STATS 402 - Interdisciplinary Data Analysis Building Change Detection using FCN Model on Satellite Aerial Images: A Comparative Analysis of 2012 and 2016 Datasets Final Report

1st Jingheng Huan
Duke Kunshan University
Beijing, China
jh730@duke.edu

2nd Luyao Wang
Duke Kunshan University
Suzhou, China
lw337@duke.edu

3rd Fanbin Xu
Duke Kunshan University
Chengdu, China
fx31@duke.edu

***Abstract*—This research presents a solution for detecting changes in urban structures using high-resolution satellite imagery, a pressing issue in the fields of urban development, disaster management, and environmental conservation. The dynamic nature of urban environments necessitates real-time, precise monitoring systems. To address this, we developed a dual-stage model that combines the robustness of Fully Convolutional Networks (FCNs) and the structure of a Siamese network.**

Our model is a synergistic combination of two independent FCN models trained on aerial images from two distinct timeframes: 2012 and 2016. These models specialize in predicting building labels, thus identifying and understanding the structural details in the images. The predictions from these two models are then fed into a Siamese network, which acts as a comparison mechanism, efficiently identifying differences between the two image sets. The output effectively pinpoints structural changes, providing a visual representation of urban evolution.

This research contributes significantly to the field by overcoming the limitations of conventional change detection techniques, which often struggle with high computational requirements and inadequate feature extraction. Our model not only provides a more accurate and efficient solution for monitoring urban development but also serves as a valuable tool in related fields. It paves the way for future research, bringing us one step closer to achieving real-time, automated monitoring of structural changes in urban areas worldwide. Despite its effectiveness, the model revealed certain limitations, particularly in handling complex scenarios where changes are not binary but involve overlapping structures. Future work could focus on refining the model to reduce false negatives and enhance precision, thereby improving its overall performance.

I. Introduction

1. Introduction

The rapid pace of urbanization and the dynamic nature of urban landscapes have necessitated the development of efficient and precise monitoring systems. These systems play a crucial role in tracking changes in building structures, which is a significant aspect of urban development, disaster management, and environmental conservation. The advent of high-resolution satellite imagery has revolutionized this field, providing a wealth of data that can be harnessed for detailed analysis and monitoring. However, the sheer volume and complexity of this data present a formidable challenge, requiring innovative solutions for effective change detection. This research work is motivated by the pressing need to address this challenge and contribute to the global imperative of real-time, automated monitoring of urban change.

The limitations of conventional change detection techniques, such as high computational requirements and inadequate feature extraction, have been a significant hurdle in achieving this goal. Traditional methods often struggle with the task of accurately identifying and understanding the structural details in satellite images. Furthermore, these methods are typically not equipped to handle the dynamic nature of urban environments, which exhibit constant change. This research work aims to overcome these limitations by proposing a novel dual-stage model that leverages the powers of Fully Convolutional Networks (FCNs) and the unique structure of a Siamese network.

FCNs have emerged as a robust tool in the field of image analysis, capable of extracting hierarchical features from raw data. Their unique architecture, which consists solely of convolutional layers, allows them to accept input images of any dimension and generate output feature maps of equivalent spatial dimensions. This makes them particularly suitable for analyzing high-resolution satellite images, which often contain a wealth of detail that needs to be preserved during the analysis process.

The Siamese network, on the other hand, is known for its ability to compare and differentiate between different inputs. In the context of this research, it serves as a comparing mechanism that can efficiently identify differences between two sets of satellite images taken at different timeframes. This ability to pinpoint structural changes over time provides a visual representation of urban evolution, which is invaluable for monitoring urban development.

The proposed model is not merely a singular, standalone system but a synergetic combination of two independent FCN models trained on aerial images from two distinct timeframes: 2012 and 2016. These models specialize in predicting building labels, thus identifying and understanding the structural details in the images. The predictions from these two models are then fed into a Siamese network, which acts as a comparing mechanism, efficiently identifying differences between the two image sets. The resultant output effectively pinpoints structural changes, providing a visual representation of urban evolution.

