvre, S., Tuia, D., Wegner, J.D., Produit, T., Nassar, A.S.: Toward Seamless Multiview Scene Analysis From Satellite to Street Level. Proceedings of the IEEE **105**(10), 1884–1899 (October 2017)
13. Lin, T., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5007–5015 (June 2015)
14. Ltd, N.A.P.: Nearmap MapBrowser (2018), https://apps.nearmap.com/maps/
15. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. arXiv:1704.03976 [cs, stat] (June 2018)
16. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 3235–3246. Curran Associates, Inc. (2018)
17. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet Classifiers Generalize to ImageNet? In: International Conference on Machine Learning. pp. 5389–5400. PMLR (May 2019)
18. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4353–4361. IEEE, Boston, MA, USA (June 2015)
19. Zbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1592–1599. IEEE, Boston, MA, USA (June 2015)
20. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond Empirical Risk Minimization. arXiv:1710.09412 [cs, stat] (October 2017)
21. Zhang, M., Xu, G., Chen, K., Yan, M., Sun, X.: Triplet-Based Semantic Relation Learning for Aerial Remote Sensing Image Change Detection. IEEE Geoscience and Remote Sensing Letters pp. 1–5 (2018)