CalCROP21: A Georeferenced multi-spectral dataset of Satellite Imagery and Crop Labels

Rahul Ghosh*
University of Minnesota
Minneapolis, US
ghosh128@umn.edu

Praveen Ravirathinam*
University of Minnesota
Minneapolis, US
pravirat@umn.edu

Xiaowei Jia
University of Pittsburgh
Pittsburgh, US
xiaowei@pitt.edu

Ankush Khandelwal
University of Minnesota
Minneapolis, US
khand035@umn.edu

David Mulla
University of Minnesota
St. Paul, US
mulla003@umn.edu

Vipin Kumar
University of Minnesota
Minneapolis, US
kumar001@umn.edu

* These authors contributed equally

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671569

***Abstract*—Mapping and monitoring crops is a key step towards sustainable intensification of agriculture and addressing global food security. A dataset like ImageNet that revolutionized computer vision applications can accelerate development of novel crop mapping techniques. Currently, the United States Department of Agriculture (USDA) annually releases the Cropland Data Layer (CDL) which contains crop labels at 30m resolution for the entire United States of America. While CDL is state of the art and is widely used for a number of agricultural applications, it has a number of limitations (e.g., pixelated errors, labels carried over from previous years and errors in classification of minor crops). In this work, we create a new semantic segmentation benchmark dataset, which we call CalCROP21, for the diverse crops in the Central Valley region of California at 10m spatial resolution using a Google Earth Engine based robust image processing pipeline and a novel attention based spatio-temporal semantic segmentation algorithm STATT. STATT uses re-sampled (interpolated) CDL labels for training, but is able to generate a better prediction than CDL by leveraging spatial and temporal patterns in Sentinel2 multi-spectral image series to effectively capture phenologic differences amongst crops and uses attention to reduce the impact of clouds and other atmospheric disturbances. We also present a comprehensive evaluation to show that STATT has significantly better results when compared to the resampled CDL labels. We have released the dataset and the processing pipeline code for generating the benchmark dataset.**

***Index Terms*—Remote Sensing, Spatio-temporal data, Semantic Segmentation, Large Scale dataset**

I. Introduction

With the rise in world's population, food supplies must scale up to keep pace with the growing demand. Hence it is critical to ensure that farm lands are being used efficiently from an environmental perspective.In particular, mapping and monitoring crops is a key step towards forecasting yield, guiding sustainable management practices, measuring the loss of productive cropland due to urbanization and evaluating progress in conservation efforts.

Indeed advances in Earth observation technologies have led to the collection of vast amount of accurate and reliable remote sensing data. There are many Land Cover Land Use (LULC) maps that are present for agriculture [12], [15], [20], [21], [23] but only a small subset of these provide crop labels at a pixel level [3], [12], [21]. Of these, the most used is the Cropland Data Layer(CDL). In the United States, the Department of Agriculture's (USDA) Cropland Data Layer (CDL) provides a publicly available land-cover classification map annually at 30m resolution which includes major crop commodities for the conterminous United States (CONUS) [5]. CDL product has driven the advancement of research in areas ranging from agricultural sustainability studies [8], [11], to environmental issues [2], [7], land conversion assessments [18], [22], crop rotations [4], [16], farmer surveys [14] and many more [6]. While CDL is the state-of-the-art spatially explicit identification product for crops, it has a number of limitations [17], [19]. First, the CDL is created using a pixel based classification algorithm and hence contains pixelated errors in crop labels. Second, each pixel is not updated every year and labels for some pixels are not updated from previous years which sometimes leads to incorrect labels. Third, CDL is known to have low accuracy in classifying many minor crops such as alfalfa, hay, tree crops, and many vegetable crops [13]. Finally, CDL labels are created using Landsat images, which are at 30m resolution, leading to mixed pixels errors. The Sentinel constellation provides images at a finer resolution (10m) and more frequent temporal scale (5days vs 15 days) and thus offers the possibility of creating crop labels at 10m resolution.

To overcome the limitations of existing pixel level crop maps and datasets and to facilitate deep learning research in RS-based crop mapping, this paper presents a new semantic segmentation benchmark dataset for crops, CalCROP21. Specifically, the input images were created using a Google Earth Engine based robust image processing pipeline on the multi-spectral temporal images collected by the Sentinel-2 constellation in the Central Valley of California in 2018. A novel spatio-temporal semantic segmentation [10] method was used to generate better quality labels using resampled CDL as initial labels. The efficacy of this methodology relies on several

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

key assumptions. First, the noisy coarse resolution CDL labels are still of good enough quality to be used for training a classifier. Second, a classifier that makes use of space and time is more effective in dealing with the training label noise than one that ignores such information. Third, labels at the geographical farm boundaries can be mixed and their labels at the coarse resolution are not trustworthy, whereas labels at the interior of a region are likely to be more confident.