The implications of this research are broad and significant. Apart from providing a more accurate and efficient solution for monitoring urban development, our model could serve as a valuable tool in related fields such as urban planning, disaster management, and environmental conservation. Our study paves the way for future research, bringing us one step closer to achieving real-time, automated monitoring of structural changes in urban areas worldwide.

II. Related Work and Literature Review

2. Related Work/Literature Review

Change detection in urban environments using high-resolution satellite imagery is a rapidly evolving field, with numerous methodologies and techniques being developed and refined over time. The cornerstone of these techniques is pixel-wise comparison, a method that calculates the disparity between each pixel in one image and the corresponding pixel in the other. This can be achieved through various computational methods, including but not limited to absolute difference, squared difference, and normalized correlation. The resultant difference image serves as a quantitative measure of the similarity or dissimilarity between the two images. However, pixel-wise comparison presents certain drawbacks, namely its computational demand and its limited effectiveness in scenarios involving image modifications such as scaling, rotation, or distortion.

In the quest for more efficient and accurate change detection techniques, researchers have explored the potential of machine learning and deep learning methodologies. One such approach is semantic segmentation, an advanced technique that facilitates the accurate segmentation of each object or region within an image and assigns appropriate labels to every pixel. The implementation of semantic segmentation enables the delineation of each object's boundaries within the image, thereby yielding a more nuanced and comprehensive understanding of the visual content.

To realize semantic segmentation, Fully Convolutional Networks (FCNs) [2] have been widely adopted. FCNs are an innovative tool that autonomously extracts hierarchical features from raw data during the process of change detection. In contrast to conventional convolutional neural networks (CNNs)—which typically comprise a sequence of convolutional and pooling layers succeeded by one or more fully connected layers—FCNs consist does not contain fully connected dense layers. This unique architecture permits FCNs to accept input images of any dimension and generate output feature maps of equivalent spatial dimensions. Despite these advancements, numerous models continue to grapple with challenges related to misalignment and the detection of alterations in smaller structures.

To address these challenges, researchers have proposed various modifications and enhancements to the FCN architecture. One such modification is the introduction of dilated convolutions, which allow the network to incorporate a larger context without increasing the number of parameters or computational complexity. Another approach involves the use of atrous spatial pyramid pooling (ASPP) [3], which captures multi-scale information by applying parallel dilated convolutions with different rates.

In addition to these modifications, researchers have also explored the potential of integrating FCNs with other deep learning architectures. For instance, the combination of FCNs with Long Short-Term Memory (LSTM) [4] networks has been proposed for the task of semantic segmentation in videos. This approach leverages the temporal modeling capabilities of LSTM networks to capture the temporal dependencies between consecutive frames, thereby improving the accuracy of segmentation.

Another innovative approach involves the integration of FCNs with siamese networks [5]. Siamese networks, which consist of two or more identical subnetworks that share the same weights and architecture, have been widely used for tasks involving image comparison, such as face recognition and signature verification. By integrating FCNs with siamese networks, researchers have been able to develop models that are capable of comparing the semantic features of two images, thereby improving the accuracy and efficiency of change detection.

Despite these advancements, there remain several challenges and limitations that need to be addressed. For instance, most existing models struggle to handle complex scenarios where changes are not binary but involve overlapping structures. Moreover, these models often fail to capture the temporal dynamics of urban environments, which can lead to inaccurate or incomplete change detection. To address these challenges, future research could focus on developing more sophisticated models that can handle complex scenarios and capture temporal dynamics. Furthermore, the potential of other deep learning architectures, such as Generative Adversarial Networks (GANs) [6] and attention mechanism [7], could be explored for the task of change detection.

