ISPRS Journal of Photogrammetry and Remote Sensing 216 (2024) 66–89

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

CropSight: Towards a large-scale operational framework for object-based crop type ground truth retrieval using street view and PlanetScope satellite imagery

Yin Liu [a], Chunyuan Diao [a,*], Weiye Mei [b], Chishan Zhang [a]

[a] Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[b] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Keywords:
Crop type ground truth
Street view imagery
PlanetScope
Deep learning
Uncertainty

ABSTRACT

Crop type maps are essential in informing agricultural policy decisions by providing crucial data on the specific crops cultivated in given regions. The generation of crop type maps usually involves the collection of ground truth data of various crop species, which can be challenging at large scales. As an alternative to conventional field observations, street view images offer a valuable and extensive resource for gathering large-scale crop type ground truth through imaging the crops cultivated in the roadside agricultural fields. Yet our ability to systematically retrieve crop type labels at large scales from street view images in an operational fashion is still limited. The crop type retrieval is usually at the pixel level with uncertainty seldom considered. In our study, we develop a novel deep learning-based CropSight modeling framework to retrieve the object-based crop type ground truth by synthesizing Google Street View (GSV) and PlanetScope satellite images. CropSight comprises three key components: (1) A large-scale operational cropland field-view imagery collection method is devised to systematically acquire representative geotagged cropland field-view images of various crop types across regions in an operational manner; (2) UncertainFusionNet, a novel Bayesian convolutional neural network, is developed to retrieve high-quality crop type labels from collected field-view images with uncertainty quantified; (3) Segmentation Anything Model (SAM) is fine-tuned and employed to delineate the cropland boundary tailored to each collected field-view image with its coordinate as the point prompt using the PlanetScope satellite imagery. With four agricultural dominated regions in the US as study areas, CropSight consistently shows high accuracy in retrieving crop type labels of multiple dominated crop species (overall accuracy around 97 %) and in delineating corresponding cropland boundaries (F1 score around 92 %). UncertainFusionNet outperforms the benchmark models (i.e., ResNet-50 and Vision Transformer) for crop type image classification, showing an improvement in overall accuracy of 2–8 %. The fine-tuned SAM surpasses the performance of Mask-RCNN and the base SAM in cropland boundary delineation, achieving a 4–12 % increase in F1 score. The further comparison with the benchmark crop type product (i.e., cropland data layer (CDL)) indicates that CropSight is a promising alternative to crop type mapping products for providing high-quality, object-based crop type ground truth of diverse crop species at large scales. CropSight holds considerable promise to extrapolate over space and time for operationalizing large-scale object-based crop type ground truth retrieval in a near-real-time manner.

1. Introduction

Food security will be increasingly challenged in the upcoming decades with ongoing climate change and population growth. A range of agricultural policies have been designed to address the challenge through optimizing crop management practices (e.g., crop rotation and tillage) and enhancing crop productivity (Bennett et al., 2012). Crop type maps are critical in informing these agricultural policy decisions by providing crucial data on the specific crops cultivated in a given area (Schmedtmann and Campagnolo, 2015; Som-ard et al., 2022). Currently, large-scale crop type maps could be efficiently generated with the remote sensing technologies and machine learning/deep learning classification models (Belgiu and Csillik, 2018; Vuolo et al., 2018; Cai et al., 2018; Griffiths et al., 2019; Oliphant et al., 2019; Dakir et al.,

* Corresponding author.
E-mail addresses: yinl3@illinois.edu (Y. Liu), chunyuan@illinois.edu (C. Diao).

https://doi.org/10.1016/j.isprsjprs.2024.07.025
Received 2 March 2024; Received in revised form 25 June 2024; Accepted 26 July 2024
Available online 1 August 2024

2020; Jia et al., 2021; Tran et al., 2022; Blickensdörfer et al., 2022). As the training data for these classification models, crop type ground truth plays an essential role in crop type mapping by providing labeled examples of various types of crop species. Such crop labels enable classification models to make informed crop type predictions over wide geographical regions through characterizing distinct remote sensing features (e.g., crop phenology patterns, tillage practices, and harvest time) associated with each crop species. The large-scale acquisition of crop type ground truth is imperative for ensuring the quality of crop type mapping as well as advancing a diversity of subsequent agricultural applications (e.g., crop phenology and condition monitoring, management practice optimization, and crop yield estimation).

