Supplementary Material for "Change Event Dataset for Discovery from Spatio-temporal Remote Sensing Imagery"

1 Overview

In this supplementary material we present more information about the dataset (including a datasheet for the dataset) and extensive results that could not fit in the main paper. In sec. 2 we include a datasheet for our dataset. In sec. 3 we present more details about our method such as architecuture details and hyperparameter value selection. In sec. 4 we look at the statistics of our two benchmarks CalFire and CaiRoad. In sec. 5 we present ablations and qualitative results for our unsupevised change detection method. In sec. 6 we present results for our change grouping method. In sec. 7 we look at the retrieval results on CaiRoad qualitatively to understand the challenging nature of the dataset. In sec. 8 we discuss results analyzing baselines with different temporal aggregation instead of averaging. In sec. 9 we provide more information about our label collection interface for CaiRoad. In sec. 10 we show examples of longer spatio-temporal change events that were hard to present in the main paper due to space constraints. In sec. 11 we share preliminary results of using our pipeline on other video domains.

The data is publicly available at `https://www.cs.cornell.edu/projects/satellite-change-events/`. Our code for accessing Sentinel-2 images, creating change events and baselines can be found at `https://github.com/utkarshmall13/satellite-change-events`.

2 Datasheet

We include a datasheet for our dataset following the methodology from "Datasheets for Datasets" [7]. In this section we include the prompts from [7] in blue and in black are our answers.

2.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to foster research on the problem of automatic discovery and semantic understanding of change events in satellite imagery. More specifically, the dataset should aid in developing systems that can automatically detect change events in satellite imagery and assign to each a semantic label that indicates the nature of the event, e.g., forest fires, road construction etc.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Utkarsh Mall, Bharath Hariharan and Kavita Bala at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

This work was funded by National Science Foundation (CHS-1900783) and (IIS-2144117), DARPA Learning with Less Label Program (HR001118S0044), IARPA (2021-20111000006), and Tata Consultancy Services Research.

Any other comments?

The dataset contains RGB bands from Sentinel-2 satellite imagery. So the spatial resolution of the data 10 metres. Users should keep in mind that changes smaller than the resolution be undetectable. For example, changes to roofs of houses, movements of traffic will not be detected. The datasets should be used for larger changes such as forest fire, crop changes etc.

2.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

An individual instance in the dataset is called a change event. A *change event* is a group of pixels over space and time that were changed by a single event, such as road construction or drying up of a reservoir of water, among others. These changes can be of arbitrary shape and size in both temporal and spatial dimensions. Except for disregarding changes due to pixel noise or illumination changes (due to change in the sun's angle), we do not place any restriction on the kind of change; this is to account for the differing needs of myriad applications. We are interested in *detecting* these change events, but also categorizing them into *classes* (e.g., road construction).

How many instances are there in total (of each type, if appropriate)?

In CaiRoad benchmark, there are a total of 28015 change event instances with 2259 of them labeled as road construction and the rest being other events. In CalFire benchmark, there are 2172 change events, with 204 forest fire events and 1968 other events. Also see Fig. 4 for more information and statistics of the dataset.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all the change events discovered by our methodology within a selected temporal span (2015-2021) and selected regions and is not a random sample. The reason for using 6 years of RGB Sentinel-2 imagery is that the latest Sentinel instrument has been active only since late 2015. For CaiRoad benchmark, the city of Cairo was chosen as the city in focus because it has seen a lot of new construction in the past decade. However within the area of city every region is given equal importance. For CalFire benchmark, the regions were chosen based on the locations of forest fires and are thus not random. The reason for using such a selection criterion is that fires are very rare events, and as a first step we wanted to get a larger representative set of fire samples in the benchmark.

What data does each instance consist of? Raw data (e.g., unprocessed text or images)or features? In either case, please provide a description.

We describe a change event instance using an ordered pair $\langle V_{1\cdots l}, C_{1\cdots l-1}\rangle$. Here $V \in R^{l\times x\times y\times c}$ are 3-D volumes of sequences of satellite images. $C \in \{0,1\}^{l-1\times x\times y}$ is the change between consecutive frames for the change event. C has a value of 1 whenever a change has happened otherwise it is 0. l, x, y are the span of changes in time and space and c is the number of channels in satellite images.

In the dataset, change events are represented by a sequence of images and a sequence of masks. So each of the frame in $V_{1\cdots l}$ is represented by an RGB satellite image. The change masks $C_{1\cdots l-1}$ are represented using a set of binary images.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each change event is annotated with a semantic label. In the CaiRoad benchmark each instance is labelled with it being a road or not. In the CalFire benchmark instances have forest fire labels.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

All the information is included in the instances.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Relationships between instances are not provided explicitly except for same labels. But since the metadata of instances contains information about location and time of the change events, some information such as proximity in space/time, event being part of same forest fire/road construction plan can be extracted.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We provide a train-test split on both our benchmarks for the linear classification task. We keep the training testing split to be around 50-50. Having a larger testset is especially important for CalFire benchmark where the number of positive class examples are not too many (204).

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There aren't any redundancies in the dataset, since each change event spans over a unique location or time. There are a few sources of error that arise due to our algorithm. For example, we know that the change detection algorithm we use is not 100% accurate and thus not all the changes can be recovered by our method (also the method might predict changes even when no change is happening).

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained as we provide all the change events their associated images and labels. The frames in change events are crops of the satellite images from Sentinel-2 imagery, which are publicly available. This dataset is free to use for non-commercial usage and available to public. For example, using our methodology additional bands can be collected from this dataset to augment our dataset.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals nonpublic communications)? If so, please provide a description.

No, as stated in the previous response Sentinel-2 imagery is free to use for non-commercial usage and available to public.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The dataset contains satellite images at medium resolution (10m) and we do not believe it contains anything offensive, insulting, or threatening.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset

No. It does not identify any subpopulations

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No. It does not contain information about individuals.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No. It does not contain any sensitive information.

Any other comments?

None

2.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The change events were discovered by using the unsupervised change event disovery method. The discovery method was applied on publicly available Sentinel-2 images from the city of Cairo and state of California. Please refer for sec. 3.2 (main paper) for the detailed information about the method. The labels were collected using human annotators.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

The raw satellite images were collected using Google Earth Engine APIs [1]. Our self-supervised method were trained and used on university servers with GPUs (GeForce GTX TITAN X). The manual annotation of labels was done using the Prolific [2] crowdsourcing platform.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is not a sample of a larger dataset.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A total of 312 different human annotators were shown change events and asked if the change event corresponds to the construction of a road or not. On average it took about 3.45 seconds for users to understand if a change event was a road construction or not. Compensation for a 100 change events was 1.85 USD. Average wage was approximately 15 USD / hour. Every change event was labeled by 3 separate annotators. An event was considered a road construction event if 2 out of 3 annotators agreed. In total these annotations cost approximately 800 USD. Also see Fig. 9- 11 for the interface and tutorial that is shown the annotators.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset contains all the change events discovered by our methodology between 2015 and 2021. The reason for using 6 years of RGB Sentinel-2 imagery is that the latest Sentinel instrument has been active only since late 2015.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The study was exempted from IRB as we do not collect any individual/personal information from users.

[1] `https://developers.google.com/earth-engine`

[2] `https://www.prolific.co/`

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Our dataset does not contain information about individuals.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Our dataset does not contain information about individuals.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Our dataset does not contain information about individuals.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Our dataset does not contain information about individuals.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Our dataset does not contain information about individuals.

Any other comments?

None

2.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

A big portion of change events (80%) occur due to occurence of clouds in images. Such a dataset would be less useful for future researchers as it is heavily biased towards cloud based changes. To resolve this problem we perform change grouping only on change detections not happening because of clouds. Automatic cloud detections can be found in most remote sensing datasets and we use these standard cloud masks [1]. Alongside cloud filtering, we additionally filter out very small change events (< 40 voxel) as they usually correspond to voxels that were not grouped by the region growing.

Was the raw data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the raw data.

Alongside the change events we also provide the full stack of raw Sentinet-2 images that we use to get change events with our dataset.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

The filtering is done using the methodology presented in the paper. Our code can be found at `https://github.com/utkarshmall13/satellite-change-events`.

Any other comments?

None

2.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset presented a novel task and has not been used for any tasks yet.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

N/A

What (other) tasks could the dataset be used for?

Our datasets can be used create benchmarks for other types of change events as well. For example, one can use information from agriculture surveys (or human annotations) to label change events such as growing/harvesting of crops.

While the benchmarks are primarily aimed at representation learning for change events, it can be used for other applications too. For example, the labels can be used to train a forest fire detection model or can be used as a dataset for supervised road change detection.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Our dataset does not contain information about individuals, so it should not result in unfair treatments of individuals or groups.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Any other comments?

None

2.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publically available on the internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset can be downloaded from Cornell's server at `https://www.cs.cornell.edu/projects/satellite-change-events/`. The dataset currently does not have a DOI but we are planning to get one.

When will the dataset be distributed?

The dataset is available since June 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is available under Creative Commons Attribution-NonCommercial 4.0 International License..

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Since our dataset is derived from Sentinel-2 images. Please also refer to Sentinel-2 terms of service[3].

[3] `https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/TermsConditions`

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No, there aren't any restrictions on the dataset.

Any other comments?

None

2.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted and supported by the web servers at Cornell. The CS department at Cornell will be maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Utkarsh Mall can be contacted via email (ukm4@cornell.edu). More updated information can be found on the dataset webpage.

Is there an erratum? If so, please provide a link or other access point.

No

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

The updates to the dataset will be posted on the dataset webpage.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Our dataset does not contain information about individuals.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers

In case of updates, we plan to keep the older version of the dataset on the webpage.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

Others may do so and should contact the original authors about incorporating fixes/extensions.

Any other comments?

None

3 Implementation Details

In this section we present the implementation details of our framework that we could not present in the main paper. First, we describe our architecture, training details, and hyperparameters for unsupervised change detections (sec. 3.1). Then we look at implementation details for change grouping (sec. 3.2) and the feature representation for change events (for baselines) (sec 3.3).

3.1 Unsupervised Change Detection

We give details on our architecture for unsupervised change detection and explain how we set up the training for self-supervised learning.