III. The Proposed Methods

A. *Data Preparation*

Data preparation was carried out meticulously for our segmentation model. We subdivided both original satellite images (2012 & 2016) into a set of 1200 satellite images from an open database [1]. Despite their high-resolution nature, to make them computationally manageable, we downscaled the images to 256x256 pixels, ensuring that crucial details were preserved. The images underwent normalization to maintain data consistency across the model. This step involved scaling the pixel intensities so that they fell within a predefined range. We further improved our dataset by applying data augmentation techniques, such as flipping and rotation, to enhance its diversity. This strategy aids in making the model more versatile in handling different image orientations and perspectives. After this, the data was partitioned into training and testing sets, serving as the foundation for the subsequent model training and evaluation phases.

B. Semantic Segmentation

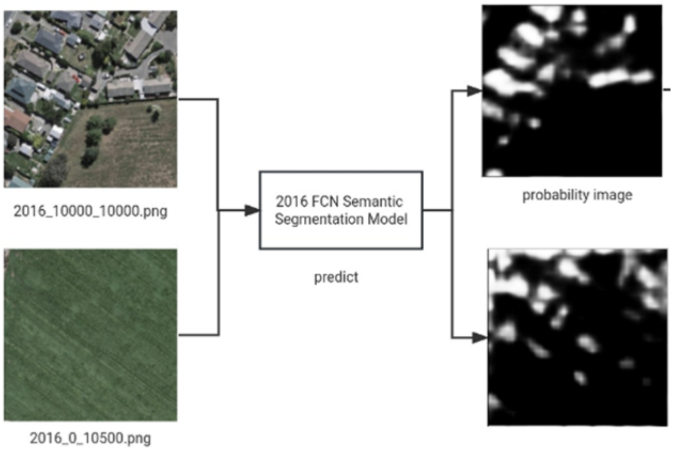


Fig. 1. Using 2016 FCN semantic segmentation model to predict probability images.

Semantic segmentation, a key application of Fully Convolutional Networks (FCNs), plays a pivotal role in our project, particularly in the analysis of satellite imagery. We employ the FCN-ResNet50 model, a built-in feature of the PyTorch library under torchvision.models.segmentation, to perform this task. The FCN-ResNet50 model is a deep learning architecture that combines the strengths of ResNet50, a robust model for image classification, with the spatial preservation capabilities of FCNs, making it ideal for semantic segmentation. Prior to input into the FCN model, images undergo a normalization process to ensure pixel intensity values fall within a standard range, thereby facilitating more efficient learning and improving model performance.

The output from the semantic segmentation stage, segmented images, are then fed into a siamese network for further processing. These segmented regions serve as critical features that the siamese network leverages to distinguish between similar and dissimilar images. The siamese network, a type of neural network architecture known for its ability to learn discriminative feature spaces, uses a pair of identical subnetworks to compare different inputs. By integrating the FCN-ResNet50 model for semantic segmentation with the siamese network for comparison, we've created a powerful pipeline that significantly enhances the overall performance of our model, leading to more accurate and reliable results in image analysis.

C. Baseline Model

Our project utilized a baseline model comprising a Siamese network and a custom image dataset, both of which were instrumental in achieving our objectives. The Siamese network,

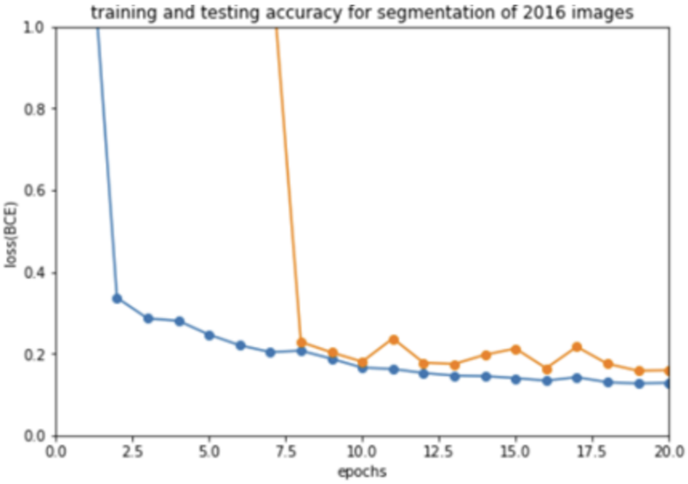


Fig. 2. Binary Cross-Entropy Loss of training and testing accuracy for segmentation on 2016 images.

