# Forecasting Urban Development from Satellite Images

**NANDO METZGER**[1]

*Abstract: Forecasting where and when new buildings will emerge is a rather unexplored niche topic, but relevant in disciplines such as urban planning, agriculture, resource management, and even autonomous flight. In this work, we present a method that accomplishes this task using satellite images and a custom neural network training procedure. In stage A, a DeepLapv3+ backbone is pretrained through a Siamese network architecture aimed at solving a building change detection task. In stage B, we transfer the backbone into a change forecasting model that relies solely on the initial input image. We also transfer the backbone into a forecasting model predicting the correct time range of the future change. For our experiments, we use the SpaceNet7 dataset with 960 km² spatial extension and 24 monthly frames. We found that our training strategy consistently outperforms the traditional pretraining on the ImageNet dataset. Especially with longer forecasting ranges of 24 months, we observe F1 scores of 24% instead of 16%. Furthermore, we found that our method performed well in forecasting the times of future building constructions. Hereby, the strengths of our custom pretraining become especially apparent when we increase the difficulty of the task by predicting finer time windows.*

## 1 Introduction

Understanding the evolution of land use (e.g., constructions and change of land type) is crucial in fields like urban planning, agriculture, natural resource management, anticipating housing market prices, and even autonomous driving and flying. One of the critical components in purely vision-based positioning systems is having an up-to-date map of the environment. Especially in emergencies, it is crucial to have a map with minimal uncertainty to find safe landing sites. These uncertainties are reducible by regularly conducting resource-intensive survey flights. However, a system that can anticipate the changes could assist in indicating the locations of future survey flights and potentially save unnecessary flights in regions with no change.

While existing approaches aim to forecast Land use/Land cover class changes, none explicitly targets building footprint as the output variable. Moreover, there appears to be a lack of research using advanced deep learning methods that aim to forecast the time of when the changes will occur. In this work, we aim to forecast where and when changes in building footprints are going to happen. We consider the problem of change forecasting as a binary segmentation problem with labels "change" and "no change", whereas the forecasting range of the model is already implicitly defined by the selection of the training data (i.e., using a consistent forecasting range in the training samples). We use satellite data series as the primary input data source because they provide global, uniform coverage.

We develop a 2-stage training strategy that is centered around the state-of-the-art segmentation network DeepLabv3+ (CHEN et al. 2018). The stages are as follows:

[1] ETH Zürich, Professorship for Photogrammetry and Remote Sensing
E-Mail: nando.metzger@geod.baug.ethz.ch

 DOI: 10.24407/KXP:179604993X

- **Stage A**: We train a Siamese change detection network on the task of change detection (See Fig. 1 for the task overview).
- **Stage B**: We use the pretrained backbone of stage A, to build two variants of forecasting models. The first model variant deals with the task of change forecasting, the other with the task of time range forecasting task (see Fig. 1 for the task overview).