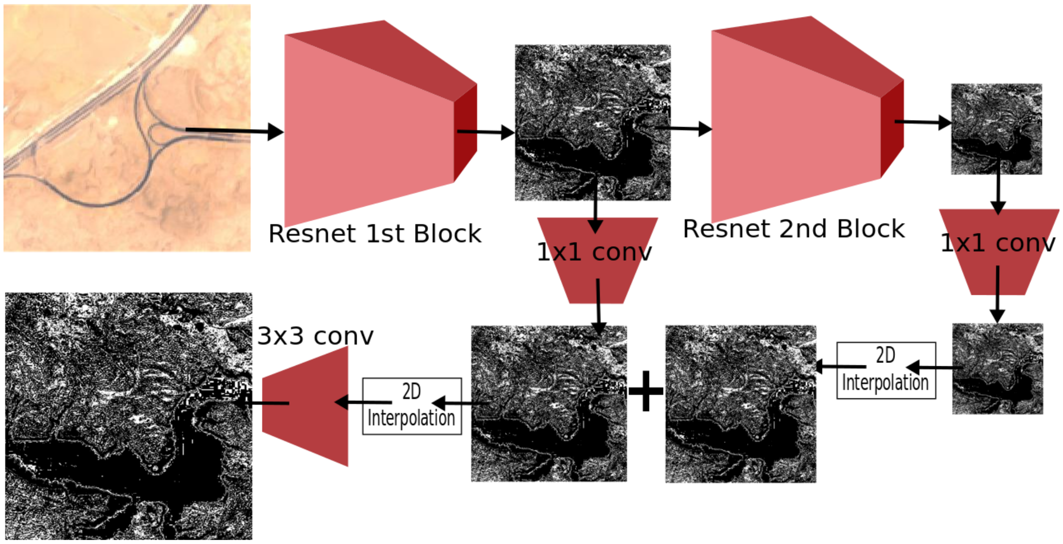


Figure 1: Architecture of our model for learning features for unsupervised change detection.

Architecture. Since changes in satellite images are local we use a shallower network with a smaller receptive field as f (First two blocks or ResNet-18 [9]) We use the Feature Pyramid architecture [10] to scale up the resolution of the output map. We use augmentations such as color jitter as photometric transformations. For geometric transforms, we use random rotations, translations, scaling, horizontal/vertical flips, random crops. We make sure that in a batch all the examples are from different locations to avoid contrasting features from nearby locations.

Fig. 1 shows the architecture of the model. We use features after each block and combine them by upscaling the second layers.

Training. Training the SimCLR loss requires large batches to allow the model to see and contrast between a diverse set of images. Since in our loss function we treat feature vectors from the same images as contrastive examples, larger sized batches of images cannot be used during training. This is because the gradient calculation becomes very large. To enable larger batch sizes, we subsample the output feature maps. A uniformly spaced subsampled grid is used instead of the full feature map. Without this, learning the representation takes longer to train to reach the same level of performance.

We use a temperature of $\tau = 0.07$, learning rate of $1e^{-3}$, and a subsampling rate of 16 (1 in every 16 feature vectors are selected). The temperature is relatively lower than what is used in SimCLR training on natural images as lower temperature works better. We believe this is due to satellite images having fewer diverse objects/textures than natural images. So the temperature must be lowered to have peakier output. The subsampling rate of 16 allows us to have a batch size 16 times larger.

3.2 Change Grouping

Hyperparameter selection: For eq. 2 in the main paper, we use euclidean distance metrics for all distances d_x, d_t and d_f. For our resolution of satellite image (10m, 1 month), δ_{st} is set to 20 voxels, so any voxel within a 20-voxel (200 metres in spatial dimensions) distance can be considered a neighbor. We set $c_t = 4$ (making potential neighbors to be within (20/4) = 5 months), so voxels have to be closer in time to be considered a potential neighbor. Since δ_f can vary with the type of features and type of dataset (for example, the variance between classes), we set δ_f to be some fraction of the average distance between any two voxels.

$$\delta_f = c_f \cdot \sum_{i=0}^{i=V} \sum_{j=i+1}^{i=V} d_f(v_i, v_j)$$

We set $c_f = 1$ for all our experiments. We implement this formulation of region growing as finding connected components on a graph with the voxel neighborhood defined using eq. 2 (main paper).

The benchmark is not very sensitive to changes to hyperparameters within a reasonable range (60m-300m $\delta_{st} = 3, 30$, 2-8 months $c_t = 10, 2.5$). Reducing δ_{st} to 60m leads to slightly more events 2203 (from 2172). The reason for only 1.4% increase is that the region growing is not very sensitive to these hyperparameters. The majority of components that were connected with $\delta = 20$ still remain connected with $\delta = 3$. Additionally experiments on these datasets with small differences from the original dataset do not significantly change performance plots and number. Therefore the benchmarks are not very sensitive to the hyperparameters.

Additional Cleaning: A big portion of change events (80%) occur due to the occurence of clouds in images. Such a dataset would be less useful for future researchers as it is heavily biased towards cloud based changes. To resolve this problem we perform change grouping only on change detections not happening because of clouds. Automatic cloud detections can be found in most remote sensing datasets and we use these standard cloud masks (namely COPERNICUS/S2_CLOUD_PROBABILITY on EarthEngine) [1]. We select change pixels that have $< 10\%$ probability of being clouds. We additionally filter out very small change events ($<$ 40 voxel) as they usually correspond to voxels that are not able to be grouped by the region growing.

3.3 Feature Representation for Change Events

Training. We use a temperature of $\tau = 0.07$, learning rate of $1e^{-3}$, for training the network. The backbone network is Resnet-18 before the global average pooling. We also start from a pre-trained ImageNet model as that helps training. Note that for representation learning the global average pooling is replaced with weighted average pooling.

4 Statistics for CalFire and CaiRoad Benchmarks

For CalFire we used a total 25688 Sentinel-2 images and 24830 for CaiRoad. We use the R (664.5nm), G (560nm), and B (496.6nm) bands from Sentinel-2. The information about geolocation of individual tile in the dataset can be found in the metadata of the dataset.

Fig. 2 shows the distribution of change events found by our method and present in our benchmarks CalFire and CaiRoad respectively. It shows distributions of different properties of change events like temporal span, maximum slice area, maximum spatial bound and total area (see the figure for descriptions of these properties). Change events in CalFire are on average longer temporally than CaiRoad and they also have larger areas. The figure also shows both, the statistics of positive events (lighter color) and all events (darker color).

Note that for both CaiRoad and CalFire, the average and median road construction/fire event is larger than the average of all events. The fire events have significantly larger average areas. The road events have relatively larger spatial bound which can be expected because road constructions are likely to have smaller area but larger length.

5 Evaluating Change Detection

Qualitative Evaluation of Change Detection. We first look at the change detection results qualitatively. Fig. 3, shows examples of changes detected by our method on pairs with drastic illumination changes. In the first row, the two images have a big illumination change where the image on the left is very bright. However our method can correctly recognize that as illumination change and ignore it. In the third row, our method learns to ignore changes due to shadows. In the middle row our method correctly detects new construction, while being invariant to illumination in the left. Our method is very robust to big illumination changes, while being able to detect real changes.

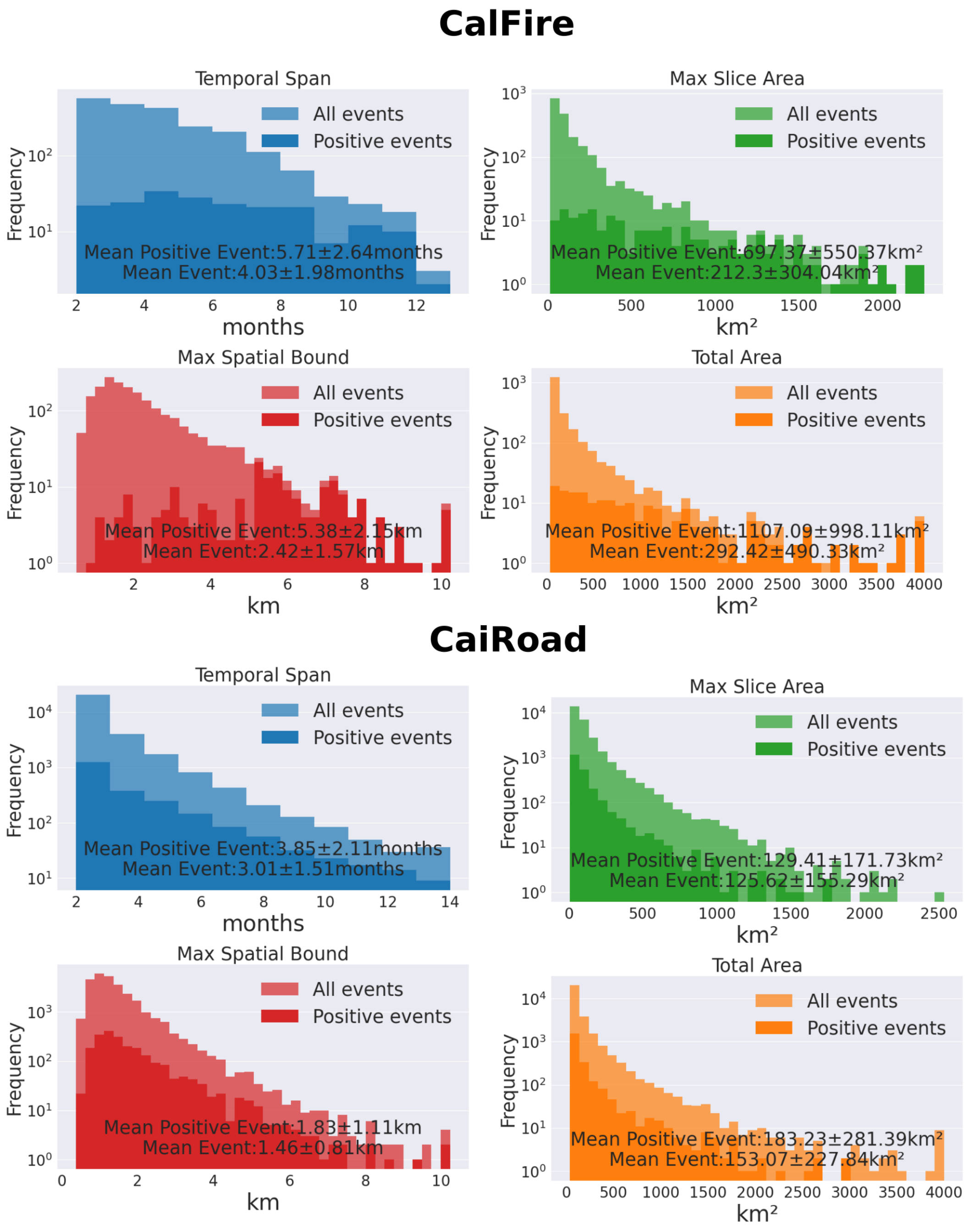


Figure 2: Distribution of change events in our benchmarks. Each plot is a histogram of temporal span, maximum slice area, maximum spatial bound and total area. Temporal span is how long a change event is in time space. It also shows distribution for both positive and all events. Total area is the sum of area of change masks summed over time. Maximum slice area is the area of the temporal slice with maximum area. Maximum spatial bound is the maximum length of change along latitude or longitude. The number in the plot shows the mean and standard deviation.