To summarize, our contributions in this paper are as follows:

- We provide a first large scale semantic segmentation dataset that includes both input images as well as labels for a diverse array of crops at 10m resolution for Central Valley, CA. Specifically, each pixel in the image is labeled as one of 21 crop or 7 non crop classes. Of the 21 crop classes, many are minor crops (Eg, Onions, Garlic) for which existing datasets such as CDL are known to have poor quality labels. To the best of our knowledge this is the first large scale pixel wise crop map that provides labels for minor crops as well.
- We use a novel spatio-temporal deep learning method that makes use of the phenotypic differences among crops at multiple time steps. This method uses resampled CDL labels (that are noisy and are at 30m resolution) and produces higher quality labels at 10m resolution.
- We validate the quality of the labels via a detailed quantitative and qualitative evaluation.
- We provide the processing pipeline code for further use by the community in collecting images and generating results for a different year and using different temporal frequency for any region in the world.

For a full length version of this paper which more figures and description please refer to [9].

II. Related Works

Several benchmark datasets are available for land use and land cover (LULC) mapping. While many datasets are available for LULC there are few that cover agricultural area and even fewer that include pixel wise crop classification, To the best of our knowledge there is no dataset on crop semantic segmentation that includes minor crops. For example, in the context of cropland mapping after CDL (described in Section 1), BigEarthNet [21] is the one most relevant dataset which provides 590,326 image patches from 125 Sentinel-2 tiles and associates each image patch with a subset of 43 Corine Land Cover classes over Europe. However, there are several limitations in this dataset which make it less ideal for crop monitoring. First, the topology of the classes included do not distinguish between different crops but rather between broad vegetation types (forest, agriculture, grassland, etc). Second, due to the association of labels to entire image patch it captures the presence and not the area of a certain category of land cover.

III. Data Sources

We use freely available multi-spectral satellite images and data products to create our dataset. Specifically, we use Sentinel-2 as the input images and the Cropland Data Layer as initial labels. Here we describe the data sources involved in creating the dataset.

A. Input Satellite Imagery

In our dataset, we use the multi-spectral images captured by the two polar-orbiting satellites as part of the Sentinel-2 mission operated by the European Space Agency (ESA). Due to Its high revisit time of 5 days, phenological characteristics of different crops can be observed compared to using single snapshot (or few snapshots) for the whole season. The multi-spectral images has 13 bands in the visible, near infrared, and short wave infrared part of the spectrum, each having a spatial resolution of 10, 20 or 60m. The captured images are available in the form of tiles, each of which have a unique ID and covers an area of 10,000 sq km.

B. Crop Labels

The Cropland Data Layer (CDL) is an annual publicly available land cover classification map for the entire US. With over 200 classes, CDL provides land cover maps covering the entire conterminous United States (CONUS) at 30-meter spatial resolution with a high accuracy up to 95% for classifying major crop types (i.e., Corn, Soybean, and Wheat). The CDL data products are free to download from Google Earth Engine [1]. Although CDL is a very useful product that has led to the development of many downstream applications, the product is plagued with noise that arise due to the reasons discussed in Section 1. In particular, it has high accuracy (up to 95%) for classifying major crops(e.g., Corn, Soybean, Wheat), but it is known to have poor accuracy for minor crops (e.g., Alfalfa, Hay and Tree crops) [13].

IV. Processing Pipeline

We use Google Earth Engine to build a robust image processing pipeline to create biweekly Sentinel image composites. Using the obtained biweekly composites and CDL labels, we develop a novel spatio-temporal deep learning method to improve upon the original CDL labels. In the following, we describe these steps in details.

A. Generation of bi-weekly Sentinel-2 multispectral composites

Many land covers, e.g., different types of crops, are indistinguishable at a single time step. In particular, different crops have different seeding time and harvesting time, which is also affected by weather conditions. Hence, different crops show discriminative signatures at different points of time [10]. Hence, we consider all the images available in a year for this dataset. However, these images often have clouds and other atmospheric disturbances. Here, we generate bi-weekly image composites using a robust Google Earth Engine based pipeline to reduce the impact of these atmospheric disturbances. Specifically, we collect all available images within a 2-week period and score every pixel of each tile using a the quality band (QA60), which presents information as to whether

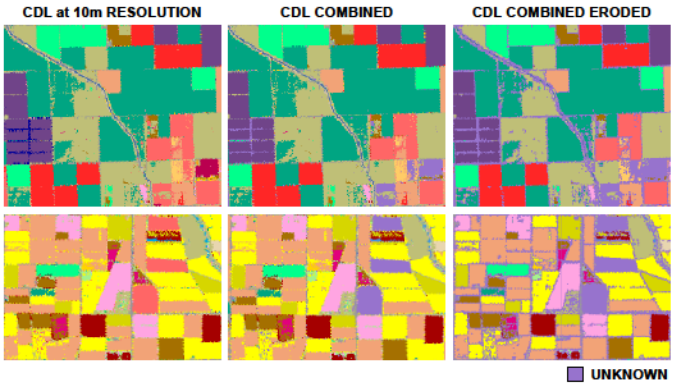

Fig. 1: Evolution of labels through each preprocessing step are shown for two randomly chosen regions (rows). The first columns shows the raw CDL labels resampled at 10m resolution. The second columns shows a revised set of labels, where similar classes have been combined and untrustworthy classes have been grouped as unknown class (in light purple). The third column, which we use for training, shows one level of erosion done classwise at boundaries and also removal of connected components of size less than 4 pixels.

a pixel is cloud-free, dense cloud or cirrus cloud. Based on the amount of cloud-free pixels the collection of images are sorted and finally the best images merged to create a cloud-free mosaic.Thus we obtain 24 mosaics in a year per tile, each in the projection of the zone (for California crop belt, 8 tiles are in EPSG:32610 - WGS 84 / UTM zone 10N and 3 are in EPSG:32611 - WGS 84 / UTM zone 11N). All the bands are resampled to 10m spatial resolution and then exported from Google Earth Engine. Since some Sentinel tile images may be slightly larger than the area they are supposed to cover, we use GDAL to clip the images and reproject them so that every pixel is of 10m$\times$10m resolution.

