A. Annotation Guidelines

For the manual annotation process, we provided annotators with a detailed set of guidelines to ensure consistency in labeling persistent changes. Annotators were instructed to focus on long-lasting changes that persisted for over a year, disregarding short-term or seasonal variations. They were given examples of desertification, urban expansion, and forest loss, and asked to ignore temporary changes such as crop rotations or seasonal foliage variations. If either annotator encountered uncertainties during the labeling process, we would review the time series, discuss our observations, and reach a consensus. We found that inter-annotator agreement was high, particularly in cases of clear, persistent changes. Furthermore, as the dataset is fairly small, we randomly sampled and double-checked each of the annotated time series.

Initially, we manually annotated 300 images of size 512x512 pixels with binary labels indicating whether persistent changes were present (1) or not (0). For images labeled as 0 (no change), we made the assumption that changes were uniformly absent throughout the image. As a result, these images were split into 16 smaller patches, each of size 128x128 pixels, and all patches were automatically labeled 0. For images labeled as 1 (indicating changes), all 16 patches were individually annotated to capture the finer details of the changes across smaller regions.

In addition to the manual annotation process, annotators had access to longitude and latitude information for each image, along with integrated Google Maps and OpenStreetMap views within the annotation interface (shown in Figure 5). This integration allowed them to cross-reference geographical context, improving their ability to identify persistent changes. For example, if annotators were unsure about changes in an image, they could determine that the area was in Australia and recognize that fires between certain months may have affected the region. This spatial context greatly improved annotation accuracy for geographically complex or ambiguous cases.

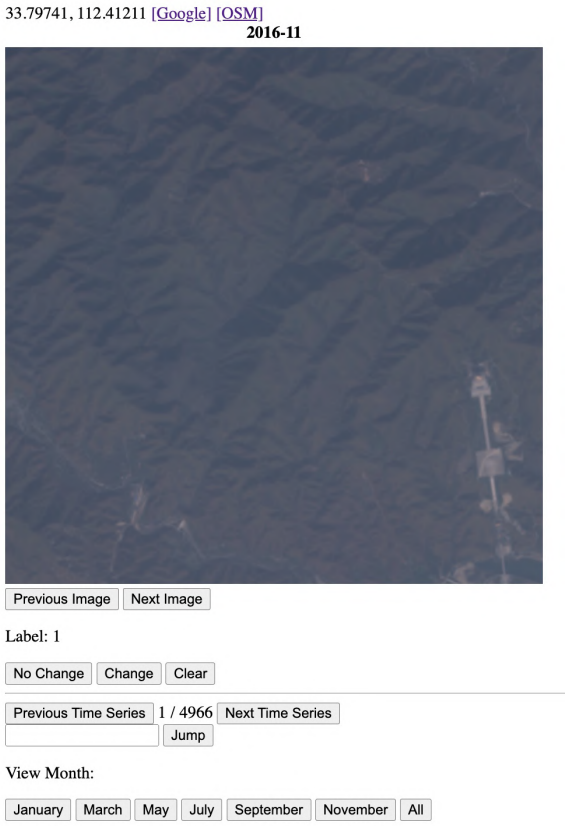

Figure 5. The annotation interface used during the manual labeling process. Annotators could navigate through time series images, view specific months, and classify changes using a set of keyboard shortcuts. Furthermore, they could adjust the number of images they wanted to view at a time (i.e., 3 at a time instead of 1).

B. Qualitative Examples and Model Comparison

The qualitative analysis of OPTIMUS, CaCo, and SeCo reveals important distinctions in how these models handle persistent and cyclic changes across diverse environments. OPTIMUS performs exceptionally well in detecting long-term, persistent changes by leveraging the temporal progression of images, allowing it to filter out short-term, cyclic variations like crop rotations or seasonal shifts in snow cover. As shown in Figure 6, OPTIMUS is effective in identifying clear, non-reversible transformations, making it suitable for both urban and natural environments where such variations dominate.