Field survey is the most traditional crop type ground truth collection method. It can record specific crop types planted in the fields. Yet, conducting field surveys over large-scale geographical regions is labor-intensive, time-consuming, and costly. As an alternative, historical crop type maps are utilized as substitutes for ground truth data to train the crop type classifier, with samples drawn from these maps (e.g., cropland data layer (CDL) of US (Cai et al., 2018; Wang et al., 2019; Johnson and Mueller, 2021; Lin et al., 2022; Zhang et al., 2022), northeast China crop type map (Di Tommaso et al., 2021), crop map of England (CROME) (Luo et al., 2022), carta uso agricolo (CUA) of Italy (Gallo et al., 2023)). Crop type maps could provide abundant historical crop type "ground truth" across agricultural areas for characterizing corresponding crop species distributions and patterns over extended regions and years. However, the accuracy of such "ground truth" varies spatially and temporally, depending on the quality of the field-based ground truth and remote sensing observations, as well as the classification methods employed in generating crop type maps. Additionally, the availability of "ground truth" is limited by the release timing of the crop type maps. The existing crop type maps are typically generated using year-round remote sensing observations to capture the full range of crop growing phenology characteristics (Yang et al., 2023). Consequently, the "ground truth" for the current year can only be retrieved in the following year, posing a delay and challenge to crop mapping and subsequent tasks.

Another promising method of retrieving crop type ground truth is using street view imagery (e.g., Google Street View and Baidu Total View). With coverage spanning over half of the world's populated regions (Goel et al., 2018), street view imagery provides a valuable large-scale image source for various data collection tasks (e.g., harvesting date estimation (Jiang et al., 2024), land use classification (Cao et al., 2018), real estate valuation (Xu et al., 2022), and pedestrian count collection (Yin et al., 2015)). By visually identifying the crop plants cultivated in the roadside agricultural fields, crop type ground truth labels could be directly extracted from the street view imagery, which replaces actual field observations required in field surveys with virtual audits (Fatchurrachman et al., 2022; Hu et al., 2022). Compared to acquiring "ground truth" from crop type maps, street view imagery-based methods largely reduce the issues of varying crop type labeling accuracy and latency over space and time. Street view imagery facilitates the reliable identification of crop types with embedded detailed visual features (e.g., crop morphology and crop leaves) from a human (horizontal) viewpoint, which cannot be provided by other commonly used data sources such as top-down view aerial or satellite imagery (Biljecki and Ito, 2021). Street view imagery also enables the prompt acquisition of crop type ground truth labels during the growing season. Upon the online availability of cropland-related street view images, the direct identification of crop types becomes feasible, allowing for the timely collection of ground truth data.

To streamline the collection of crop type ground truth from street view imagery, deep learning models (e.g., residue network and inception v3 network) have been widely utilized in identifying crop types of the imagery (Ringland et al., 2019; Wu et al., 2021; Paliyam et al., 2021; Yan and Ryu, 2021; d'Andrimont et al., 2022). Yet street view images (e.g., Google Street View images) are typically collected in a huge amount and encompass diverse landscapes of varying characteristics (e.g., urban areas, forests, agricultural fields, and water bodies). To efficiently train a deep learning classification model in large scale crop type retrieval, a more focused approach on only agricultural street view images could be utilized to reduce the complexities and confusion from non-agricultural landscapes. However, even with agricultural landscapes as the focus, the presence of non-crop elements in street view images such as weeds and obstructions may still hinder the capability of deep learning models in crop type identification. The timing of collecting agricultural street view images also introduces complexity in crop identification. The images may depict either bare land or lands with cover crops during non-growing seasons. Addressing these issues necessitates a systematic approach to the collection and processing of street view images for crop type labeling. However, the development of such a targeted and operational method for large-scale street view imagery-based crop type ground truthing remains underexplored in current research.

The accuracy of crop type information retrieved by deep learning models is also closely tied to the quality of street view images. A range of factors, such as varying lighting conditions, occlusions, viewing angles, or distances from the crop, can introduce ambiguities or uncertainties to crop type ground truthing from street view imagery (d'Andrimont et al., 2018; Hou and Biljecki, 2022). The identification of crop types from street view imagery can further be complexed by crop morphological or structural variations due to seasonal changes as well as visual similarities among crop species. To effectively tackle these complexities, quantifying the uncertainty and the confidence level of each street view image prediction becomes urgently needed in crop type ground truthing (Abdar et al., 2021b; Yordanov et al., 2023). With the measure of uncertainty, low-confidence predictions could be filtered out to avoid wrongly assigning the same label to different crop types. Such uncertainty assessment is crucial for providing reliable crop type labels for downstream tasks, yet has rarely been considered in current studies.