This finally produces 24 georeferenced files each of which has a shape of (10980,10980,10) for a tile, with 10 signifying the number of bands (10m and 20m) used. Since our objective is to map the entire crop belt in the Central Valley of California, we found that 11 Sentinel-2 tiles covers this crop belt, namely T10SEH, T10SEJ, T10SFG, T10SFH, T10SFJ, T10SGF, T10SGG, T10TEK, T11SKA, T11SKV, and T11SLV, giving a total of 264 tif files. As a preprocessing step we first clip the bottom and top 2%ile of each channel of the satellite images and then apply max-min normalization. Following the preprocessing of the images, we split each tile into 100 grids each of size 10km$\times$10km (1098$\times$1098 pixels). We combine all the 24 composite images corresponding to same grid together to form an array of shape (24,1098,1098,10): 24 timestamps, (1098,1098) pixels and 10 channels. We have 1,100 grid arrays in total, each of which is named as "TILEID_YEAR_ROW_COL_IMAGE.npy", e.g., "T11SKA_2018_5_6_IMAGE.npy" corresponds to the 5th row and 6th column (indexed from 0) of the tile T11SKA in 2018.

B. Pre-processing of CDL

We use Google Earth Engine to fetch the CDL labels and crop them using each georeferenced Sentinel-2 tile, which produces a label image at 30m resolution for each tile. We then resample the labels to 10m resolution to create 11 label tiles of shape (10980,10980). CDL provides labels for more than 200 crop classes, many of which are completely absent or rarely present in the California Central Valley region. In our dataset we exclude these absent classes in the California Central Valley region. In addition, CDL provides state-wise validation metrics for their labels using ground-truth labels. We also exclude those classes for which the number of pixels used for CDL validation is too few as their labels cannot be trusted. Specifically, we include a crop class in our dataset if it fulfils the following conditions: The crop class has at least 1 million pixels in the study region and The crop class has at least 100 validation pixels used by CDL

For non-crop classes we only apply the first condition (e.g., wetlands, grass, forests, hay, urban etc.) as their validation metrics are not provided by CDL.Following these steps we are left with 34 classes: {Corn, Cotton, Rice, Sunflower, Barley, Winter Wheat, Safflower, Dry Beans, Onions, Tomatoes, Cherries, Grapes, Citrus, Almonds, Walnut, Pistachio, Garlic, Olives, Pomegranates, Alfalfa, Hay, Barren, Fallow and Idle, Deciduous Forests, Evergreen forest, Mixed Forests, Clover and wildflower, Shrubland, Grass, Woody wetlands, Herbaceous Wetlands, Water, Urban, Double Crops}. For training and evaluation purposes we combine the different forest classes to a super class "Forest Combined", wetland classes to "Wetlands Combined", and combine {Grass, Shrubland, Clover, Wildflower} to "Grass combined". We also do not use Double Crops in our study and label all those pixels as unknown class. Following the preprocessing of the labels we are left with 21 crop classes and 7 other classes and we refer to this label set as CDL-combined.

Since the CDL is originally at 30m resolution, which we resample to 10m (to match with the resolution of input images), the boundary pixels are mixed and thus they could contain regions of multiple classes. Given the uncertainty of labels at spatial boundaries between any two classes, we perform 1 pixel erosion for each class and replace these eroded pixels with unknown class and remove connected components of a class that are less than or equal to size 4. These labels are called CDL-combined-eroded. Fig 1 shows the progression of the labels through these preprocessing steps. Similar to the image data, post erosion we segment and store the label in arrays of shape (1098,1098) and have the naming convention as TILEID_YEAR_ROW_COL_PREPROCESSED_CDL _LABEL.npy.

C. Grid curation

As described earlier, in our dataset we have 1,100 grids of 1098$\times$1098 pixels in size covering the entire crop belt in California's Central Valley. Many of these grids are predominantly covered by non-crop classes, and hence are removed from the dataset resulting in 367 acceptable grids. Specifically, a grid

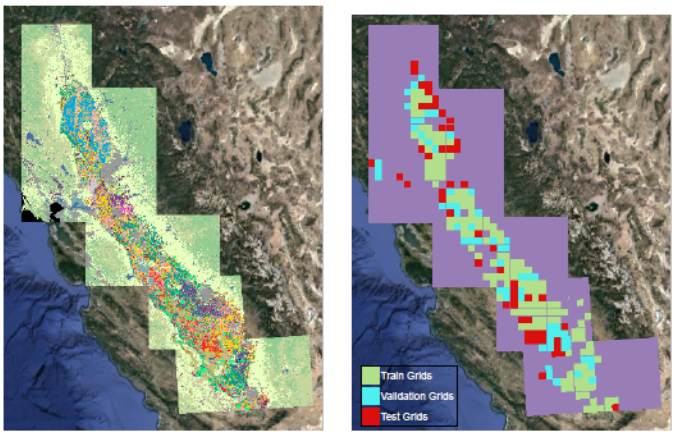

Fig. 2: Overlay of 11 tiles in Central Valley of California region we used in our study along with Distribution of Train/Val/Test grids. Green, light blue, and red represent the regions used for training, validation and testing, respectively. The purple regions denote the non-agricultural land and are not used in our experiments.