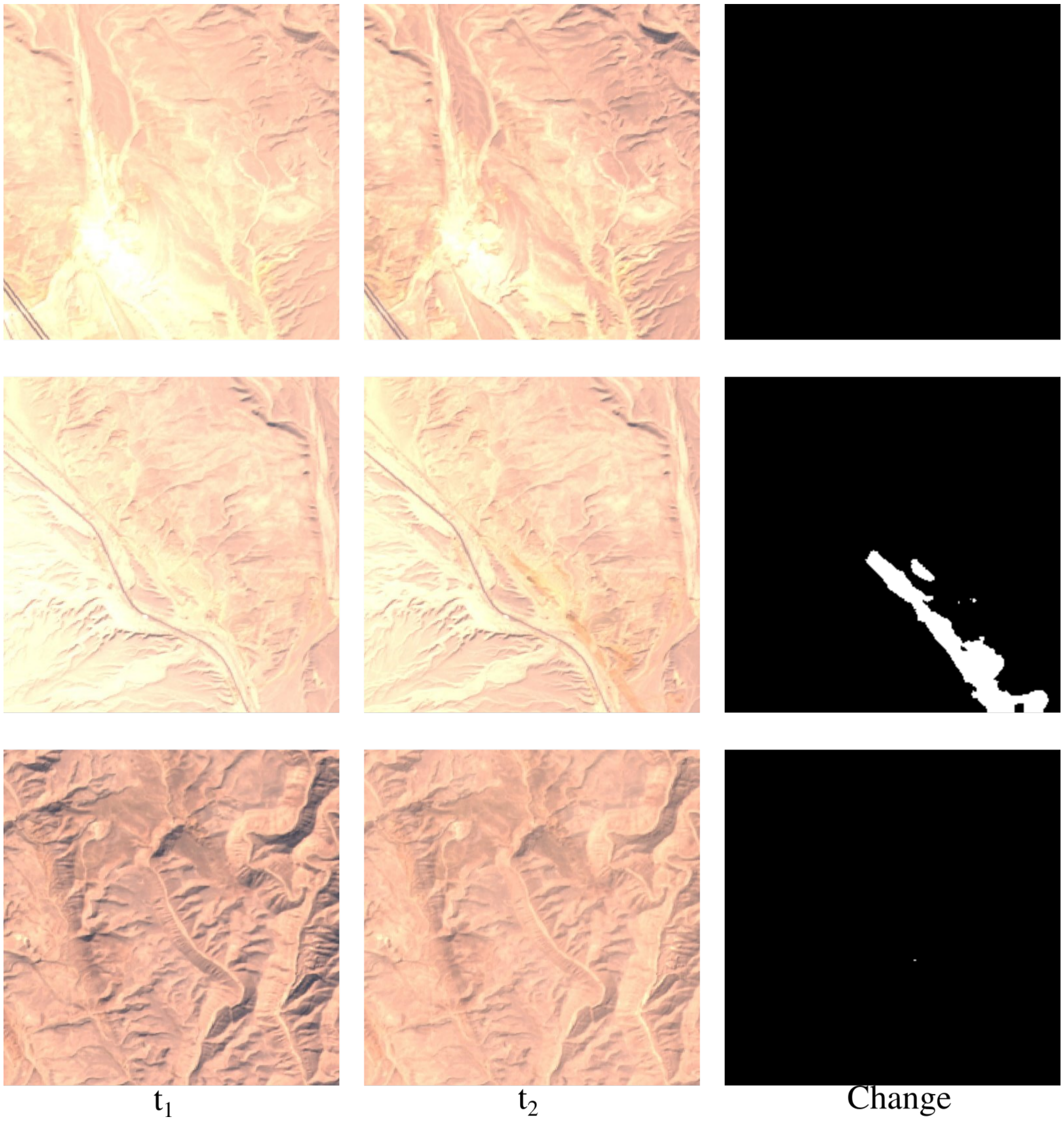


Figure 3: Results of change detected by our method on pairs with drastic illumination changes. Our method is invariant to illumination changes while being aware of other changes.

Dataset for Change Detection Table 1 gives information of different change detection datasets. Several change detection methods evaluate on datasets such as SZTAKI, OSCD and a larger dataset LEVIR-CD. These datasets have binary information of change or no change and the change masks are annotated by humans. DynamicEarthNet has temporal landcover segmentation labels, which can be used to create land cover change masks. Overall 1725 change masks can be created over 75 locations. Many of these change masks can be empty as landcover does not necessarily change at many location over smaller period of time. We also compare these datasets to our benchmarks. Note that the comparison is not very fair as our benchmarks are created automatically and labelled semi-automatically, whereas these datasets are manually created. Additionally, our benchmarks have change events with semantic labels in addition to the binary change masks.

Baseline Implementation. For all the baselines we use the same hyperparameters and inference methods as present in the original implementation of these papers when evaluated on the OSCD

Dataset	Dataset Size	Image Size	Resolution	Semantic Info.	Annotation
SZTAKI [3]	12	940×640	1.5m	✗	Manual
OSCD [4]	24	600×600	10m	✗	Manual
LEVIR-CD [5]	637	1024×1024	0.5m	✗	Manual
DynamicEarthNet [13]	1725	1024×1024	10m	✓(landcover)	Manual
Calfire	2172 (events)	242×242 (mean)	10m	✓(1 class)	Automatic
CaiRoad	28015 (events)	146×146 (mean)	10m	✓(1 class)	Automatic

Table 1: Information on various change detection dataset with their datasset size, image size, resolution and semantic information available in them.

Method	F-score	Cohen's κ-score
Larger Receptive Field	0.259	0.223
No intra-image contrast	0.299	0.263
No inter-image contrast	0.272	0.242
Ours	**0.321**	**0.287**

Table 2: Performance of our unsupervised change detection method on the OSCD benchmark compared to ablations of the model. Having a larger receptive field hurts the performance as changes are very local and larger recepetive fields capture larger regions. Doing SimCLR without equivariance transforms *i.e.*, only contrasting augmented patches to patches in other images leads to a loss in performance as well. Finally, using batches with multiple images allows the loss function to see diverse examples to contrast with. Hence, not having inter-image contrast leads to lower performance.

benchmark. So the comparison is fair. Additionally, we optimize the data preprocessing step using GPUs for all the baselines resulting in faster preprocessing. So the inference runtimes are faster than the original implementations.

Justification for False Positives. In the main paper we evaluated our self-supervised change detection method on OSCD dataset. While the performance of our method was better than all the baselines and existing methods for unsupervised change detection, there an an issue. The F1-score remains low for all the methods on this dataset, while the recall is high. The reason for low F1 score is that the true change labels are not exhaustive and only urban changes are annotated in OSCD. Thus, the false positives involve other kinds of changes, like natural disasters or seasonal changes. Fig. 4, shows different examples of changes (zoomed in red and green) which are not annotated in the OSCD dataset as they are not urban changes. But these are still detected by our change detection method. Changes like this result in false positives in the OSCD benchmarks but actually do denote a change. Our methods high recall shows that it recovers a big fraction of the annotated changes while having a higher F1 score than the baselines.

Ablation. We perform ablations on our unsupervised change detection method to demonstrate the importance of having our proposed architecture and the training method. Table 2 shows the comparison of our method against 3 ablations on the OSCD benchmark [4]. The first ablation considers a model with a larger receptive field that captures more global features which are not required for change detection, as demonstrated by the lower performance. The second ablation shows a model with no intra-image contrast; this is a model similar to image-based SimCLR. We do not perform geometric transforms; instead we simply take features of patches and contrast them with features of other patches. As seen by the lower performance of this ablation, learning equivariance to geometric transforms leads to a better performing model. The third ablation considers no inter-image contrast; it only performs contrastive learning within a single image. This model cannot see diverse enough regions in a single batch to contrast from and hence is not very good.

Fig. 5 shows an example of change detected by our method vs. KPCA-MNet on Cairo. KPCA-MNet detects many changes that are because of illumination changes and thus are not very useful, whereas our method can successfully suppress these changes and find the important changes like the new road.