The crop type information retrieved from geo-tagged street view images is mostly pixel-based with a singular coordinate point denoting a crop ground truth location (Yan and Ryu, 2021; Laguarta et al., 2024). Object-based crop type ground truth, encompassing both crop type and associated cropland boundaries, serves as a crucial spatial unit for policy support in crop monitoring for farmers' subsidies, yet it remains unexplored. Compared to pixel-based crop type ground truth, object-based ground truth has enriched agricultural field characteristics and exhibited enhanced classification performance in crop type mapping tasks (Ok et al., 2012; Kussul et al., 2016). This is due to the fact that object-based ground truth could characterize agricultural fields more thoroughly and diversely, with both within-object (e.g., spectral, texture, and shape characteristics) and between-object information (e.g., connectivity, contiguity, distances, and directional relationships among adjacent objects) represented (Zhang et al., 2018). In addition, when mapping large-scale crop type distributions using remote sensing technique, the aggregation of the remote sensing information from multiple pixels within object-based agricultural boundaries could help alleviate cloud cover issues of crop fields with potentially more available remote sensing observations (Cai et al., 2018). Therefore, retrieving both crop types from street view images and associated cropland boundaries becomes vital in enhancing the value of collected ground truth.

The overarching goal of this study is to develop a large-scale operational framework, named CropSight, to retrieve the rich crop type ground truth information at the object-based cropland level. CropSight leverages the Google Street View (GSV) images and PlanetScope satellite images to scalably retrieve crop types along with their corresponding cropland boundaries. Focusing on dominant crop species in four designated study areas across the United States, our study encompasses three specific goals: (1) design a systematic operational approach to collect enhanced cropland field-view images from extensive GSV database for subsequent crop type labeling; (2) develop an uncertainty-aware image classification model, UncertainFusionNet, to retrieve crop types from GSV imagery and quantify associated uncertainty (3) devise a fine-tuned

Segmentation Anything Model (SAM) to automate the extraction of cropland boundary corresponding to each GSV crop type image from PlanetScope imagery. CropSight's performance is evaluated via three perspectives, including its capability in crop type classification from field-view images, its effectiveness in cropland boundary delineation with PlanetScope images, and the reliability of its collected crop type data relative to established benchmark products (i.e., CDL).

2. Study area and data

2.1. Study area

Our study encompasses four unique agricultural regions (i.e., Illinois, Southern Midwest, Texas, and California) in the US, each with different dominant crop species (Fig. 1 and Table S1). In Illinois (study area A), the agricultural landscape is dominated by corn and soybean, occupying more than 95 % of the cultivated land. These two primary crops, typically rotated year on year, serve as the study crops for this region. California (study area B) has diverse agricultural production owing to its unique combination of climate and geography. In light of this agricultural diversity, we select more crops including almond, corn, rice, wheat, grape, and pistachio. Cotton and tomato are excluded from our analysis due to insufficient GSV data for these crop species (Table S1). In southern Texas (study area C), cotton, corn, and sorghum are selected as the study crops because they are the three leading crop species, collectively accounting for over 90 % of the cropland in this region. The Southern Midwest (study area D) is dominated by corn, soybean, rice, and cotton, collectively making up over 90 % of its cropland, and these four crops are selected as the focus of our study in this region. The four study areas exhibit large variations in dominant crop species, reflecting the diversity in climate, soil conditions, and water availability across regions. This diversity in regional crop species makes these areas suitable for evaluating our devised crop type ground truth retrieval framework CropSight at large scales. The study period spans from 2013 to 2022, based upon the availability and quality of GSV images.