is included if it follows both of the following conditions: The grid has at least 50% pixels that are not unknown and out of the valid pixels the grid has at least 50% pixels that belong to crop classes

D. Label Improvement using STATT

As described earlier, CDL based labels cannot be used directly as reference labels. To improve the quality of CDL labels, we used the STATT model proposed by the authors in [10] which uses spatial as well as temporal information to effectively model the phenology of crops and reduce the effects of clouds and other atmospheric disturbances. Specifically STATT uses a UNET style architecture to extract spatial features and a bidirectional Long-Short Term Memory (biLSTM) to model temporal progression of the crop specific growing and harvesting patterns. Further it uses attention networks to aggregate the hidden representations for each time-step based on their contribution to the classification performance. Finally, using these attention scores, the spatial features by the convolutional encoders at multiple resolutions are aggregated and passed using skip connections to the convolutional decoder to generate segmentation maps. A comparison of STATT with alternative approaches that model either the spatial or temporal information, or both (but not as effectively as STATT) is available in [10].

To demonstrate the efficacy of this method in improving the quality of CDL labels, we divided the grids into train, validation and test set. To make sure we create a training set that is balanced amongst classes and is also spread uniformly across space, we adopt a gridwise count based data splitting strategy which sorts grids based on number of crop pixels. With this approach we created a training set of 210 grids, validation set of 84 grids and test set of 73 grids. The color coded final distribution of the sets along with raw labels can be seen in the Fig. 2.

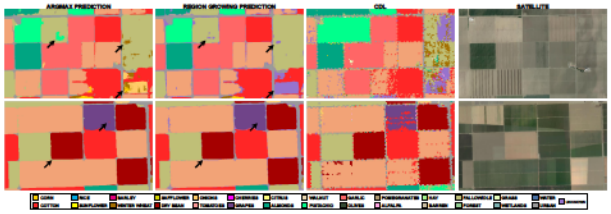

Fig. 3: Argmax predictions vs Region Growing predictions on certain patches in test area which demonstrate the advantage of Region Growing over Argmax. (Arrows represent places of improvement by Region growing over Argmax)

Following the approach as outlined in [10], STATT extracts patches of size 32×32 pixels from the training grids. Using this input patch of size 32×32, we output labels for a patch of size 16×16. For this task, we use three convolutional blocks in our encoder each having two convolutional layers. Thus there are six convolutional layers having 64,64,128,128,256,256 channels and filters of size 3×3. To downsample the output of the convolutional blocks STATT uses max-pooling of size 2×2 after the first and second convolutional blocks. In the decoder, STATT has two convolutional blocks each of which consistsof two convolutional layers. The four convolutional layers of the decoder have 128,128,64,64 channels respectively. To upsample the output we add transposed-convolutional layers before the first and second convolutional block of the decoder having 128,64 channels respectively and kernel size of 2×2. Finally, STATT has a fully-connected layer with input dimension of 64 and output dimension equal to the number of classes i.e. 28. The model was trained using the training dataset for 50 epochs and the validation performance was used as the model selection criteria.

The output of the model are softmax probabilities over the classes for each pixel thus having shape of (16,16,34). By combining all the patches within a grid, we create probability grids of shape (16,16,34). Usually for multi-class classification the decision is made by predicting the class for which the model gives the highest probability. We refer to it as the $argmax$ prediction. In a multi-class classification setting, confusion between classes can easily occur when dealing with a large number of classes. Furthermore, class confusion also happens at the geographical boundary of different classes (e.g. fields with different crops or roads around field).

We use a region growing strategy to post-process the pixel-wise probability outputs instead of directly taking $argmax$ outputs. Specifically, for each class, the pixels that have highest probability value greater than 0.9 are considered as confident anchor pixels. Starting from these anchor pixels, we include all the pixels in their neighborhood which have at least 0.3 probability of belonging to the same class as the anchor pixels. Since the region growing strategy produces class-wise prediction maps, clashes between two or more class at certain pixels are bound to happen, in which case we assign unknown values to those pixels. We observe that majority of such clashes occur near the boundaries which is expected due to the reasons that were described above. As illustrated in

Fig 3, this method is very effective in removing noise within fields and also removing confusion at boundaries by replacing them with "unknown". We store the STATT labels in arrays of shape (1098,1098) and have the naming convention as TILEID_YEAR_ROW_COL_STATT.npy.

