SIAMESE ATTENTION U-NET FOR MULTI-CLASS CHANGE DETECTION

Sol Cummings[1,2], *Lukas Kondmann*[2,3], *Xiao Xiang Zhu*[2,3]
[1]PASCO CORPORATION
[2]Technical University of Munich (TUM)
[3]German Aerospace Center (DLR)

ABSTRACT

Recent developments in deep learning have pushed the capabilities of pixel-wise change detection. This work introduces the winning solution of the DynamicEarthNet Weakly-Supervised Multi-Class Change Detection Challenge held at the EARTHVISION Workshop in CVPR 2021. The proposed approach is a pixel-wise change detection network coined Siamese Attention U-Net that incorporates attention mechanisms in the Siamese U-Net architecture. Moreover, this work finds the location of the attention mechanism within the network is crucial in achieving higher performance. Positioning the attention blocks in the up-sample path of the decoder filters noisy lower resolution features and allows for more fine-grained outputs. The impact of architectural changes, alongside training strategies such as semi-supervised learning are also evaluated on the DynamicEarthNet Challenge dataset.[1]

Index Terms— Change Detection, Siamese Neural Network, Semantic Segmentation

1. INTRODUCTION

Automatic change detection is a crucial task in remote sensing that can aid in analyzing and monitoring the Earth's surface. An increase in the availability of image data due to ever-shortening revisit times and rapid advancements in deep learning have allowed for significant improvements [1] in automatic change detection.

Pixel-wise change detection involves the localization and classification of areas within images that have undergone a change. Previous methodologies have approached the localization and classification independently. [2] focus on the classification of change using Siamese Neural Networks for image patches, where features from multiple inputs are compared within a network after feature extraction using shared weights. However, change detection at a pixel level must rely on a separate localization network.

The U-Net architecture [3] and subsequent improvements [4] exhibit high performance in object localization, specifically semantic segmentation. Additionally, these networks generalize to remote sensing imagery [5], where targets can be small relative to image sizes.

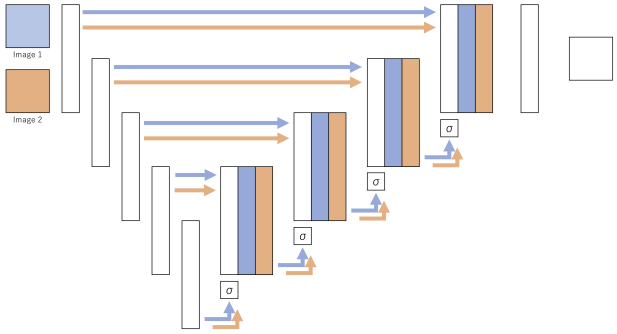

Fig. 1. Structure of the Siamese Attention U-Net. Feature maps of the two inputs are demarcated in different colors. Attention blocks are signified as σ.

More recent works such as [6] propose a Siamese U-Net architecture, where semantic segmentation and change classification is done in an end-to-end manner. This allows for efficient and high performing pixel-wise change detection with a single network.

With ever-increasing amounts of available data, methodologies that make use of unlabeled data have gained traction. Specifically, recent improvements in semi-supervised learning that make use of large unlabeled datasets [7] have improved classification performance on ImageNet.

This work outlines the winning approach out of over 100 submissions in the DynamicEarthNet [8] Weakly-Supervised Multi-Class Change Detection Challenge held at the EARTHVISION Workshop in CVPR 2021, which incorporates improvements to the Siamese U-Net architecture. In particular, the network is adapted to better suit small objects for change detection. Additionally, this work shows that improvements to training strategies such as loss functions and semi-supervised learning can boost performance. Section 2 will outline the dataset, network architecture, and training strategies. Section 3 will discuss results from experiments, and section 4 is the conclusion.

[1]Code is available at: https://github.com/solcummings/earthvision2021-weakly-supervised

2. METHOD

2.1. Dataset

The dataset from the DynamicEarthNet Weakly-Supervised Multi-Class Change Detection Challenge is used for subsequent experiments. This dataset consists of images from Planet and Sentinel-2 satellites alongside monthly pixel-wise ground truths from January 2018 to December 2019 for 75 locations. The ground truths are labelled with changes to impervious surfaces, agriculture, forest/other vegetation, wetlands, soil, water, snow/ice, or no-change surfaces.

The dataset is categorized into three parts: the training dataset with 55 locations, public validation dataset with 10 locations, and test dataset with 10 locations. Among the training dataset locations, only 10 have ground truth labels. 80% of the training dataset is randomly sampled for training and the remaining 20% is used for internal validation.