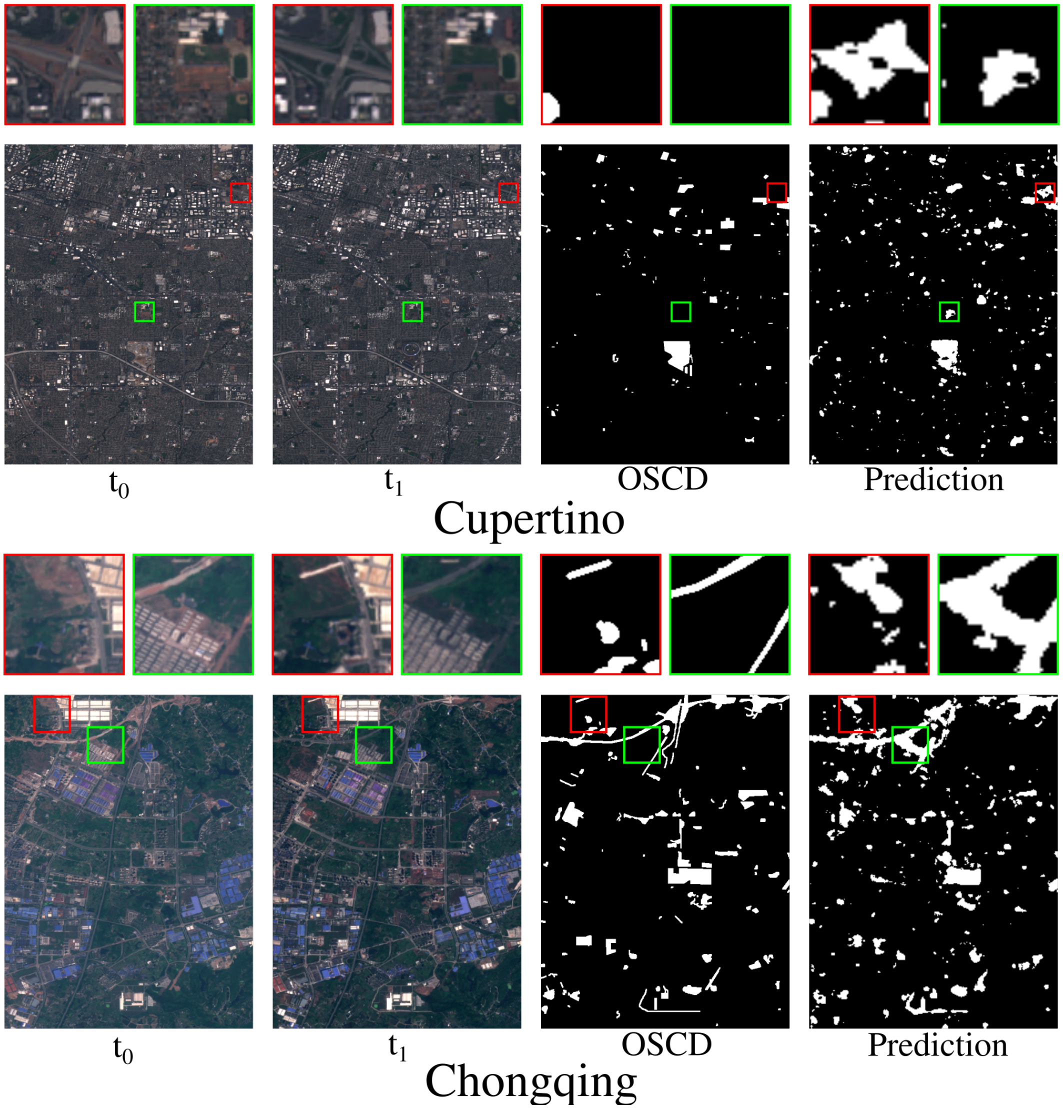


Figure 4: Results of our change detection method on Cupertino and Chongqing. Many changes detected by our method (zoomed in red and green) are evaluated as false positives as OSCD only labels urban changes. But changes detected by our methods are also real changes.

Method	Recall	F-score
CVA [11]	0.340	0.122
DCVA [12]	0.495	0.141
PCANet [6]	0.358	0.118
KPCA-MNet [14]	0.464	0.140
Ours	**0.604**	**0.155**

Table 3: Performance of our unsupervised change detection method in comparison to baselines on the LEVIR-CD building detection dataset. Even though the F-score is low (because not all true labels present in the ground truth) our method has the best F-score. Our method also has better recall.

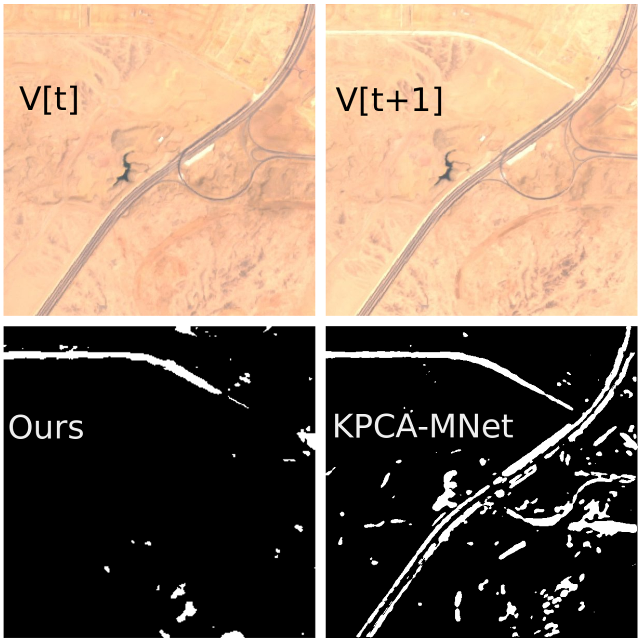


Figure 5: Change detected by our method vs. KPCA-MNet for the pair of images. KPCA-MNet detects many spurious changes due to changes in illumination (even after radiometric correction), whereas our method is invariant to them.

Performance on LEVIR-CD. We also evaluate our method on another change detection benchmark. LEVIR-CD [5] is a dataset with annotated building changes from google earth images. This dataset is much larger than the OSCD benchmark and contains a total of 673 pairs of changes in train+test set. In this dataset, although all different types of changes happen, only building construction events are marked.

Table 3 shows the performance of our method in comparison to baselines on the LEVIR-CD dataset. The F1-scores are lower for all the methods because the precision is low: this is because the detected changes might be correct but might still be marked as false positives because they do not involve building construction. Nonetheless our method has the highest F1-score. Further, our method produces highest recall by a significant margin showing that our method is better at retrieving building changes annotated by humans.

6 Evaluating Change Grouping

Fig. 6 shows examples of change grouping using our method with features from our self-supervised network and KPCA-MNet. Each color represents a different segment. KPCA-MNet are not trained to be scale, translation or rotation invariant. As a result features closer to the edge of an object are treated differently from features towards the center. This results in oversegmentation as can be seen

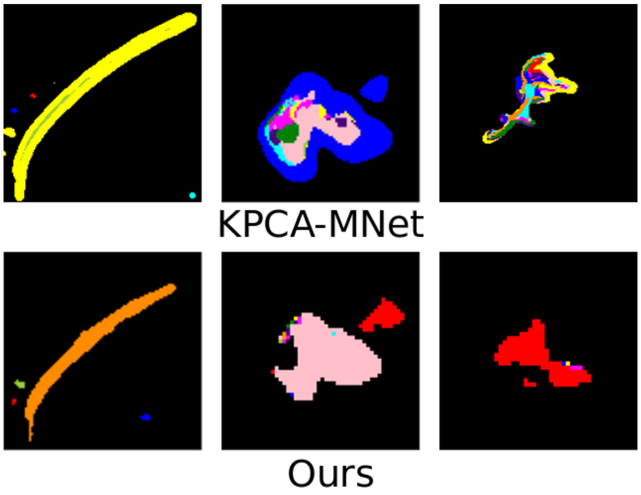


Figure 6: Example of changes grouped into segments when using our method vs. KPCA-MNet. Each color represents a different segment. Features in our method are invariant to rotation, scale and translation and thus lead to better grouping of changes. Whereas KPCA-MNet oversegments the changes where it is not required.

in all three cases. Note that all three changes are around the same region, but the change masks are different because the change detection method is also different.

7 Qualitative Retrieval Experiments.

We look at some retrievals done by the best performing method "SimCLR: Change Events" on CaiRoad. Fig. 7 shows the query events (left) and retrievals using the query event (right). We observe many false positives due to the challenging nature of the dataset (shown in red). In some cases the false positives are because a change event overlaps road construction. Examples of such change events are the first retrieval in the second row and second example in the third row. Another reason for incorrect retrival is a change event overlapping with roads without construction (the last example in row 3). Sometimes when a change event is near a road construction such as the third and fourth example in the first row. The reason for similar features of such event to a road construction event is a large receptive field of the CNNs. Even with the masking on the feature map, regions near the mask can influence the feature.

All these different reasons make the benchmark very challenging for existing vision algorithms. More work in the future is required to learn better feature representations for change events that can deal with these issues.

8 Baselines with Different Temporal Aggregation.

In the experiments in the main paper, in order to encode temporal information, we averaged the features across time for all these methods. We now also look at alternative approaches for temporal feature aggregation. We try three alternative to averaging for the nearest neighbor retrieval task.

- Dynamic Time Warping: Without the temporal aggregation the features for a change event can be treated as a multivariate time-series. We use dynamic time warping (DTW) [2] a method to compare time-series to measure similarities between change events.
- Dynamic Time Warping (Normalized): Since we use kNN neighbors with $k > 1$. If the query change events are of variable length the DTW similarity metric will vary a lot. So we normalize the metric with query change events lengths.

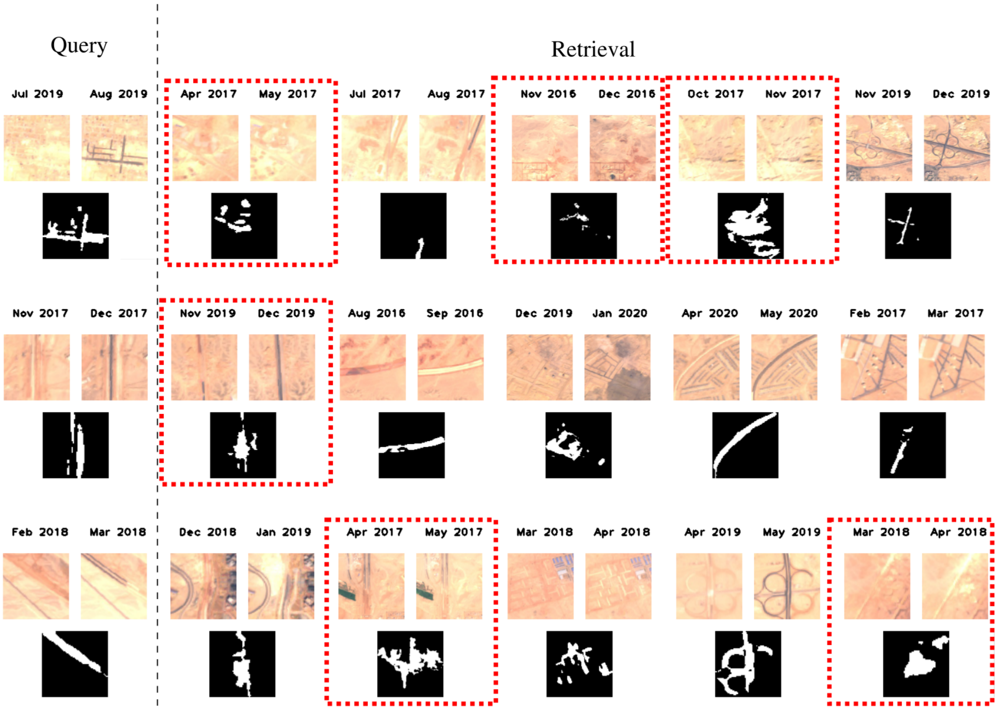


Figure 7: Query (left) and top-5 retrievals change events, using the selected query. Only the biggest temporal slice of the change events are shown for visualization. The method falsely retrieves many examples that are not road constructions (shown in red).