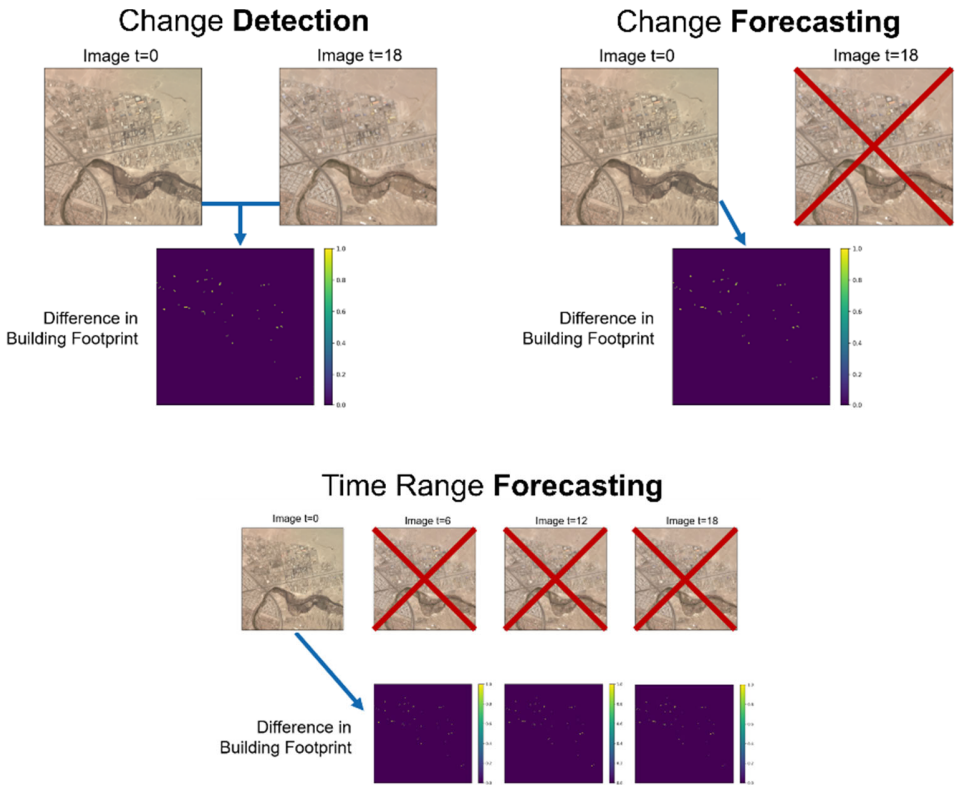


Fig. 1: Visual overview of tasks. The red crosses stress the unavailability of a variable in each task.

## 2 Methodology

### 2.1 Data

To validate our method, we make use of the SpaceNet7 dataset. It is published by VAN ETTEN 2021 and created for the building tracking competition featured at the NeurIPS 2020. The dataset consists of 60 locations, each containing a series of 24 Planet Labs satellite image mosaics of 4×4 km at timestamps that are one month apart. The ground sampling distance of the images is 4 m, and the total covered area of this dataset is 960 km$^2$. The dataset also contains a set of human-labeled building footprints, where each image of the time series is labeled individually.
For this work, we derive a dataset that considers image pairs from the same locations but at different time steps and calculate the changes in the corresponding building footprints. Note, that for simplicity reasons, we neglect the cases where buildings are removed. Using this method, we obtain about 11,000 unique image pairs. If we divide them into training patches of 224×224 pixels,

we get 237,000 mutually exclusive pairs. However, for the task of change forecasting, one implicitly defines the forecasting range by the choice of the training sub-datasets with a consistent forecasting range. For example, when we want to forecast for the smallest possible range (i.e., one month), the subset of pairs that are one month apart contains 5,000 samples, whereas, with the largest forecasting range of 24 months, we can only obtain 200 samples. The dataset also exhibits a severe label imbalance. The average percentage of positive pixels in the whole dataset amounts to 0.3%, and only every seventh patch has more than 0.5% positive labels.

## 2.2 Experimental Setup

We allocate the locations in the dataset into a training set with 42 locations (70%), a validation set for parameter tuning with 6 locations (10%), and a test set to calculate the final metrics with 12 locations (20%). As our evaluation metric, we use the optimal F1 score. The optimal F1 score has appealing theoretical properties, but it uses knowledge about the true labels to classify the predictions. Therefore, we also implement our adaptive threshold by a moving average over the last 500 optimal training thresholds. We found that the discrepancies between the optimal F1 score and the F1 score from our adaptive threshold are diminishingly small, such that we only report the optimal F1 score.

## 2.3 General Idea

The task of change forecasting is naturally ill-posed. One satellite image alone does not define for which future time range the binary classification of "change"/ "no change" should be predicted. It is therefore important that the model is trained on a subset of pairs with a fixed temporal difference, which implicitly defines the forecasting range. However, using these small subsets, one cannot exploit the full potential of the data. To still use all samples, we propose to first train the components of the model to perform the change detection task, where the time range is explicitly defined by the two input images of each pair. We call this pretraining procedure stage A, whereas our main task of change forecasting is called stage B (Fig. 2).