This work does not use the Sentinel-2 images, and instead solely uses the higher resolution Planet images for accurate semantic segmentation. Due to the large size of the original images and limited GPU VRAM, image chips with side lengths of 128 pixels are cropped from the original image. The image chips are further randomly cropped to side lengths of 112 pixels, randomly flipped vertically and horizontally, and randomly rotated 90 degrees before entering the network during training. The public validation dataset and test datatset are predicted on at original image size, with test time augmentations (TTA) of vertical and horizontal flipping and 90 degree rotations.

2.2. Network Architecture

Siamese U-Net architectures with and without attention blocks are experimented with. The network architecture of the Siamese Attention U-Net is illustrated in Figure 1. Features of the two image inputs are extracted and down-sampled through the encoder of the network, with shared weights. The features are then concatenated and up-sampled through the decoder of the network, similar to [6]. In order to investigate the influence of low resolution features from the decoder on segmentation outputs, bilinear, transposed convolution, and sub-pixel convolution [9] up-sampling methods are explored. Attention blocks from [4] are added to the network to improve overall performance. However, the position of the attention blocks within the network has a significant influence on the network's performance.

In [4], attention blocks are attached to the skip connection paths from the encoder to disambiguate noise from higher resolution features. However, due to the small size of targets in remote sensing semantic segmentation, the higher resolution features from the encoder network often carry more semantically important information, whereas lower resolution features may introduce noise. In order to prioritize higher resolution features and retrieve only relevant information from low resolution features, the attention blocks are embedded in the up-sampled path. The proposed attention block is shown in Figure 2.

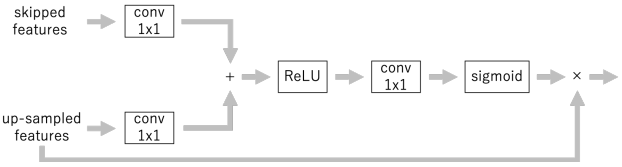

Fig. 2. Structure of the attention block. Unlike [4], the features from the up-sampled path are attended.

2.3. Loss Functions

Evaluation on the public validation dataset is conducted using mean IoU across all classes. Due to the evaluation metric being associated with IoU, Jaccard loss [10] is experimented with. Additionally, Jaccard loss is compared to Dice loss [11], which is a common loss function in semantic segmentation tasks.

2.4. Weakly Supervised Learning

[7] show the effectiveness of semi-supervised learning on classification tasks. However, aspects such as the augmentations used for noisy training cannot be automatically applied to semantic segmentation tasks, let alone remote sensing change detection tasks. Still, experiments in [7] show improvements in performance when using hard pseudo labels. Thus once a network is trained on the training dataset, hard pseudo labels are applied to the public validation and test dataset and a new network is trained on the combined dataset.

2.5. Implementation Details

Networks are trained until there is no update to the internal validation loss, mean F1-score, or mean IoU. These trained networks are then used to predict on the public validation dataset. Experiments are conducted on a single Nvidia RTX2070 with a batch size of 30 in mixed precision. The learning rate is scheduled from 1e-2 to 1e-4 using a cosine annealing scheduler with warm restarts.

3. RESULTS

The impact of up-sampling methods in the decoder of Siamese U-Nets without attention are shown in Table 1. Scores improve when using more sophisticated operations such as the transposed convolutions or sub-pixel convolutions when compared to bilinear up-sampling. Sub-pixel convolutions increase mean IoU scores the most at a small cost in parameter count. Nonetheless, variations in scores display the need

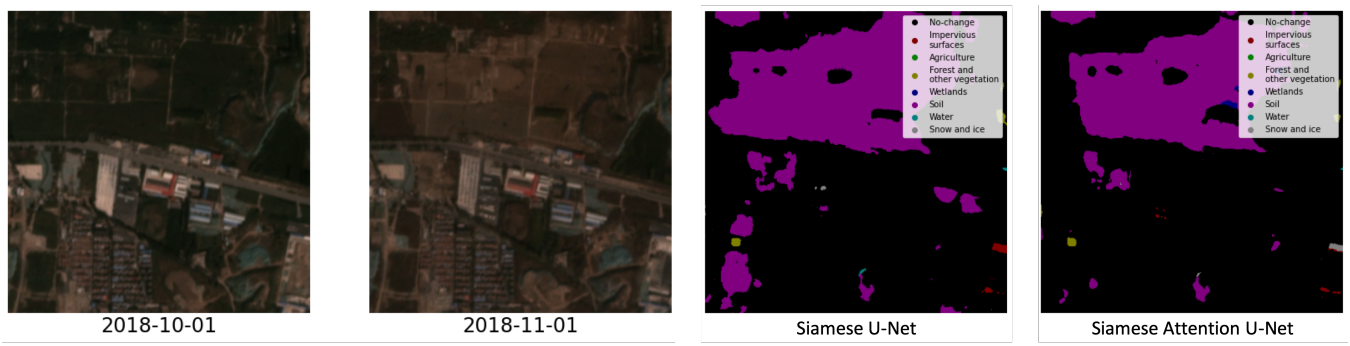

Fig. 3. Comparison of results from the Siamese U-Net and Siamese Attention U-Net on the test dataset

for an exploration of effective up-sampling techniques in the decoder.