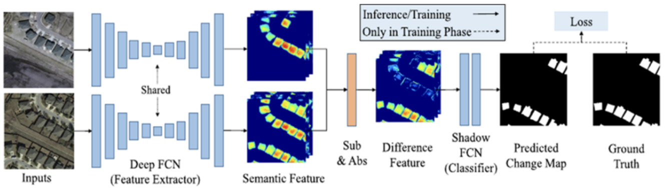
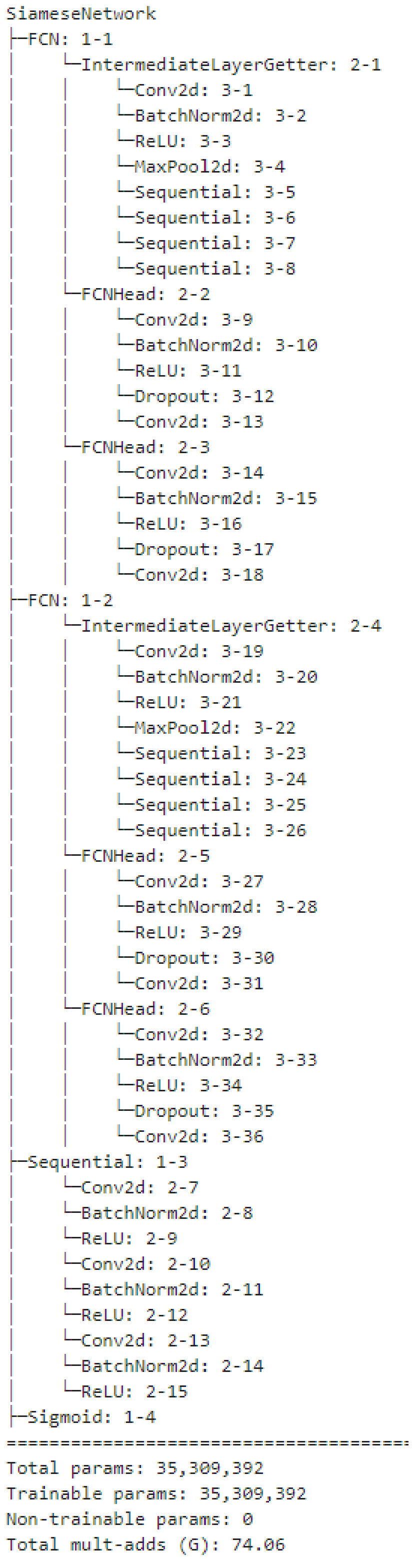


Fig. 3. Baseline model by using FCN siamese network [6].

a key component of the baseline model, was employed to extract distinctive features and differentiate between similar and dissimilar pairs of images. This was accomplished through its dual FCN structure. As illustrated in the corresponding figure, the output from the Siamese FCN feature extractor is subsequently fed into another FCN classifier, which then predicts the change map. The preprocessing steps for the change detection dataset encompasses normalization, flipping, and rotation, all of which serves as data augmentation techniques to ensure the robustness of the model. The baseline model was subjected to extensive training over multiple epochs, with the batch size and learning rate meticulously calibrated to optimize our computational resources. During the training phase, the baseline model was provided with two images for comparison as input, and the predicted change map was compared with the ground truth. Regrettably, this baseline model did not yield significant learning, as evidenced by a training binary cross-entropy loss of approximately 0.66. This suboptimal result indicated that our baseline model was unable to extract meaningful features from the pair of satellite images. Consequently, we modified our approach by using the predicted change map from the preceding FCN segmentation model as input, and then predicting the change map based on this input.

It gives a better result than the previous one which is directly training on satellite images.