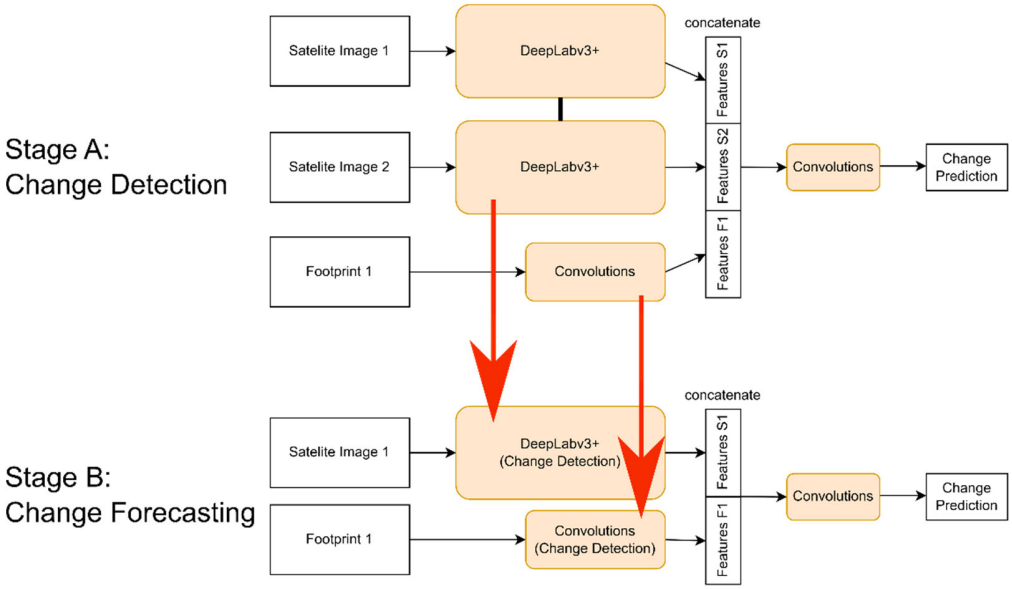


Fig. 2: Modular model architectures and general workflow.

### 2.4 Stage A: Change Detection

In stage A, we train a Siamese change detection network with the DeepLabv3+ backbone on the task of change detection using ordered pairs of satellite images. On the one hand, we can use all the available training pairs in this stage. On the other hand, this type of pretraining already allows the adaption to the satellite images domain, in contrast to the traditional pretraining that is usually performed on the ImageNet domain. We use the building footprints belonging to the first image as an additional feature and process it using a convolutional neural network. Please refer to Fig. 2 for further detail about the architecture. As a loss function, we use the binary cross-entropy loss.

### 2.5 Stage B: Forecasting

We extract the components of the change detection model of stage A and use them as initializations for two versions of forecasting models:

- The **change forecasting model** predicts a binary change label and is trained using the binary cross-entropy loss. We train one classifier for each of the sub-datasets with forecasting ranges {1,3,6,9,12,15,18,21,24}, which helps us better understand the strengths and weaknesses of our approach.
- The **time range forecasting model** is set up in a multitasking learning approach, where the model needs to predict the time of the change on an ordinal scale (*n* time windows) and the binary change label. We fix the maximal range to *24* months, and in a first set up, we split it into two prediction windows – e-g, *n=2*, such that the two ordinal labels are defined as *1-12*, and *13-24* months respectively. Furthermore, we also test a setup of three labels (*n=3*) where the labels are defined as *1-8*, *9-16*, and *17-24* months. The output dimension of the model is *n+1*, where the last output represents the "no change"-class. Finally, we train the model using a multi-class- and a binary cross-entropy loss for time range forecasting and change forecasting, respectively.

In both cases, we first freeze the backbone and footprint convolutions and train the remaining model for 5,000 batch iterations. We then reduce the learning rate by a factor of 10 and train the entire model.

## 3 Results and Discussion

### 3.1 Change Forecasting

In the comparison of Fig. 3 (left), we notice a consistent improvement by using our pretraining strategy compared to using the traditional ImageNet-pretraining strategy. Most notably, the model experiences a much more significant absolute increase in performance of +8% F1 score for the forecasting range of 24 months. The baseline models are only trained on the sub-datasets that correspond to their forecasting range, while our training strategy implicitly allows the models to access the information contained in the whole dataset through the pretraining of stage A. Therefore, it makes sense that the performance gap gets more significant as the forecasting range progresses because that is also when the availability of change forecasting samples is scarce.

Furthermore, we provide the precision-recall curve in Fig. 3 (right) for the classifier that is assigned to the 24 months prediction range. It shows what tradeoffs may be considered when implementing a real-world system.
Fig. 4 helps to get an understanding of which visual cues the model is using to make predictions. It is apparent that the models mainly go for construction sites and cluttered or empty spaces in the vicinity of already existing buildings.