2.2. Data

To date, the most comprehensive and easily accessible repository of street view imagery is Google Street View (GSV), published by Google Maps (Anguelov et al., 2010). Since its launch in 2008, GSV has effectively covered numerous countries, offering a vast collection of street-level images spanning a long time period. The extensive and open access nature of GSV makes it an invaluable resource for gathering ground truth data across large geographical regions. Specifically, GSV images provide panoramic views of surroundings, captured by cameras mounted on vehicles as they pass through roads. Each GSV panoramic image is accompanied by rich metadata including the Pano ID, Heading, Latitude, Longitude, Month, and Year. The Pano ID is a unique identifier for each panorama. The Heading refers to the direction that the vehicle with a camera faces when it takes the panorama image. The Latitude, Longitude, Month, and Year provide geographic and temporal information when each panorama is captured. These metadata are vital for optimizing the selection and preprocessing of panoramic images, subsequently aiding in the identification of crop types from these images. GSV metadata and panoramas are retrieved and downloaded from

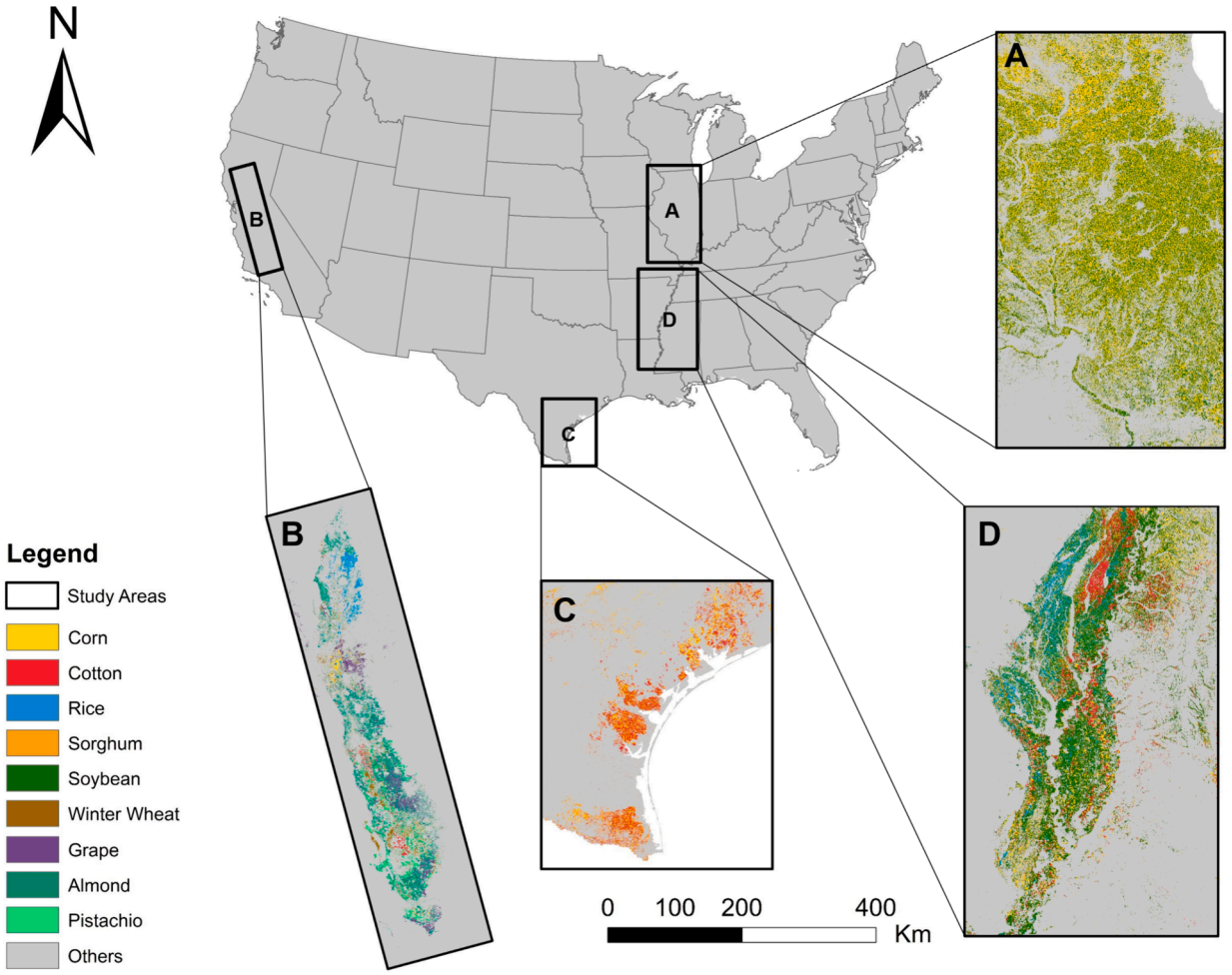

Fig. 1. Four distinct study areas (A-D) situated in key production hotspots across the US, each characterized by different dominant crop types.