- Time-Series Transformer (TST) [15]: This is a transformer based representation of time-series information. This has been shown to better handle higher dimension multivariate data in comparison to DTW.

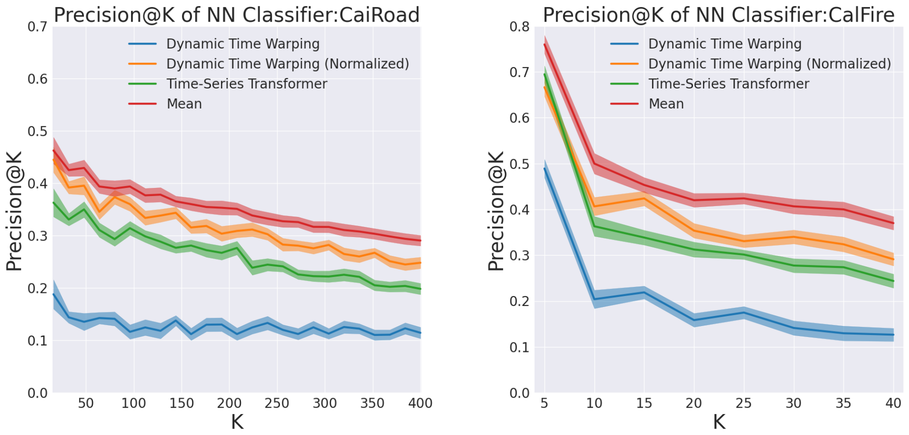
Figure 8, shows the performance of these various methods in comparison to the mean aggregation that we use in the main paper. Surprisingly the simplest technique of averaging works the best. We suspect that while technique like DTW or TST might work better with very long time-series in our case the time-series are not very long (they have a maximum of 14 datapoints). This leads to relatively poor performance when using these techniques. There is a big room for improvement on our benchmark and we posit that better temporal aggregation might be one potential direction of improvement in the future.

9 Interface.

As stated in the main paper, to collect labels for the CaiRoad benchmarks we use human annotations. In this section we present the interface shown to the annotators in Prolific.

The annotators first land on an instruction page as shown in Fig. 9. The instruction discuss various aspects of the annotation process. It includes instruction about the basic interface, for example, how to label a change event or undo an annotation. It also includes description of a change event and its representation, as a novice annotator will be unfamiliar with it. It also includes examples of positive and negative change events for road construction. Finally it provides some expected time to finish annotations for the task.

After reading the instructions the annotators can go the tutorial page (Fig. 10). In the tutorial page, a small tutorial task is presented with one 5 change events. After answering a question, the tutorial tells if the annotators are correct or not. It also explains why they are correct/incorrect.


Precision@K of NN Classifier:CaiRoad

Precision@K of NN Classifier:CalFire

Figure 8: Performance of various methods of aggregating temporal information. Surprisingly averaging works best.

The final annotation page (Fig. 11), is very similar to tutorial page, expect that there are many more examples and no explanations.

10 Examples of Satellite Change Events

Although using clustering methods on top of the learned representation results in noisy clusters, the examples closest to the cluster centers are meaningful. Figs. 12-16 show change events discovered by our method. The “SimCLR: Change events” method is able to capture some long-term changes as can be seen in the examples. Fig. 12 shows longer road construction events where we can see the progress of the roads being constructed over time with change masks. In Fig. 13 crops in regions are being harvested at different times leading to a longer change event. In Fig. 14 we can see crops being harvested but belonging to a different cluster due to it being in a different shape (circles) and thus a different semantic category. Fig. 15 shows the change event cluster detecting the occurrence of snowfall in California. The biggest change in the change mask can be seen when a lot of snowfall happens. Finally, Fig. 16 shows the change event cluster detecting receding water levels in lakes in California. This change event cluster can be used to measure the water level over time at a place.

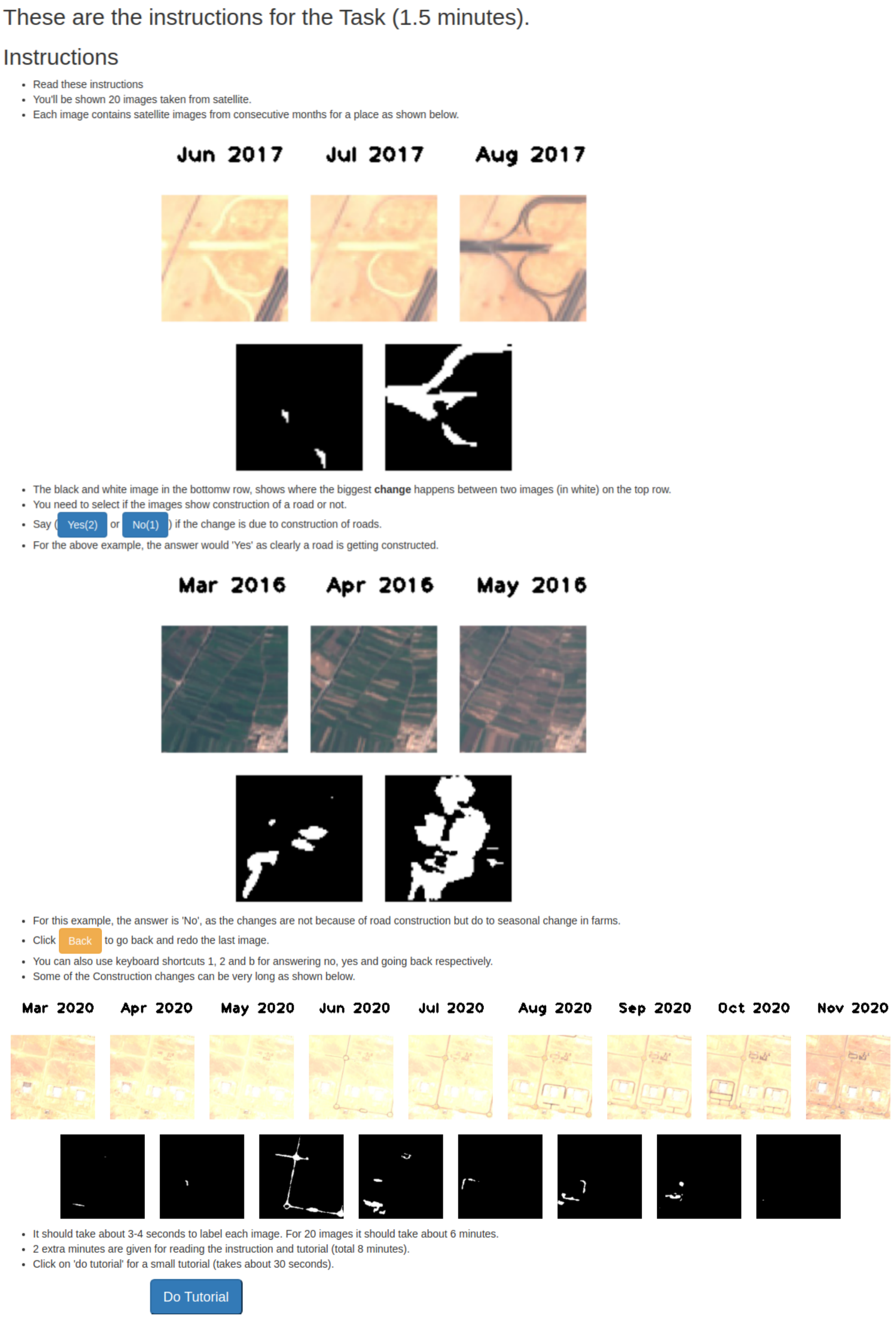
11 Video Change Events

We show that our method can potentially be applied more generally than satellite images, and can discover change events in other domains. We run our pipeline on static (almost static) camera video datasets such as talk shows [8], and cooking videos.

We show a proof-of-concept by running our pipeline on them. Note that to understand actions in videos, motion information is required (with RGB channels). So we also use optical flow channels.

Our pipeline discovers simple actions as change events on talk-show monologue videos. Fig. 17 shows examples of gestures found on the movement of hands and faces that are grouped together using clustering.

We additionally looked at cooking recipe videos as another static camera video domain. We collected data from the YouTube channel “Food Wishes”. We remove videos that are not about cooking recipes and are not shot with a static camera. Overall we get 87 such videos. We run our method to discover atomic cooking actions from this dataset. While the change events noisy as not all videos are completely shot with a static camera, we still can get some useful clusters. Fig. 18, shows a cluster discovered by our method, where the action being performed is “chopping”.



These are the instructions for the Task (1.5 minutes).

Instructions

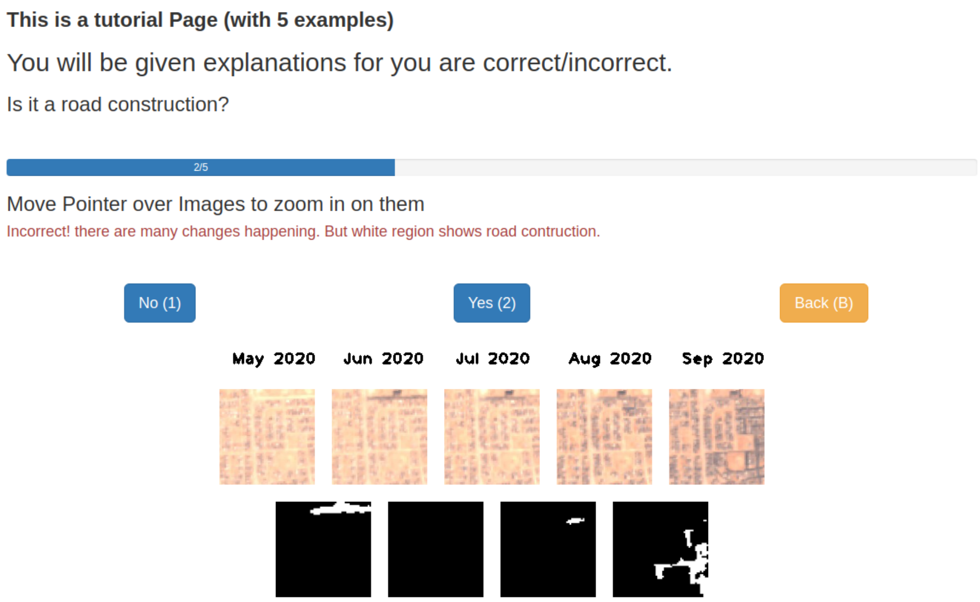
- Read these instructions
- You'll be shown 20 images taken from satellite.
- Each image contains satellite images from consecutive months for a place as shown below.