However, some failure cases highlight challenges for OPTIMUS in distinguishing between significant and subtle environmental shifts. In particular, regions like deserts or areas with strong lighting changes or surface texture shifts (e.g., sand dunes, shadows) can mislead the model, causing it to detect changes where none actually exist. These failure cases suggest that while temporal progression helps filter out short-term cyclic variations, certain natural phenomena—such as reflective surfaces or temporary shadows—can still trigger high change scores in OPTIMUS (see Figure 7). Notably, SeCo and CaCo exhibit similar patterns in these situations, as reflected in their segmentation masks.

CaCo also performs relatively well in rural environments due to its seasonally invariant representations. However, in our evaluation of both SeCo and CaCo, change scores are calculated by dividing the detected change by the total number of pixels in the image. This pixel-wise normalization makes them more effective in urban areas, where changes like new buildings or roads are concentrated and well-defined. In urban settings, larger and clearer change

maps make the pixel-wise division less problematic. However, in rural environments where changes are smaller and more dispersed, this approach dilutes the change signal, making both SeCo and CaCo less sensitive to subtle transformations.

Many failure cases, as shown in Figure 7, occur in environments where multiple change signals overlap—such as urban expansion combined with cyclic agricultural changes. In these cases, the models must distinguish between permanent, meaningful changes and temporary cyclic phenomena. OPTIMUS generally performs better while SeCo has a tendency to output black change maps, indicating no detected change when the signal is weak or dispersed.

C. Ablations

For all ablations, we report the area under the ROC curve (AUROC), which is threshold-independent. Additionally, we include the optimal F1 scores across all thresholds.

C.1. Backbones

We evaluated various backbones to determine their impact on performance. The backbones tested include Resnet-50, Resnet-152, and Swinbase-v2 [16]. For these tests, we used Satlas [1] weights and a context size of three. Due to the availability of Satlas weights, we were limited to these three backbones.

Table 3. Testing different encoder backbones

Backbone	F1 Score	AUROC
Resnet-50	**0.760**	**0.876**
Resnet-152	0.730	0.850
SwinBase-v2	0.750	0.865

The AUROC performance was comparable across all tested backbones, suggesting that the core strength of our approach lies in the method of using change scores rather than the specific backbone used. Therefore, we opted for Resnet-50 due to its efficiency.

C.2. Weight initializations

We assess the effect of different weight initializations on the Resnet-50 backbone with a context size of three. We tested three initialization methods: Random, ImageNet [6] [1], and Satlas weights, as shown in Table 4.

Table 4. Testing different initializations

Initialization	F1 Score	AUROC
Random	0.723	0.851
Imagenet	0.746	0.857
Satlas	**0.760**	**0.876**

We acknowledge that this is not a comprehensive ablation analysis, as random and Imagenet weights were never ideal for the specific task. The primary aim was to demonstrate that using Satlas weights, which are pre-trained on satellite images, improves performance compared to Random and Imagenet weights.

C.3. Context sizes

For an ablation study, we evaluate the impact of context size on the performance of OPTIMUS, as described in Section 4, by varying context size from one to five. For each configuration, OPTIMUS is retrained on the entire dataset to adjust for the new input size.

Table 5. Testing different context sizes

Context size	F1 Score	AUROC
1	0.696	0.809
2	0.745	0.850
3	**0.760**	**0.877**
4	0.739	0.850
5	0.689	0.817

Table 5 presents the results of varying the context size. A context size of three was optimal, which is the size used for all other results in this paper. This is likely because three images provide a balance between robustness and variance in the temporal context given to the model. Performance decreases with context fewer than three due to reduced robustness to outliers. Conversely, performance drops with context more than three due to increased variability within the set, which complicates predictions.

C.4. Change Measures

For an ablation study, we evaluate using two different change measures, pivot score and Spearman coefficient. All of these were done on the Resnet-50 backbone, with Satlas weights, and with context size three.

Table 6. Testing different change measures

Measure	F1 Score	AUROC
Spearman	0.748	0.860
Pivot	**0.760**	**0.877**

Pivot scores were slightly more effective than the Spearman coefficient in detecting progressive changes that aligned with human annotations. This may be because humans are better at identifying abrupt changes, which the pivot score captures more effectively.