This baseline model served as a critical benchmark for



```
SiameseNetwork
├─FCN: 1-1
│    └─IntermediateLayerGetter: 2-1
│    │    └─Conv2d: 3-1
│    │    └─BatchNorm2d: 3-2
│    │    └─ReLU: 3-3
│    │    └─MaxPool2d: 3-4
│    │    └─Sequential: 3-5
│    │    └─Sequential: 3-6
│    │    └─Sequential: 3-7
│    │    └─Sequential: 3-8
│    └─FCNHead: 2-2
│    │    └─Conv2d: 3-9
│    │    └─BatchNorm2d: 3-10
│    │    └─ReLU: 3-11
│    │    └─Dropout: 3-12
│    │    └─Conv2d: 3-13
│    └─FCNHead: 2-3
│    │    └─Conv2d: 3-14
│    │    └─BatchNorm2d: 3-15
│    │    └─ReLU: 3-16
│    │    └─Dropout: 3-17
│    │    └─Conv2d: 3-18
├─FCN: 1-2
│    └─IntermediateLayerGetter: 2-4
│    │    └─Conv2d: 3-19
│    │    └─BatchNorm2d: 3-20
│    │    └─ReLU: 3-21
│    │    └─MaxPool2d: 3-22
│    │    └─Sequential: 3-23
│    │    └─Sequential: 3-24
│    │    └─Sequential: 3-25
│    │    └─Sequential: 3-26
│    └─FCNHead: 2-5
│    │    └─Conv2d: 3-27
│    │    └─BatchNorm2d: 3-28
│    │    └─ReLU: 3-29
│    │    └─Dropout: 3-30
│    │    └─Conv2d: 3-31
│    └─FCNHead: 2-6
│    │    └─Conv2d: 3-32
│    │    └─BatchNorm2d: 3-33
│    │    └─ReLU: 3-34
│    │    └─Dropout: 3-35
│    │    └─Conv2d: 3-36
├─Sequential: 1-3
│    └─Conv2d: 2-7
│    └─BatchNorm2d: 2-8
│    └─ReLU: 2-9
│    └─Conv2d: 2-10
│    └─BatchNorm2d: 2-11
│    └─ReLU: 2-12
│    └─Conv2d: 2-13
│    └─BatchNorm2d: 2-14
│    └─ReLU: 2-15
├─Sigmoid: 1-4
======================================
Total params: 35,309,392
Trainable params: 35,309,392
Non-trainable params: 0
Total mult-adds (G): 74.06
```

Fig. 4. Structure of the siamese model for change detection

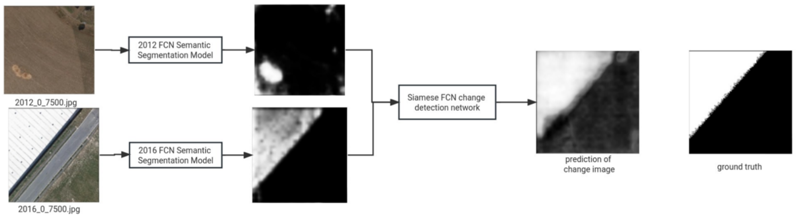

2012_0_7500.jpg
2012 FCN Semantic Segmentation Model
2016_0_7500.jpg
2016 FCN Semantic Segmentation Model
Siamese FCN change detection network
prediction of change image
ground truth

Fig. 5. FCN change detection siamese network on FCN semantic segmentation models.

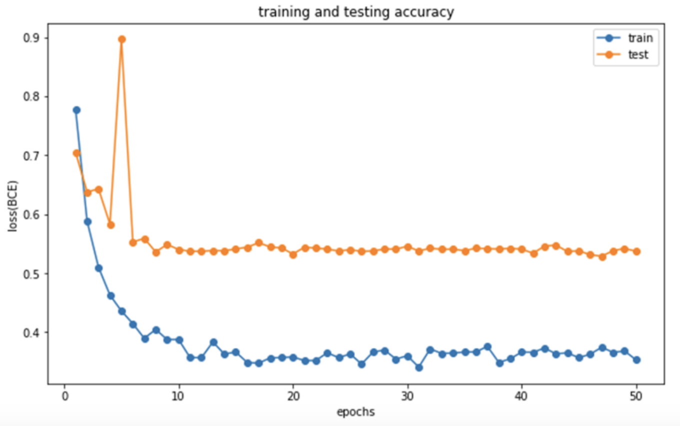


Fig. 6. Binary Cross-Entropy Loss of training directly on satellite images from baseline model.