- The black and white image in the bottomw row, shows where the biggest **change** happens between two images (in white) on the top row.
- You need to select if the images show construction of a road or not.
- Say (Yes(2) or No(1)) if the change is due to construction of roads.
- For the above example, the answer would 'Yes' as clearly a road is getting constructed.

- For this example, the answer is 'No', as the changes are not because of road construction but do to seasonal change in farms.
- Click Back to go back and redo the last image.
- You can also use keyboard shortcuts 1, 2 and b for answering no, yes and going back respectively.
- Some of the Construction changes can be very long as shown below.

- It should take about 3-4 seconds to label each image. For 20 images it should take about 6 minutes.
- 2 extra minutes are given for reading the instruction and tutorial (total 8 minutes).
- Click on 'do tutorial' for a small tutorial (takes about 30 seconds).

Figure 9: The instruction page of our inferface for collecting labels from human annotators on CaiRoad. The instruction shows basic interface usage, positive and negative examples, talk about expected time taken.



This is a tutorial Page (with 5 examples)

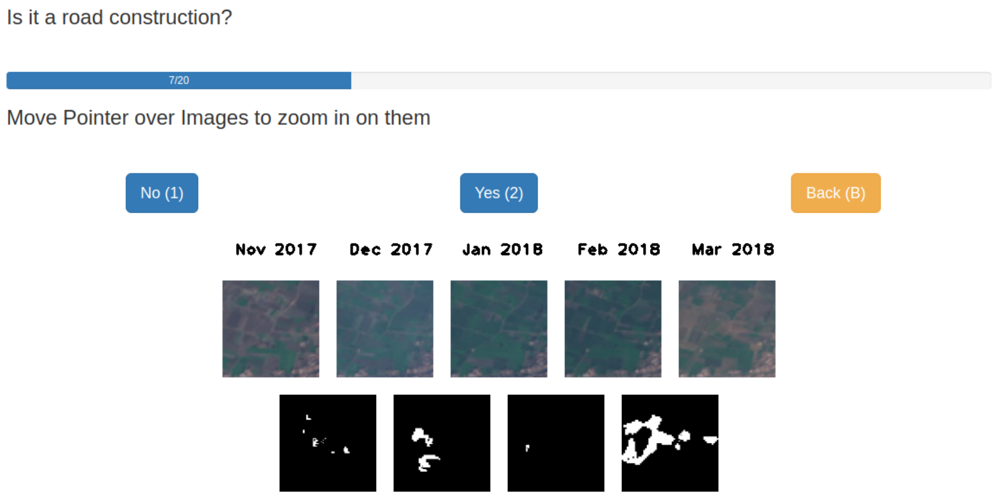
You will be given explanations for you are correct/incorrect.

Is it a road construction?

Move Pointer over Images to zoom in on them

Incorrect! there are many changes happening. But white region shows road contruction.

Figure 10: The tutorial page of our inferface for collecting labels from human annotators on CaiRoad. If the annotators are incorrect in tutorial, we provide them with reasons (shown in red).



Is it a road construction?

Move Pointer over Images to zoom in on them

Figure 11: The annotation page of our inferface for collecting labels from human annotators on CaiRoad. A large number of examples are shown on the annotation page.

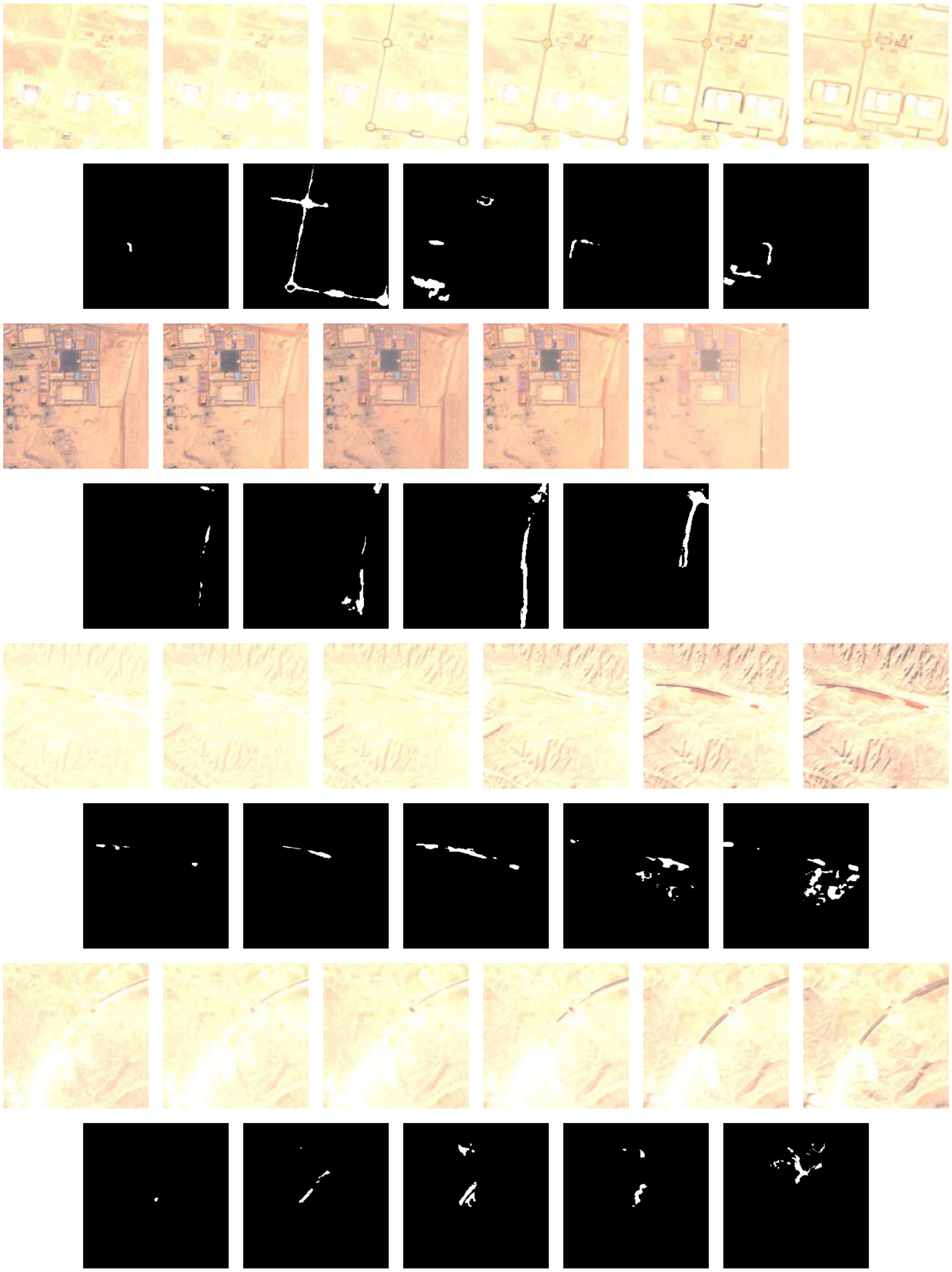


Figure 12: Longer Construction of road events discovered by “SimCLR: Change events”. Each pair of rows shows a single change over time, with its change mask right below it. Different rows show different examples that are closest to a cluster center. The first two examples show roads construction in an urban/industrial area. Whereas the last two rows show construction in remote areas. Note that our method can find similar change events “road construction” in different looking areas.