Up-sample method	mIoU
bilinear	0.2586
transpose	0.2606
sub-pixel	**0.2623**

Table 1. Mean IoU scores on the public validation dataset of up-sampling methods in Siamese U-Net. Experiments are carried out using Jaccard loss and without TTA.

Experiments on the attention blocks in Siamese U-Nets are shown in Table 2. The network without attention blocks outperforms that of the original attention block configuration proposed in [4], where skip connection features are attended. On the other hand the Siamese Attention U-Net, which attends on up-sampled features, outperforms both. This implies the need for prioritization in up-sampled features instead of skip connection features.

Attention	mIoU
none	0.2635
skip connection	0.2603
up-sample	**0.2658**

Table 2. Mean IoU scores on the public validation dataset of Siamese U-Nets with and without attention blocks. Experiments are carried out using Jaccard loss, sub-pixel convolution up-sampling, and TTA.

A comparison of the results from the Siamese U-Net and the proposed Siamese Attention U-Net on the test dataset is shown in Figure 3. The Siamese Attention U-Net creates more fine-grained segmentation results due to preserving information in high resolution features while reducing noise in low resolution features.

Experiments on the loss function are shown in Table 3. Dice loss and Jaccard loss have comparable performances. While multiple loss objectives are typically combined using a hyperparameter when training a single network, this hyperparameter is often manually determined and requires multiple training runs. This work instead trains multiple networks on each loss function and averages their predictions. Due to the loss functions optimizing for slightly different metrics, ensembling networks trained on each is effective. Additionally, the influence of each loss objective can be controlled with a weighted average when ensembling predictions, as opposed to retraining a single network using a different hyperparameter.

Loss function	mIoU
jaccard	0.2658
dice	0.2668
ensemble	**0.2676**

Table 3. Mean IoU scores on the public validation dataset of loss functions for Siamese Attention U-Net. Experiments are carried out using sub-pixel convolution up-sampling and TTA.

Experiments on semi-supervised learning are shown in Table 4. Pseudo labels on the public validation and test dataset improve the performance on the public validation dataset across loss functions. However, this process requires two consecutive training steps; the first step is regular training on the original dataset, and the second step is training on the original dataset along with hard pseudo labels generated from the first step. Although there is a gain in mean IoU scores, semi-supervised learning using hard pseudo labels is not suitable for time critical applications.

Semi-supervised	Loss function	mIoU (val)
none	jaccard	0.2658
val+test	jaccard	0.2669
none	dice	0.2668
val+test	dice	0.2674
none	ensemble	0.2676
val+test	ensemble	**0.2684**

Table 4. Mean IoU scores on the public validation dataset with or without pseudo labels of the public validation and test dataset for Siamese Attention U-Net. Experiments are carried out using sub-pixel convolution up-sampling, Jaccard loss, and TTA.

4. CONCLUSION

This work examines the effectiveness of attention blocks in the Siamese U-Net. Experiments reveal the originally proposed placement of the attention block may hinder the network, but changing the position of the block improves performance. Siamese Attention U-Net, which computes attention in the up-sample path of the decoder, is able to denoise lower resolution features and allows for fine-grained outputs. This architectural change, alongside enhancements in the loss function and semi-supervised learning boosts mean IoU scores for the DynamicEarthNet Weakly-Supervised Multi-Class Change Detection Challenge.

5. ACKNOWLEDGEMENTS

This work is jointly supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO - Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS".

6. REFERENCES

[1] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch, "Fully convolutional siamese networks for change detection," in *IEEE International Conference on Image Processing*, 2018, pp. 4063–4067. 1

[2] Sergey Zagoruyko and Nikos Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4353–4361. 1

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and computer-Assisted Intervention*, 2015, pp. 234–241. 1

[4] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018. 1, 2, 3

[5] Ryuhei Hamaguchi and Shuhei Hikosaka, "Building detection from satellite imagery using ensemble of size-specific detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 187–191. 1

[6] Vit Ruzicka, Stefano D'Aronco, Jan D Wegner, and Konrad Schindler, "Deep active learning in remote sensing for data efficient change detection," in *Proceedings of MACLEAN: MAChine Learning for EArth Observation Workshop co-located with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. RWTH Aachen University, 2020, vol. 2766. 1, 2

[7] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10687–10698. 1, 2

[8] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Caglar Senaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, and Laura Leal-Taié, "Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[9] Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," 2017. 2

[10] Md Atiqur Rahman and Yang Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *International Symposium on Visual Computing*, 2016, pp. 234–244. 2

[11] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision*, 2016, pp. 565–571. 2