our project, providing a standard against which we could measure the performance improvements offered by our proposed method. In essence, the baseline model set the stage for comparison and validation, ensuring our model's superiority was empirically grounded.

D. *Refined Model*

Several factors could have contributed to the suboptimal performance of the baseline model. One potential issue lies in the use of interpolation as the upsampling layer in the PyTorch built-in model. Interpolation, while useful for resizing images, may not provide sufficient additional information for the model to learn from. In contrast, transposed convolution, also known as deconvolution, is often a more effective choice for upsampling in deep learning models as it can learn to upsample in a way that is optimized for the model's specific task. In comparison, previous research [8] has demonstrated the effectiveness of a similar FCN siamese network for change detection. This model introduces two key improvements over our baseline model. The first is the use of transposed convolution layers, as discussed earlier. The second improvement is the incorporation of skip connections, a technique that allows the network to learn more complex and diverse patterns from the data. Skip connections also reduce the number of parameters and operations required by the network, making it more computationally efficient. Moreover, they facilitate the training of deeper networks, which are capable of capturing a broader range of features from the data.

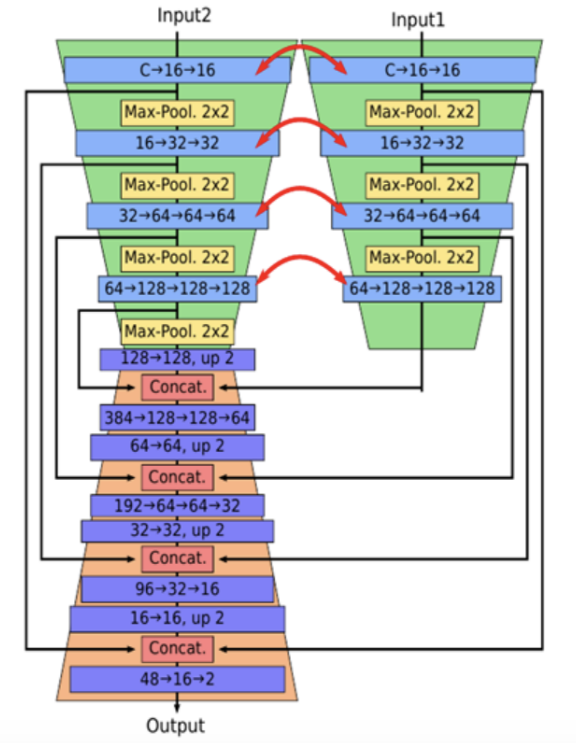


Fig. 7. Schematic of the proposed architecture for change detection. Block color legend: blue is convolution, yellow is max pooling, red is concatenation, purple is transpose convolution. Red arrows illustrate shared weights [8].

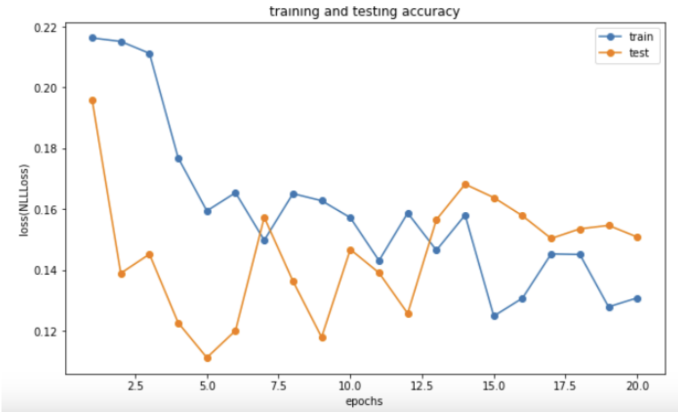


Fig. 8. Binary Cross-Entropy Loss of training and testing accuracy from refined model.