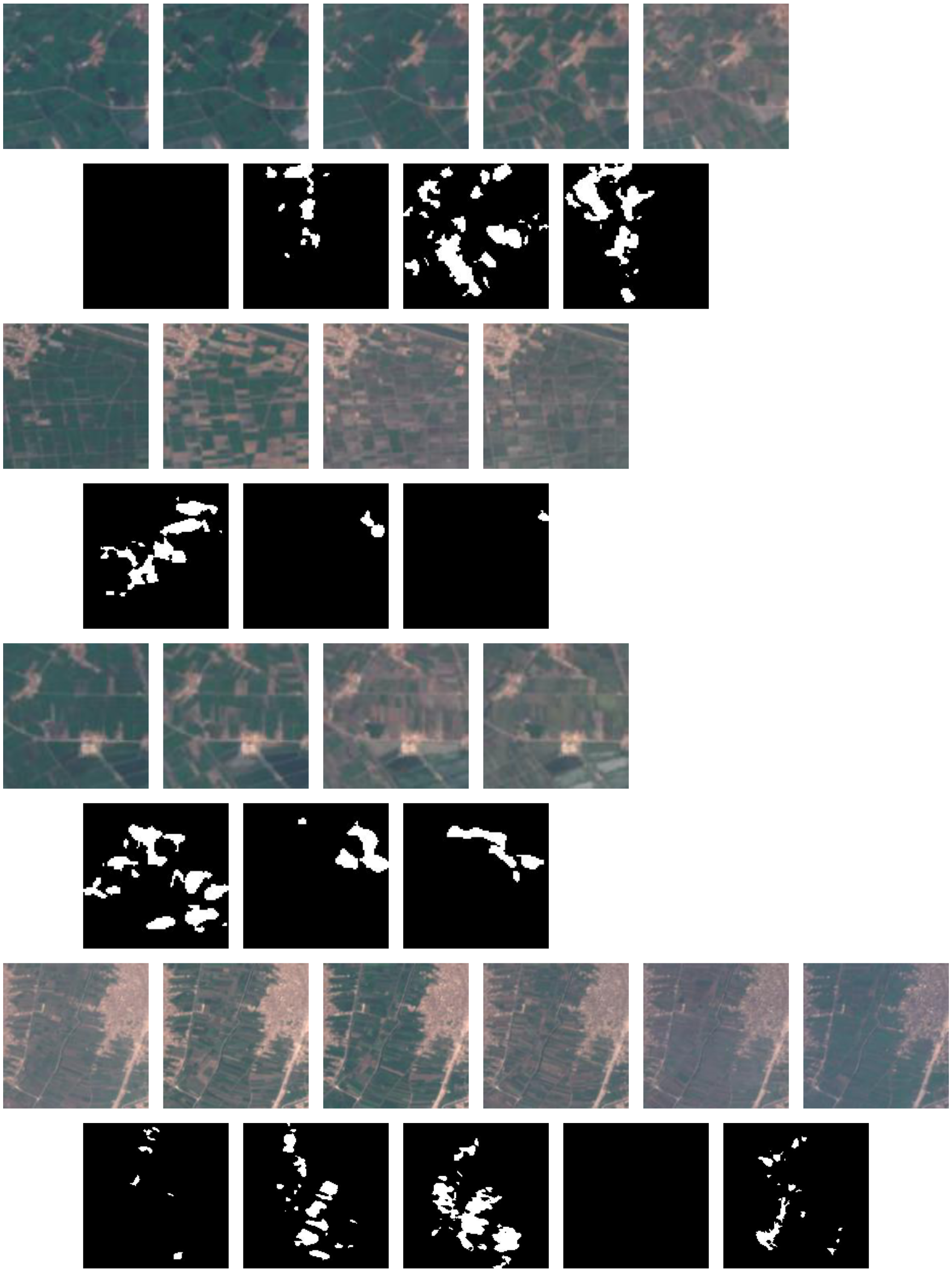


Figure 13: Longer Harvesting of crop events discovered by “SimCLR: Change events”. Each pair of rows shows a single change over time, with its change mask right below it. Different rows show different examples that are closest to a cluster center. In each example, we see crops in a region getting harvested.

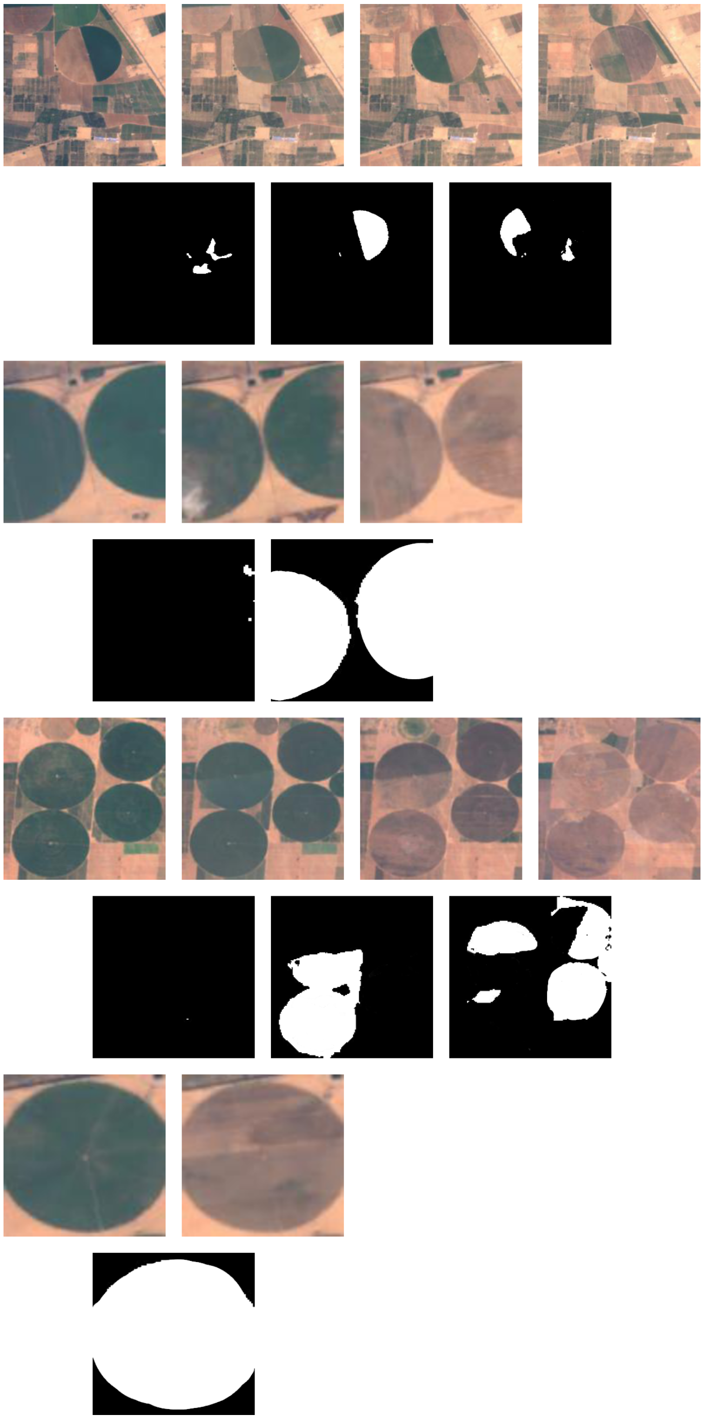


Figure 14: Longer Harvesting of crops in crop circles events discovered by “SimCLR: Change events”. Each pair of rows shows a single change over time, with its change mask right below it. Different rows show different examples that are closest to a cluster center. In each example, we see crops circles in a region getting harvested.

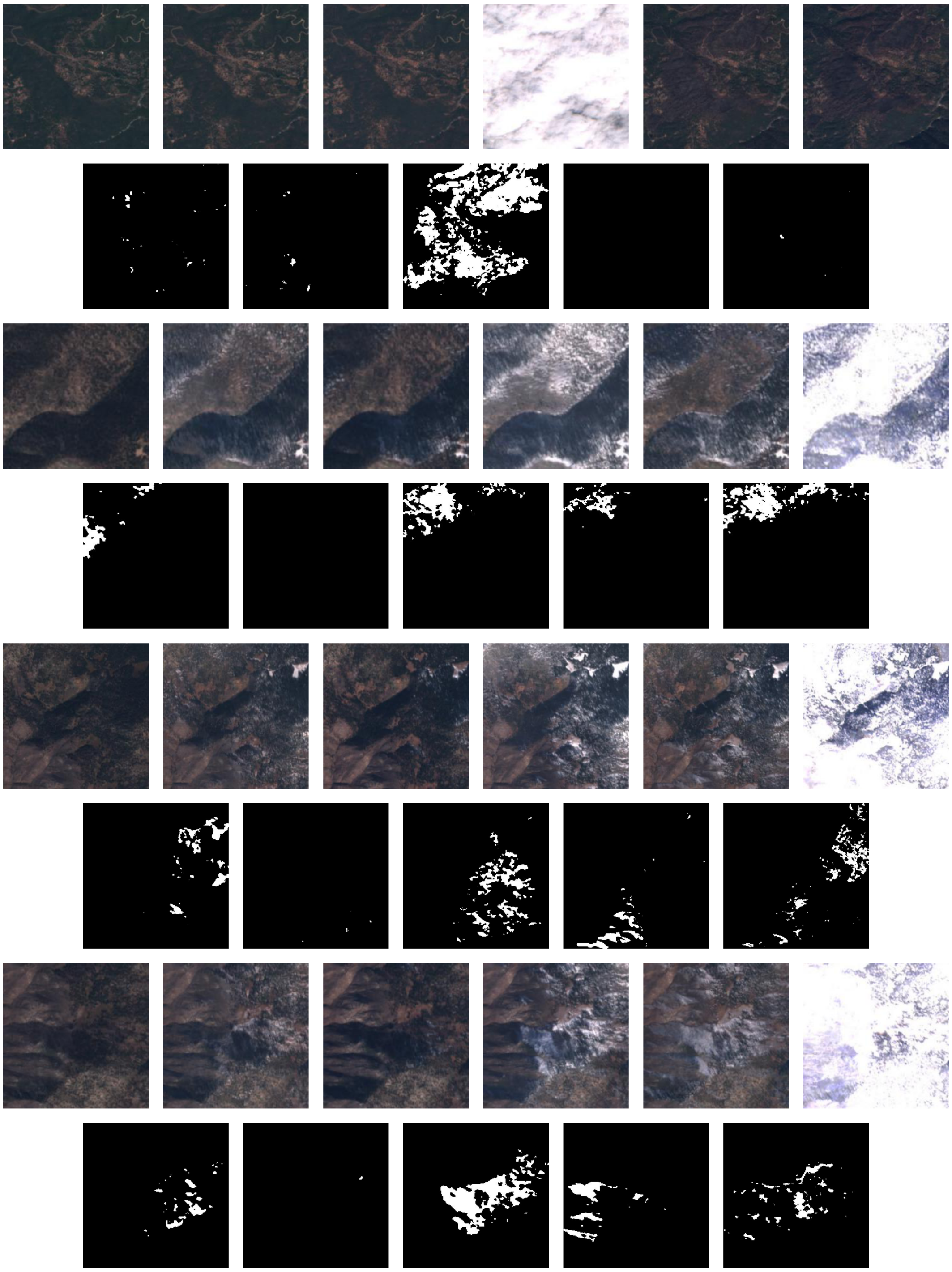


Figure 15: Snowfall events discovered by our method in California. Each pair of rows shows a single change over time, with its change mask right below it. Different rows show different examples that are closest to a cluster center. In each example, we see the occurrence of snowfall in a region where the change mask detects something.