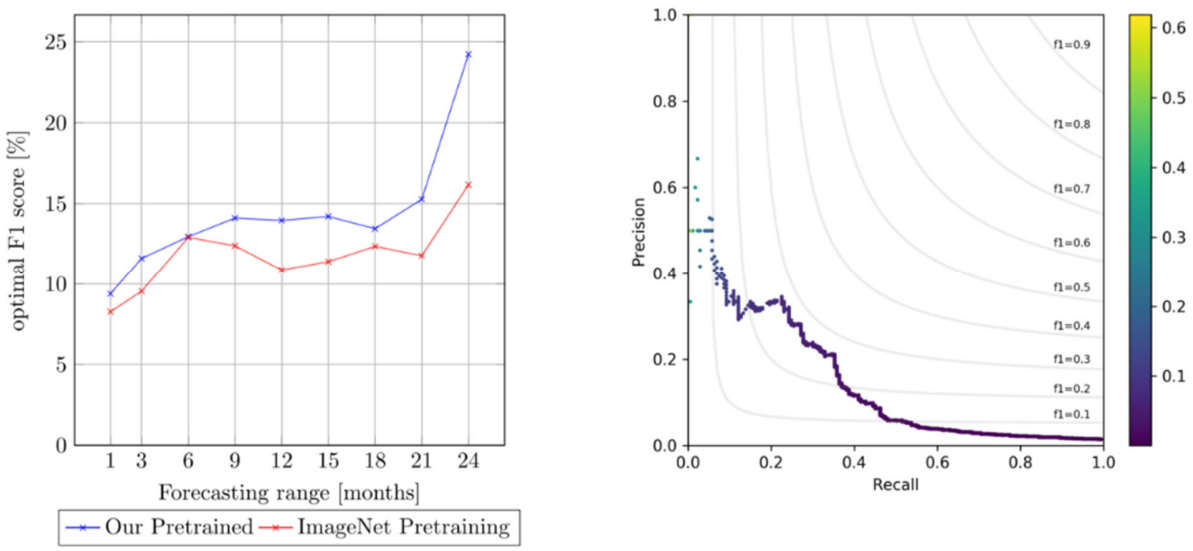


Fig. 3: Left: Performance of our pretraining strategy compared to ImageNet. Each dot denotes a different model that is specialized in the corresponding time range; Right: The precision-recall curve of the model that predicts the 24 months window. The colors of the dots resemble the corresponding thresholds

## 3.2 Time Range Forecasting

The resulting confusion matrices and F1 scores for the time range forecasting tasks with two and three target labels are presented in Tab. 1 and

Tab. 2, respectively. The results attest that all classifiers exhibit a noticeable signal on the diagonals of their confusion matrices. Our pretraining method outperforms the baseline method in terms of F1 score in both tasks, whereas the lead grows significantly for the three-label setting, which is the more challenging task. Moreover, we note that the class 9-16 months in the three-label setting creates the most considerable confusion. After all, it seems that the pretraining is effective enough to learn an operational classification.

Tab. 1: Setup with two labels (1-12 months, and 13-24 months).
The rows are the ground truth classes, while the columns are the predicted classes.

| | **Our Pretraining** (F1= 68.4%) | | | **ImageNet Pretraining** (F1= 67.6%) | |
|---|---|---|---|---|---|
| | ***(1-12)m*** | ***(13-24)m*** | | ***(1-12)m*** | ***(13-24)m*** |
| ***(1-12)m*** | 44k | 28k | | 39k | 33k |
| ***(12-24)m*** | 19k | 60k | | 14k | 65k |

Tab. 2: Setup with three labels (1-8 months, 9-16 months, and 17-24 months).
The rows are the ground truth classes, while the columns are the predicted classes.

| | **Our Pretraining** (F1= 48.2%) | | | | **ImageNet Pretraining** (F1= 42.2%) | | |
|---|---|---|---|---|---|---|---|
| | ***(1-8)m*** | ***(9-16)m*** | ***(17-24)m*** | | ***(1-8)m*** | ***(9-16)m*** | ***(17-24)m*** |
| ***(1-8)m*** | 25k | 10k | 18k | | 26k | 16k | 13k |
| ***(9-16)m*** | 4k | 13k | 22k | | 11k | 9k | 21k |
| ***(17-24)m*** | 2k | 16k | 32k | | 6k | 16k | 29k |