E. Final Dataset

In summary, our dataset covers the entire California Central Valley Crop Belt using the 367 grids of cloud filtered multi-spectral images (each in (1098,1098,10)), and we call these image grids. For each image grid, we also provide both the raw and preprocessed CDL grid as well as STATT grid of size (1098,1098). STATT labels are provided for a total of 442,456,668 pixels (ã4,000 sq. km) covering 29 classes, of which 249,946,750 pixels (ã5,000 sq. km) belong to one of the 21 crop classes and the remaining 192,509,918 pixels (ã9,000 sq. km) belong to other 8 classes including unknown.

The entire dataset including Image Grids, CDL grids, pre-processed CDL grids and STATT grids for the acceptable grids as well as the Image grids and CDL grids for the rest of the entire region can be found in the link given below [1].

V. Evaluation

In this section we present the quantitative analysis of the results of our approach on the test regions. We observe that out of the total 57,795,199 pixels, STATT and CDL labels differ in 9,785,767 pixels (16.93%). Focusing only on pixels that are labeled as crop, disagreement drops to 6.97%. Table I shows precision, recall, and F1-scores for all classes while treating pre-processed CDL labels as ground truth. We notice that F1-score is usually high for classes that have high support (see left half of Table I) and usually low for classes that have low support (right half of Table I). As we discuss in the following, STATT labels are generally more accurate than those provided by CDL.

Fig. 4 shows a comparison of the segmentation maps of STATT and the corresponding patch from the CDL layer. In all four triplets, We notice that STATT generally performs much better in detecting boundaries and removing noise. In the first triplet of the first column, we can see how a noisy field is replaced with a smooth prediction of fallow land. In the second we can see how a fallow prediction by CDL is replaced with cotton by STATT and it can be verified in the third image of the triplet, an image in July, that the field cannot be fallow as there is a crop present. In the first triplet of the second column, one can observe removal of erroneous cotton speckles present in the CDL map, and in the final triplet we can see smoothing of multiple fields by the STATT map over the region.

Further, we analyze the pixels where our map does not match with CDL. We have noticed errors in the CDL layer at numerous locations throughout the crop belt which are mainly of two types:

- incorrect labeling of complete (or large parts of) fields

[1]https://drive.google.com/drive/folders/1EnXXRHNoTyIbM-_5p-P9pH4zH3xyTqBp?usp=sharing

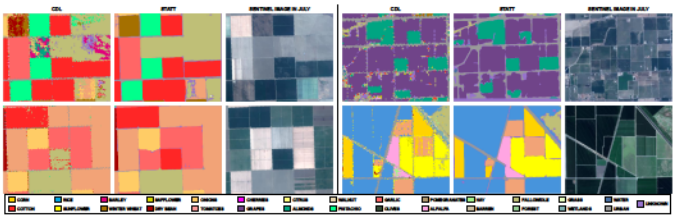

Fig. 4: Segmentation map comparisons on some patches from CDL and STATT in the test regions, Each triplet shown depicts a situation where STATT produces better labels. For description on each triplet please refer to Section V

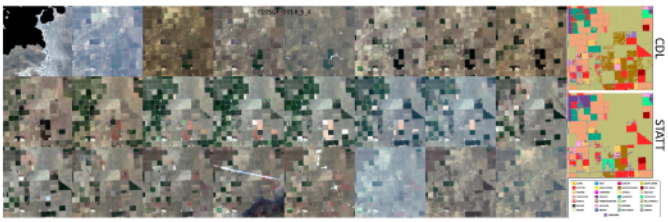

Fig. 5: Visual Analysis of Grid T10SGF_2018_5_4. One can observe from the visual images that many fields in this region are fallow throughout the year but CDL labels them as Winter Wheat (such as the field next to the triangular Cotton (red) field). However STATT does not make these mistakes and correctly labels the region as Fallow land.

- spatially discontinuous label prediction (i.e., label of a pixel differs from its surrounding pixels that all belong to a different class)

In the next few paragraphs we systemically discuss and analyze these cases on some of the fields within California Crop belt.

A. Visual analysis of a sample of patches

The first way to compare our dataset and CDL is to visually inspect images over time, check the growing time, rate of greenness and harvest time to assign a label to the field, and then check whether CDL or STATT is correct. Although this method cannot be scaled to every field due to substantial manual effort and expertise needed regarding crop growing patterns, we can still use this approach to verify disagreement between CDL and STATT where one predicts fallow and idle land and the other predicts a crop. Since no crop is grown year around in a fallow or idle land, it should be easy to verify the correct prediction. We observed numerous cases in the California Central Valley crop belt where CDL predicted a crop and STATT predicted fallow land. An example of this can be seen in Fig. 5. In this grid, We can observe numerous fields as fallow throughout the year (such as the one one next to the triangle shaped cotton field) but CDL labels them as crop (the field next to the triangle cotton field is labelled as winter wheat by CDL). Through this first method of visual analysis, we were able to verify many cases of fallow vs crop disagreement, in which the STATT prediction of fallow seems appropriate.

TABLE I: Precision, Recall, and F1-Score of STATT labels in the test region with CDL as groundtruth. We also mention support(in pixels), that is the count classwise of CDL labels used during evaluation (10000 pixels equals 1 sq.km).