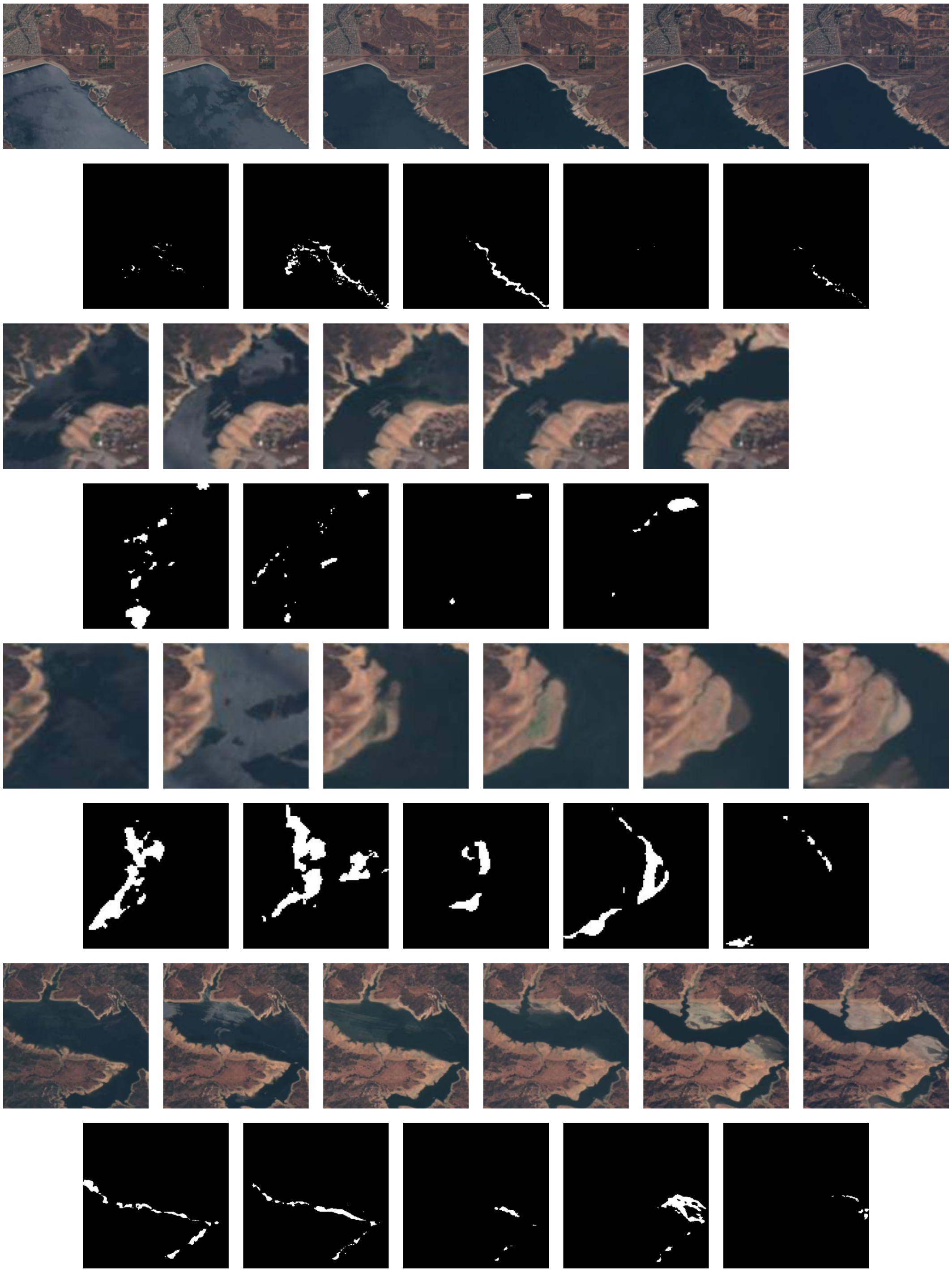


Figure 16: Receding water level discovered by our method in California. Each pair of rows shows a single change over time, with its change mask below it. Different rows show different examples that are closest to a cluster center. In each example, we see the receding water level in a region where the change mask detects something.

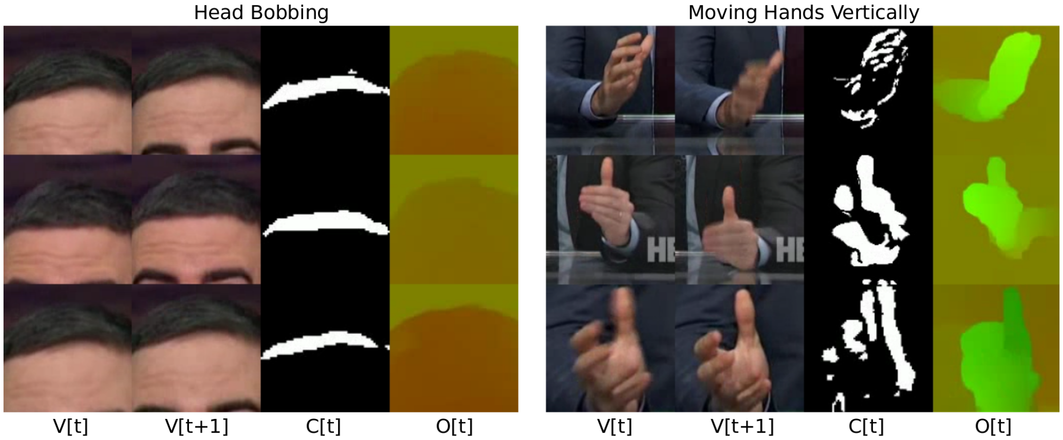


Figure 17: Discovery of gestures such as head bobbing (left) and moving open hands vertically (right) by our method from Last Week Tonight Videos. The last column shows the optical flow between consecutive frames.

References

[1] Sentinel-2. `https://sentinel.esa.int/web/sentinel/home`. Accessed: 06/09/2022.

[2] Richard Bellman and Robert Kalaba. On adaptive control processes. *IRE TAC*, 1959.

[3] Csaba Benedek and Tamás Szirányi. Change detection in optical aerial images by a multilayer conditional mixed markov model. *TGRS*, 2009.

[4] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS*, 2018.

[5] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 2020.

[6] Feng Gao, Junyu Dong, Bo Li, and Qizhi Xu. Automatic change detection in synthetic aperture radar images based on pcanet. *GRSL*, 2016.

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 2021.

[8] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *CVPR*, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[11] William A Malila. Change vector analysis: an approach for detecting forest changes with landsat. In *LARS symposia*, 1980.

[12] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-change detection in vhr images. *TGRS*, 2019.

[13] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *CVPR*, 2022.

[14] Chen Wu, Hongruixuan Chen, Bo Do, and Liangpei Zhang. Unsupervised change detection in multi-temporal vhr images based on deep kernel pca convolutional mapping network. *IEEE Transactions on Cybernetics*, 2019.

[15] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *SIGKDD*, 2021.

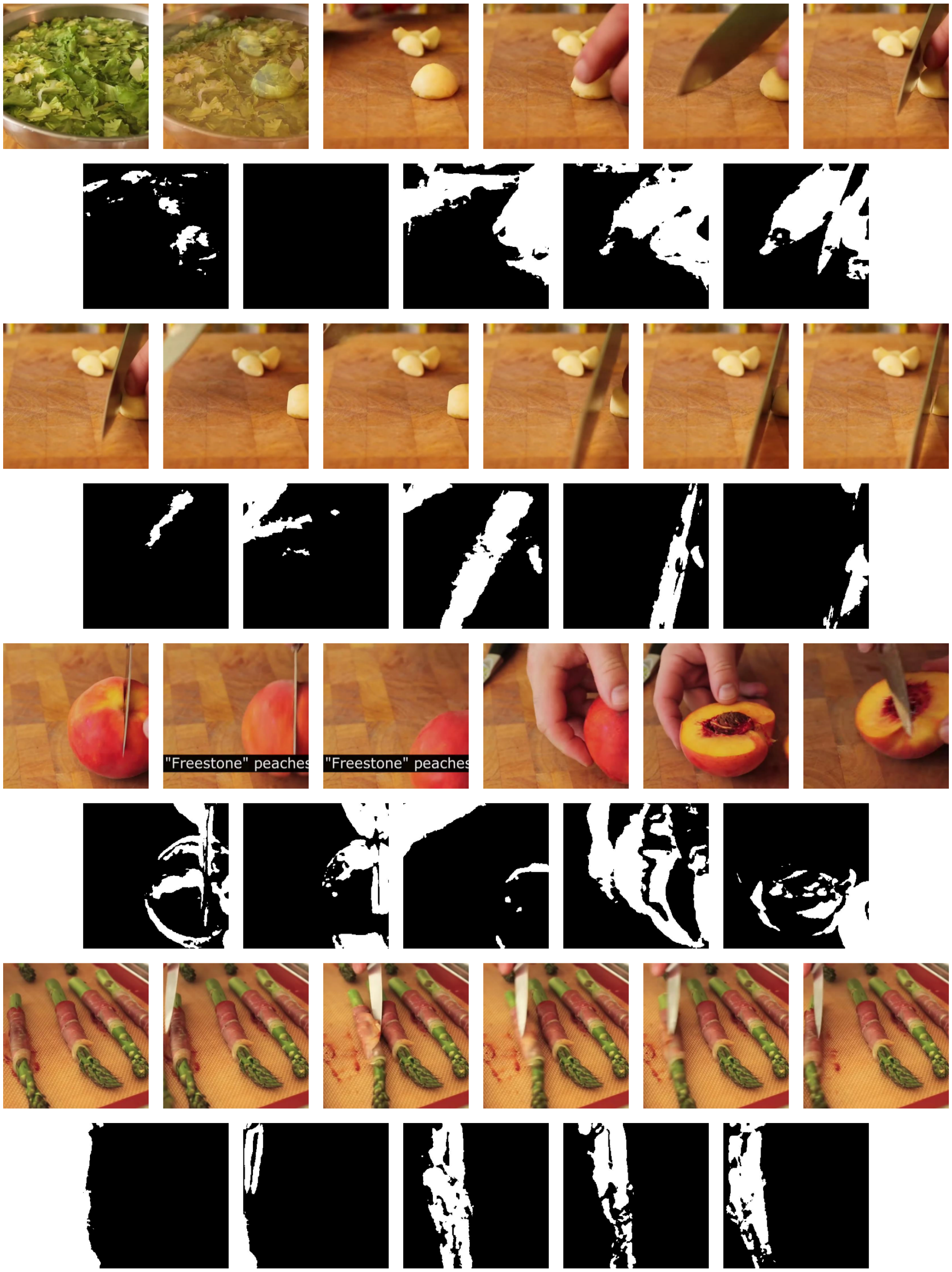


Figure 18: A cluster discovered from the cooking recipe videos. Each row is a different instance of the “chopping” action.