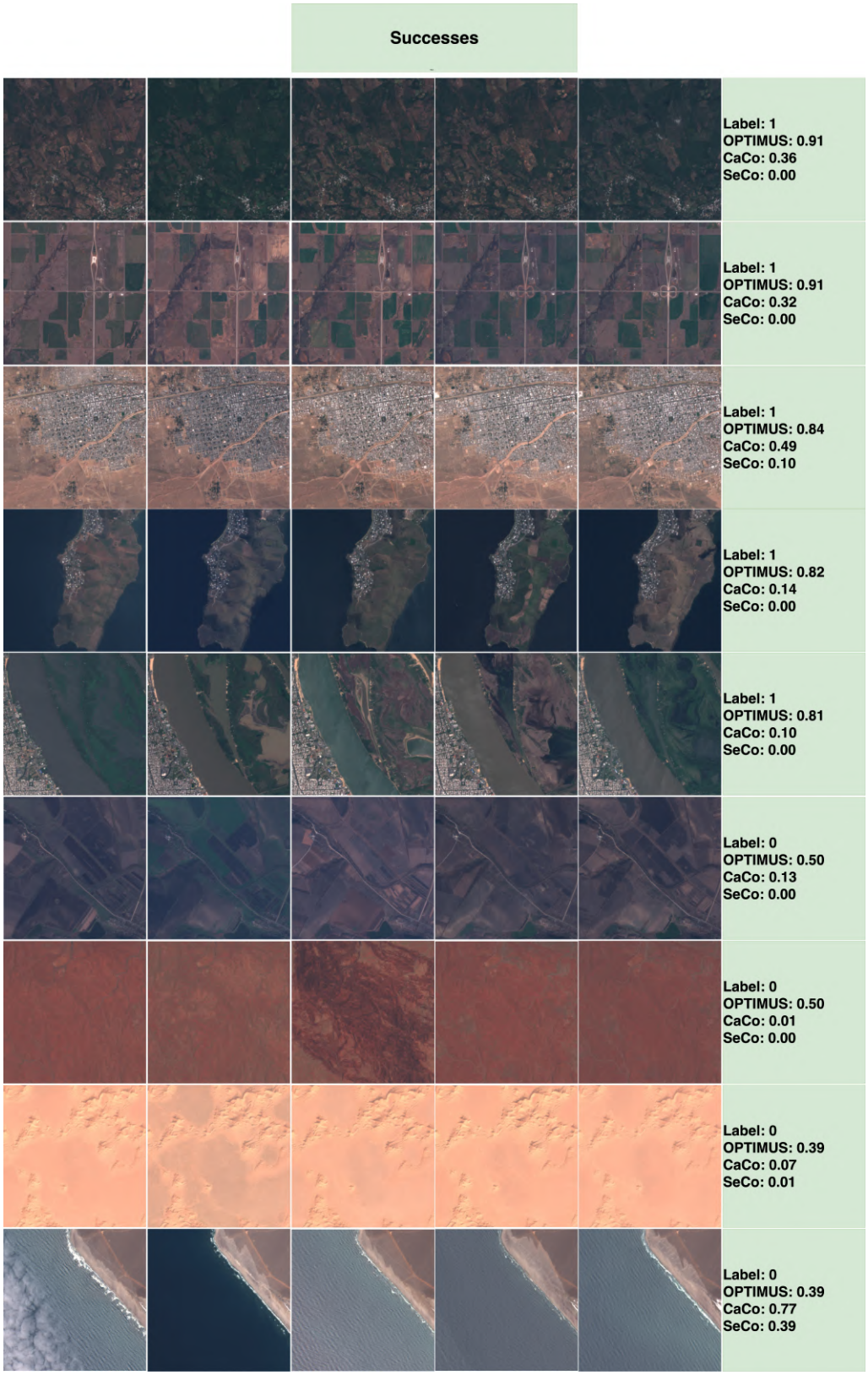

Figure 6. Examples of successful predictions by OPTIMUS, with performance scores from CaCo and SeCo models across various environments. The images are shown from left to right for the years 2016, 2018, 2020, 2021, and 2023, respectively, all captured in the month of November. Each row corresponds to a specific time series, with labels indicating whether a persistent, non-cyclic change is present (1) or absent (0).