STATT									
CLASS	Precision	Recall	F1-score	Support	CLASS	Precision	Recall	F1-score	Support
Fallow and Idle	0.6587	0.8069	0.7253	8448934	Safflower	0.8965	0.5623	0.6911	682725
Almonds	0.9172	0.9107	0.9139	8336706	Hay	0.3869	0.2360	0.2932	644262
Rice	0.9668	0.9941	0.9803	6920141	Wetlands Combined	0.7462	0.2604	0.386	396679
Grass Combined	0.7742	0.7109	0.7412	6579739	Garlic	0.9214	0.9079	0.9146	332639
Walnut	0.9167	0.8050	0.8572	3469078	Barren	0.7746	0.0735	0.1343	272560
Cotton	0.9730	0.9691	0.9710	3382440	Sunflower	0.9586	0.8479	0.8999	261263
Urban	0.8265	0.8248	0.8257	3330554	Onions	0.5718	0.7729	0.6573	260724
Grapes	0.7663	0.8952	0.8258	2855088	Pomegranates	0.8164	0.4295	0.5628	219041
Pistachio	0.9065	0.8528	0.8788	2665781	Olives	0.4636	0.6974	0.557	214675
Tomatoes	0.9018	0.9383	0.9197	2580987	Forests Combined	0.0000	0.0000	0.0000	211352
Winter Wheat	0.8753	0.5944	0.7080	1927378	Citrus	0.8559	0.8382	0.8469	173361
Alfalfa	0.7422	0.8708	0.8013	1576520	Barley	0.9435	0.0886	0.162	172028
Water	0.9300	0.9843	0.9564	935492	Dry Beans	0.8227	0.7487	0.7839	126070
Corn	0.6819	0.8472	0.7557	716901	Cherries	0.6550	0.4440	0.5293	102081
OVERALL	Precision	Recall	F1-score	Support	CROPS ONLY	Precision	Recall	F1-score	Support
MEAN	0.7732	0.6754	0.6885	57795199	MEAN:	0.8067	0.7262	0.7386	37,619,889
Weighted MEAN	0.834	0.8307	0.8251	57795199	Weighted MEAN	0.8882	0.8699	0.8731	37,619,889
ACCURACY:			0.8307		ACCURACY			0.9303	

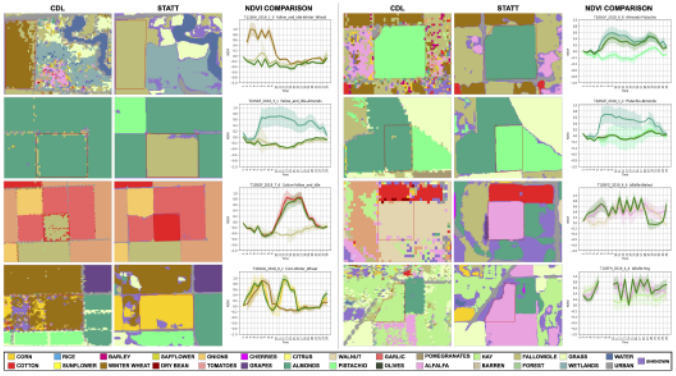

Fig. 6: Comparison of NDVI for some fields with disagreeemnt between CDL labels and STATT labels. Each triplet denotes a case where the labels produced by STATT are better than CDL, as the series of the field (represented in Green) lies closer to the series of the class denoted by STATT.

B. Visual analysis of a sample of NDVI time series

Here we resolve the disagreement between STATT and CDL labels by analyzing the NDVI time series of the field in question (i.e. a continuous region where STATT and CDL differs). In each disagreement there are two classes in analysis, the CDL prediction class and the STATT prediction class. If the NDVI series of field lies closer to the characteristic NDVI series of the STATT prediction class than the CDL prediction class then we can say that the field is actually the STATT prediction class and vice versa. Now the question arises, how do we obtain this characteristic NDVI series of different classes in our dataset? To get the characteristic NDVI series we take the median (timestamp wise) of the NDVI series for pixels agreeing with the class of interest in the grid where the field of interest is located. What we mean by pixels of agreement are those pixels where CDL and STATT agree, i.e predict the same class at that pixel. we use only agreement pixels within the grid of the field because we found that across grids crops have different NDVI series due to local farmer practices, weather conditions and cloud cover patterns.

We now plot the characteristic NDVI series for both the CDL prediction class and the STATT prediction class on the same graph. We then plot the median NDVI series timestamp wise of all the pixels in the field of interest on the same graph and check which characteristic NDVI series it lies closer to. If the NDVI series of the field is closer to the characteristic series of the class labeled by STATT compared to the characteristics series of the class labeled by CDL, then we can say that STATT label was correct, and vice versa. We found that in a vast majority of cases, whenever there is a field of disagreement, the NDVI series of the field lies closer to the STATT prediction class signature than the CDL prediction class signature. Fig. 6 shows 8 triplets for some fields where we conducted this method of analysis. The first image in the triplet is the CDL prediction and the second image is the STATT prediction, and in each of these images there is a red boundary denoting the field of interest. One can observe that in all the triplets, the predicted class within the field of interest (i.e the red boundary) differs between the CDL image and the STATT image. The third image is the NDVI plot of the three timeseries described before, i.e the CDL prediction class characteristic NDVI series (denoted by the plot in the color of the CDL prediction class), the STATT prediction class characteristic NDVI series (denoted by the plot in the color of the STATT prediction class) and the NDVI series of the field of interest (denoted by the signature in the green color). All examples in Fig. 6 show that the green line lies closer to the STATT class NDVI line thus showing superiotiy over CDL labels.