Google Maps Platform (API: https://maps.googleapis.com/maps/api/streetview/).

In order to facilitate the crop type ground truth collection process, pinpointing the GSV images with clear view of agricultural land is critical. We utilize the WorldCover product to locate specific GSV points within the GSV metadata database that have a higher likelihood of showcasing agricultural landscapes. The WorldCover is a pioneering global land cover product for 2020 and 2021 at 10 m resolution generated through Sentinel-1 and Sentinel-2 satellite data (Zanaga et al., 2022), which is suitable for application on a global scale. It provides 11 land cover classes including tree cover, shrubland, grassland, cropland, built-up, bare/sparce vegetation, snow and ice, permanent water bodies, herbaceous wetland, mangroves, and moss and lichen. In our framework, we harness the 'cropland' and 'tree cover' category to identify appropriate agricultural GSV images at four study areas.. We also utilize road network geospatial data from OpenStreetMap to help target the GSV images with high potential capturing clearer views in the roadside croplands. Additionally, we utilize the USDA's Crop Progress Reports (CPRs) to pinpoint local time windows in which there is a higher likelihood of capturing images of crops during the growing season in the local fields. These CPRs offer weekly cumulative percentages of major crops reaching specified phenological stages. GSV images collected outside of the time window derived from CPRs are considered to capture the barren landscape during the off-season and are excluded. The time window is determined from the earliest recorded date for planting to the latest recorded date for harvesting of local crop species. Given that the almond and pistachio crops in study area B are not included in the CPRs, the growing window for these species is established from April to September, in accordance with the local climate conditions in California (Bellvert et al., 2018). Alternatively, satellite time series data on croplands could be employed to monitor crop calendars and establish the optimal timing windows in regions where CPRs are unavailable.

After collecting the target cropland field-view images, the PlanetScope satellite imagery is leveraged to extract the boundary information corresponding to each crop type image. PlanetScope's high spatial resolution (3-m) ensures the visibility of crop canopy features to detect the boundaries of crop fields. Its high temporal resolution (2–3 days) guarantees the availability of high-quality satellite imagery within the month when the GSV image is collected. Its low latency (1 day) ensures rapid access to PlanetScope satellite imagery, facilitating near-real-time delineation of cropland boundaries.

Cropland Data Layer (CDL) is used as a benchmark product for the evaluation of CropSight given its extensive coverage across the US and its high accuracy. As one of the most influential crop type products of US, CDL has been widely used as the large-scale crop type "ground truth" for various agricultural and land use applications (Cai et al., 2018; Wang et al., 2019; Johnson and Mueller, 2021; Lin et al., 2022; Zhang et al., 2022). CDL provides 30-m pixel-based crop type data by harnessing satellite time series combined with agricultural field surveys and other ancillary data, processed through machine learning classifiers (Boryan et al., 2011). Complementing its crop type information, CDL also provides a confidence byproduct, which reflects the confidence level of each pixel's crop type classification.

3. Methodology

CropSight is an operational framework that harnesses GSV and PlanetScope imagery to remotely collect object-based crop type ground truth information (Fig. 2). It comprises three key components: large-scale operational cropland field-view imagery collection method (Section 3.1.1), uncertainty-aware crop type image classification model (UncertainFusionNet) (Section 3.1.2), and cropland boundary delineation model (SAM) (Section 3.1.3). The cropland field-view imagery collection method is designed to collect geotagged field-view GSV images that are suitable for crop type ground truth retrieval at large scales in an operational fashion. The UncertainFusionNet model is a novel Bayesian convolutional neural network developed to identify crop types in the acquired geotagged field-view images with uncertainty quantified. It integrates two state-of-the-art image classification models, vision transformer (ViT-B16) and residual neural network (ResNet-50), to