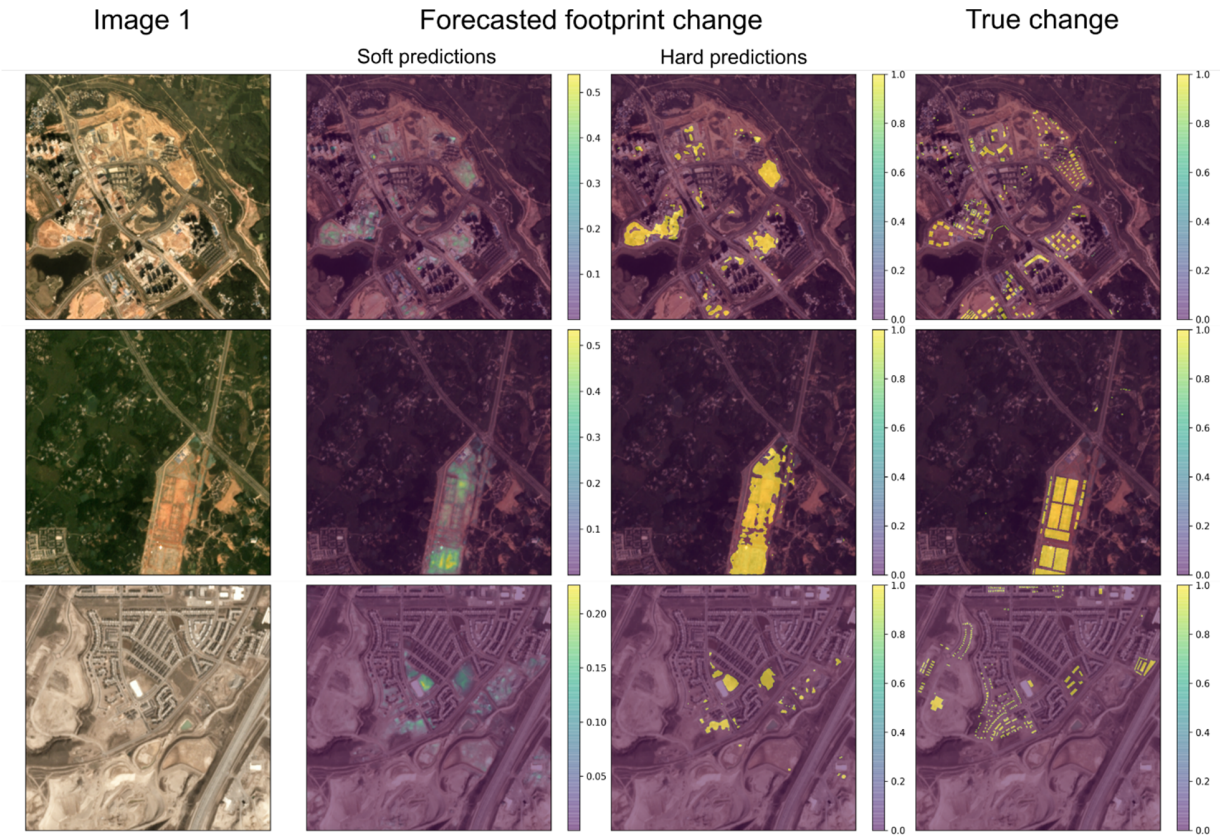


Fig. 4: Visual samples of the 18 months change forecasting task. We overlay the predictions and ground truth labels with their corresponding input image for better interpretability.

## 4 Conclusion and Outlook

The main contribution of this work is a model to forecast where and when changes in building footprints will happen. In this niche, there seems to be a scarcity of research. Furthermore, we develop a 2-stage transfer learning procedure that is superior to the traditional applied ImageNet pretraining and validate it on the task of change forecasting and time range forecasting. In stage A of our methodology, we pretrain a Siamese DeepLabv3+ on the task of change detection. In stage B, we use the pretrained components to tackle the task of change forecasting and time range forecasting. In all tasks, our method performs consistently better than traditional pretraining on ImageNet. In particular, for change forecasting, the method proved to be the most efficient for the

longer forecasting ranges - i.e., for the 24-month range, it achieves an F1 score of 24%, while the baseline achieves only 16%. In the range forecasting task, we found the largest performance increases for the more challenging case where three time ranges need to be predicted.
We argue that the tasks at hand with these particular settings are very challenging. On the one hand, very few visual cues give away a potential future change in land use. On the other hand, building change detection and forecasting tasks inherently lead to a significant imbalance in the label distribution, as new buildings are built very slowly and infrequently.
In our future work, we want to exploit the temporal dimension of the data more efficiently. We assume that time series analysis, such as convolutional recurrent neural networks (RNNs), can yield a model that recognizes spatial and temporal patterns indicating future changes. Moreover, the usage of higher resolution imagery would reveal more valuable cues to solve the task.

## 5 Bibliography

CHEN, L. C., ZHU, Y., PAPANDREOU, G., SCHROFF, F. & ADAM, H., 2018: Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV), 801-818, https://arxiv.org/abs/1802.02611v3.

VAN ETTEN, A., HOGAN, D., MANSO, J. M., SHERMEYER, J., WEIR, N. & LEWIS, R., 2021: The Multi-Temporal Urban Development SpaceNet Dataset. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6398-6407.