C. Comprehensive analysis of all pixels where STATT and CDL disagree

The previous two methods prove to be very useful while doing field analysis and using a combination of the two we can show for each field who is correct. However, neither of these methods give a global perspective nor do they quantify how much better STATT's predictions are when compared to those

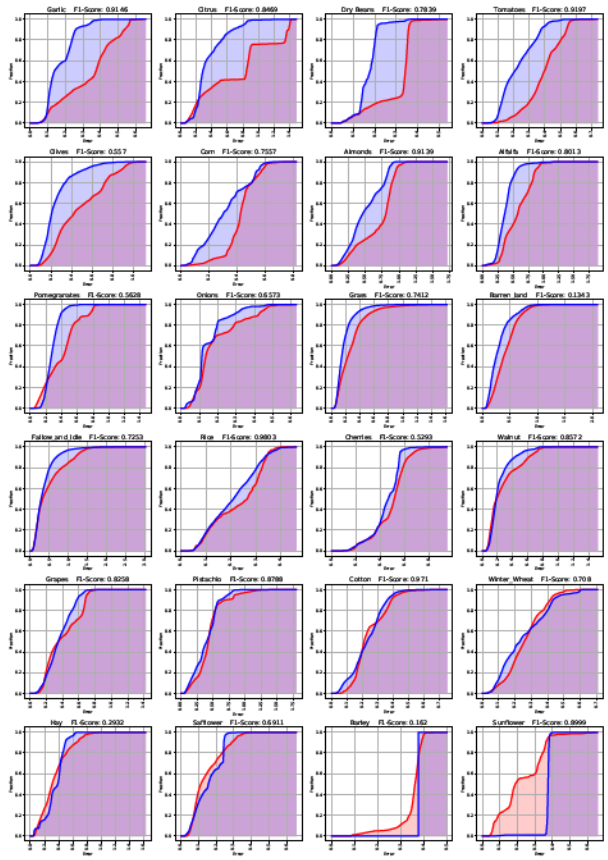

Fig. 7: Class wise Area under the curve plots. The x axis represents the NMSE and the y axis represents the $Score(\cdot)$ value for the corresponding NMSE from x axis. The red and blue curves represent CDL and STATT, respectively. We also mention F1-Score of that class in each plot.

of CDL's predictions. To address this issue, we use a third method of analysis in which we devise a function to measure closeness of pixels to ground reality, and after plotting the function, use area under the curve to establish which strategy, i.e. CDL or STATT, is better.

After obtaining the map for STATT and CDL, we calculate the characteristic NDVI series gridwise for each class using the agreement pixels as described in the previous section. Now we consider a characteristic series to be valid only if there are at least T pixels in agreement in the grid, and T is set to 100 in this work. Now for each pixel of disagreement for each class we calculate the Normalised Mean Square Error (NMSE) with the characteristic NDVI series and the NDVI series of the the pixel of disagreement. For each strategy (CDL or STATT), we first sort all the disagreement pixels according to their NMSE. Then we compute $Score(E)$ for each strategy, which is defined to be the proportion of disagreement pixels with NMSE less than a particular error E over all the disagreement pixels, i.e., $Score()$, for a particular error (E) as follows:

$$Score(E) = \frac{\#\ disagreement\ pixels\ in\ strategy\ with\ NMSE\ less\ than\ E}{Total\ No.\ of\ pixels\ of\ disagreement\ in\ strategy} \quad (1)$$

where $strategy$ represents either CDL or STATT.

The function Score represents the fraction of total disagreements pixels whose NMSE lies below a set threshold error denoted by E. The notion behind this function is that, the closer the NMSEs of the disagreement pixels are to zero, the faster Score rises as E rises. At the max error, the Score will be 1, as all NMSEs lie below the threshold. Our hypothesis is that the STATT disagreement pixels have lower errors and so Score will rise faster for STATT when compared to CDL. As a result, STATT will reach a higher Score faster and will thus have more area under the plot of Score until the max error. A plot of this function is constructed for each class with E starting from 0 and ranging up til maximum NMSE error recorded for that class, which we denote as E_{max}. which could come from a CDL disagreement pixel or a STATT disagreement pixel We then calculate Area under the curve as follows:

$$Area_{strategy} = (\int_0^{E_{max}} Score(E)\, dE)/E_{max} \quad (2)$$

We divide the area of each plot by E_{max} to keep it within the range (0,1).The plots of curve for each class can be seen in Fig. 7, with Blue representing STATT and red representing CDL. We can see that STATT has a higher area when compared to CDL in almost all the classes. We also see from the figure that in a lot of classes the blue line lies above the red line throughout the plot. This experiment solidifies our claim that STATT labels are closer to the ground reality than when compared to the labels provided by CDL.