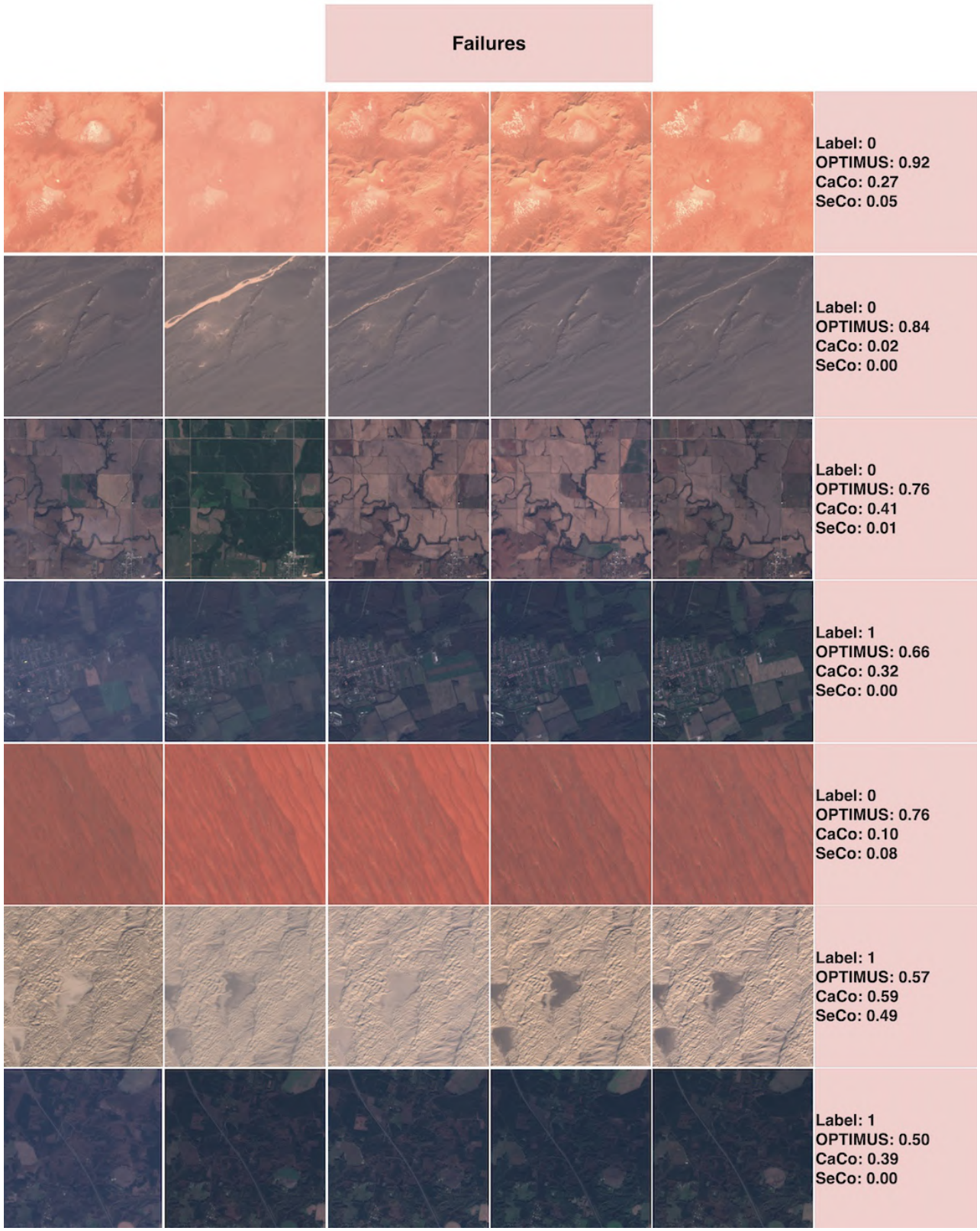

Figure 7. Most failures occur in complex environments with both seasonal and urban changes, or in situations where visual artifacts (e.g., shadows, lighting changes) mislead the models. Note that row 6 is incorrectly labeled (it should be 0); this was identified earlier, and we have since double-checked the evaluation dataset.