In light of these insights, we adjusted our approach and trained the improved model directly on the satellite images, bypassing the semantic segmentation stage used previously. This modification resulted in a model that not only achieved higher accuracy but also outperformed the previous baseline model. This demonstrates the value of iterative model development and the importance of incorporating insights from prior research to enhance model performance.

IV. Performance Evaluation

	Train	Test
Accuracy	0.9502	0.9874
Precision	0.7113	0.5634
Recall	0.7763	0.7962
F1	0.7424	0.6599

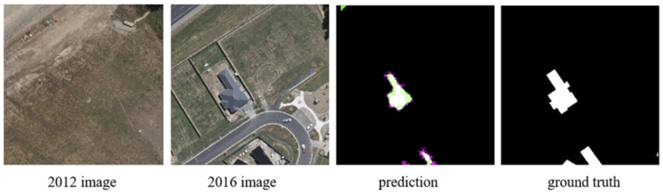


Fig. 9. The accurate prediction outcome from the refined model compared with ground truth image.

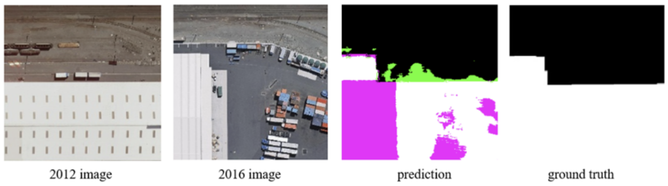


Fig. 10. Prediction outcome containing false negatives from the refined model compared with ground truth image.

In the evaluation of the refined Siamese Fully Convolutional Network (FCN) model, we have utilized two distinct figures to illustrate the model's predictive capabilities. The color-coding scheme employed in these figures is as follows: white represents true positives, black signifies true negatives, green indicates false positives, and magenta denotes false negatives.

Upon examination of the first figure, it is evident that the model demonstrates a commendable ability to identify changes between two images. This is a testament to the model's robustness and its capacity to discern variations effectively. However, the second figure reveals a higher incidence of false negatives. A notable instance of this is observed in the bottom left corner of the image, where the model incorrectly assumes no change.

This discrepancy can be attributed to the inherent limitations of the ground truth labeling. For instance, the comparison of the two satellite images from 2012 and 2016 reveals that while the buildings in the two images are not identical, there is a degree of overlap. The binary labeling system, which categorizes the situation as either 'change' or 'no change', fails to capture this nuance. This highlights the need for a more sophisticated labeling system that can accommodate such complexities.

Despite these challenges, the refined Siamese FCN model exhibits a high degree of proficiency in change detection. This is further substantiated by the performance metrics obtained from the training and testing data.

For the training data, the model achieved an accuracy of 0.9502, indicating that it correctly identified 95.02% of the instances. The precision of 0.7113 suggests that of all the instances the model predicted as 'change', 71.13% were indeed changes. The recall value of 0.7763 implies that the model was able to correctly identify 77.63% of all actual changes. The F-1 score, a harmonic mean of precision and recall, was 0.7424, reflecting a balanced performance of the model.

In the case of the testing data, the model exhibited an impressive accuracy of 0.9874. The precision was slightly lower at 0.5634, indicating a higher rate of false positives. However, the recall value was robust at 0.7962, signifying that the model was successful in identifying a high proportion of actual changes. The F-1 score was 0.6599, demonstrating a reasonable balance between precision and recall.

There exist potential avenues for enhancement in the current model, particularly in the selection of the loss function. The prevailing choice of Cross Entropy may not be the most optimal for tasks such as change detection. Alternative loss functions, such as Contrastive Loss or Triplet Loss, could potentially yield superior results.