VI. Conclusion

In this paper we presented CalCROP21, a georeferenced data set for a diverse array of crops grown in the Central Valley of California. This dataset contains multi spectral Sentinel imagery along with crop labels at 10m resolution for year 2018 that are derived using a novel spatial-temporal deep learning method that makes use of noisy CDL labels available at 30m resolution. Our extensive analysis of this dataset demonstrates the superiority of our dataset over CDL. We have also released our processing pipeline and associated datasets that can be used by the community to generate crop labels for other years and for creating similar data sets for other parts of US. We anticipate this dataset will catalyze the innovation in machine learning research on remote sensing data (e.g., classifying multiple imbalanced classes and modeling heterogeneous data over space), and also enable the use of this information for studying crop distribution and its implications by the agricultural community.

VII. Acknowledgement

This work was funded by the NSF awards 1838159 and 1739191. Rahul Ghosh is supported by UMII MNDrive Graduate Fellowship. Access to computing facilities was provided by the Minnesota Supercomputing Institute.

REFERENCES

[1] Usda nass cropland data. https://developers.google.com/earth-engine/datasets/catalog/USDA_NASS_CDL, 2021.

[2] Jason B Belden, Brittany Rae Hanson, Scott T McMurry, Loren M Smith, and David A Haukos. Assessment of the effects of farming and conservation programs on pesticide deposition in high plains wetlands. *Environmental science & technology*, 46(6):3424–3432, 2012.

[3] Mariana Belgiu and Ovidiu Csillik. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sensing of Environment*, 204:509–523, 2018.

[4] Claire Boryan, Zhengwei Yang, and Liping Di. Deriving 2011 cultivated land cover data sets using usda national agricultural statistics service historic cropland data layers. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 6297–6300. IEEE, 2012.

[5] Claire Boryan, Zhengwei Yang, Rick Mueller, and Mike Craig. Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358, 2011.

[6] USDA cropland data layer. https://www.nass.usda.gov/Research_and_Science/Cropland/docs/MuellerICASVI_CDL.pdf, 2021.

[7] R Cibin, I Chaubey, and B Engel. Simulated watershed scale impacts of corn stover removal for biofuel on hydrology and water quality. *Hydrological processes*, 26(11):1629–1641, 2012.

[8] Timothy Fitzgerald and Grant Zimmerman. Agriculture in the tongue river basin, output, water quality, and implications. 2013.

[9] Rahul Ghosh, Praveen Ravirathinam, Xiaowei Jia, Ankush Khandelwal, David J. Mulla, and Vipin Kumar. Calcrop21: A georeferenced multi-spectral dataset of satellite imagery and crop labels. *CoRR*, abs/2107.12499, 2021.

[10] Rahul Ghosh, Praveen Ravirathinam, Xiaowei Jia, Chenxi Lin, Zhenong Jin, and Vipin Kumar. Attention-augmented spatio-temporal segmentation for land cover mapping, 2021.

[11] Laura Hartz, Fritz Boettner, and Jason Clingerman. Greenbrier valley local food: The possibilities and potential. *Greenbrier Valley Economic Develeopment Corporation*, 2011.

[12] I. Hernandez, Pedro Benevides, Hugo Costa, and Mário Caetano. Exploring sentinel-2 for land cover and crop mapping in portugal. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B3-2020:83–89, 08 2020.

[13] Tyler J Lark, Ian H Schelly, and Holly K Gibbs. Accuracy, bias, and improvements in mapping crops and cropland across the united states using the usda cropland data layer. *Remote Sensing*, 13(5):968, 2021.

[14] Kathleen Painter, Hilary Donlon, Stephanie Kane, et al. Results of a 2012 survey of idaho oilseed producers. *AE Extension Series-Department of Agricultural Economics and Rural Sociology, University of Idaho*, (13-01), 2013.

[15] Otávio A. B. Penatti, Keiller Nogueira, and Jefersson A. dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 44–51, 2015.

[16] James D Plourde, Bryan C Pijanowski, and Burak K Pekin. Evidence for increased monoculture cropping in the central united states. *Agriculture, ecosystems & environment*, 165:50–59, 2013.

[17] Joshua Pritsolas and Randall Pearson. A cautionary tale: A recent paper's use of research based on the usda cropland data layer to assess the environmental impacts of claimed cropland expansion.

[18] Benjamin S Rashford, Shannon E Albeke, and David J Lewis. Modeling grassland conversion: Challenges of using satellite imagery data. *American Journal of Agricultural Economics*, 95(2):404–411, 2013.

[19] Kurtis D Reitsma, David E Clay, Sharon A Clay, Barry H Dunn, and C Reese. Does the us cropland data layer provide an accurate benchmark for land-use change estimates? *Agronomy Journal*, 108(108):226, 2015.

[20] R. Saini and S. K. Ghosh. Crop Classification on Single Date SENTINEL-2 Imagery Using Random Forest and Suppor Vector Machine. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 425:683–688, November 2018.

[21] Gencer Sumbul, Marcela Charfuelan, Begum Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Jul 2019.

[22] Christopher K Wright and Michael C Wimberly. Recent land use change in the western corn belt threatens grasslands and wetlands. *Proceedings of the National Academy of Sciences*, 110(10):4134–4139, 2013.

[23] Zhiwei Yi, Li Jia, and Qiting Chen. Crop classification using multi-temporal sentinel-2 data in the shiyang river basin of china. *Remote Sensing*, 12(24), 2020.