Contrastive Loss and Triplet Loss are distance-based loss functions that measure the similarity between images. These functions operate on the principle of taking a pair of samples as input, which are either similar or dissimilar. The fundamental objective of these loss functions is to minimize the distance between similar samples and maximize the distance between dissimilar samples in the feature space.

These alternative loss functions could potentially enhance the model's ability to discern changes by effectively capturing the similarity and dissimilarity between images, thereby improving the overall performance of the model in change detection tasks.

In conclusion, the refined siamese FCN model exhibits a high degree of effectiveness in change detection, as evidenced by its performance metrics. However, there is room for improvement, particularly in reducing the incidence of false negatives and enhancing the precision of the model. Future work could explore the potential of a better loss function or a more nuanced labeling system to better capture the complexities inherent in change detection.

V. Conclusion

In conclusion, this research presents a significant advancement in the field of change detection within urban environments, utilizing high-resolution satellite imagery. The study introduces a novel dual-stage model that synergistically combines the robust capabilities of Fully Convolutional Networks (FCNs) and the unique structure of a Siamese network. This innovative approach has been designed to address the pressing need for real-time, accurate monitoring of urban development, a requirement that has become increasingly critical in the face of rapid urbanization and environmental changes.

The model developed in this research has demonstrated a high degree of effectiveness in detecting and identifying changes in building structures over time. The performance metrics obtained from the training and testing data provide empirical evidence of the model's proficiency. The accuracy, precision, recall, and F-1 scores all indicate a strong performance, showcasing the model's ability to correctly identify a significant proportion of actual changes and to accurately predict 'change' instances.

However, the research also revealed certain limitations in the model's performance. Specifically, the model encountered challenges in handling complex scenarios where changes were not strictly binary but involved overlapping structures. This highlighted the need for a more nuanced labeling system that can better capture the complexities inherent in change detection. The incidence of false negatives and the need for enhanced precision were identified as areas for improvement.

The implications of this research are broad and far-reaching. The model developed has potential applications in a variety of fields, including urban planning, disaster management, and environmental conservation. By providing a more accurate and efficient solution for monitoring urban development, the model could serve as a valuable tool in these areas, facilitating informed decision-making and effective resource allocation.

Future work could focus on refining the model to address its identified limitations. Efforts could be directed towards reducing the incidence of false negatives and enhancing the model's precision. Additionally, the development of a more sophisticated labeling system could be explored to better accommodate the complexities of change detection.

In essence, this research represents a significant step forward in the field of change detection using satellite imagery. It not only provides a robust model for detecting changes in urban structures but also paves the way for future research in this area. The findings of this study bring us one step closer to achieving real-time, automated monitoring of structural changes in urban areas worldwide, contributing to the broader goal of creating smart, sustainable cities.

References

[1] "-OpenDataLab-," Opendatalab.com, 2023. https://opendatalab.com/Building_change_detection_dataset

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," arXiv.org, 2014. https://arxiv.org/abs/1411.4038.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," arXiv.org, 2016. https://arxiv.org/abs/1606.00915.

[4] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM Fully Convolutional Networks for Time Series Classification," arXiv.org, 2017. https://arxiv.org/abs/1709.05206

[5] S. Goel, "Change Detection using Siamese Networks - Towards Data Science," Medium, Jul. 21, 2020. https://towardsdatascience.com/change-detection-using-siamese-networks-fc2935fff82 (accessed May 17, 2023).

[6] H. Chen, W. Li and Z. Shi, "Adversarial Instance Augmentation for Building Change Detection in Remote Sensing Images," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022, Art no. 5603216, doi: 10.1109/TGRS.2021.3066802.

[7] H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," vol. 12, no. 10, pp. 1662–1662, May 2020, doi: https://doi.org/10.3390/rs12101662.

[8] R. C. Daudt, Saux, Bertrand Le, and A. Boulch, “Fully Convolutional Siamese Networks for Change Detection,” arXiv.org, 2018. https://arxiv.org/abs/1810.08462.