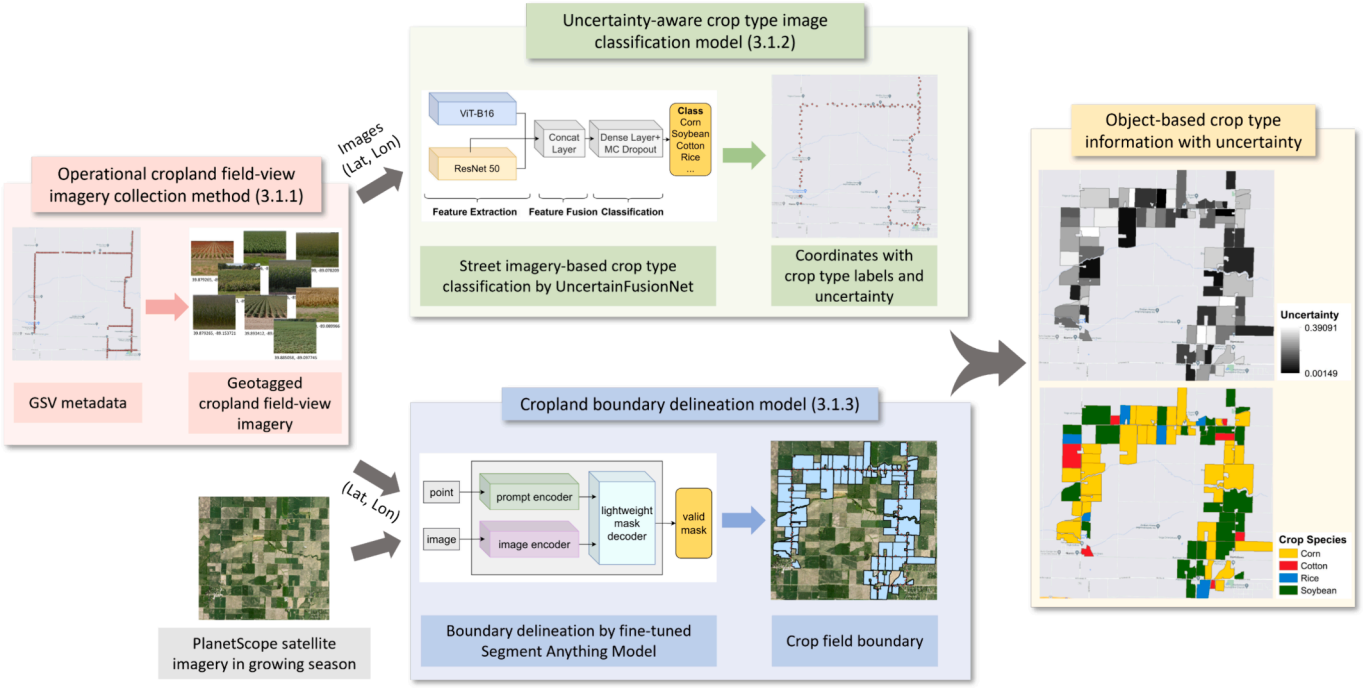

Fig. 2. Overview flowchart of CropSight. The left module (3.1.1) is employed to operationally collect cropland field-view imagery over large scales. The upper middle module (3.1.2) classifies these field-view images and estimates the prediction uncertainty metrics. The lower middle module (3.1.3) delineates the corresponding cropland boundaries. The right module displays the final retrieved object-based crop type ground truth with crop type labels, associated uncertainty metrics, and corresponding cropland boundaries.

facilitate the crop type identification from the field-view images. Additionally, the model could estimate the uncertainty associated with each prediction, a vital attribute that enables the identification and isolation of predictions that are substantially uncertain. The SAM model is fine-tuned and employed to delineate the cropland boundary corresponding to each geotagged field-view image from PlanetScope imagery. It utilizes the coordinate of a geotagged field-view image as a point prompt to guide the delineation of cropland boundary associated with this field-view image. With the geotagged GSV images and corresponding PlanetScope imagery, CropSight can retrieve object-based ground truth with crop type labels, associated uncertainty metrics, and corresponding cropland boundaries. CropSight's performance is evaluated from three perspectives, including its capability in crop type classification from field-view images, its effectiveness in cropland boundary delineation with PlanetScope images, and the reliability of its collected object-based crop type information in the context of established benchmark products (i.e., CDL). The codes, datasets, and crop type maps generated from CropSight are open source: https://github.com/rssiuiuc/CropSight/.

3.1. CropSight

3.1.1. Operational cropland field-view imagery collection method

The operational cropland field-view imagery collection method is developed to extract the cropland field-view images from the extensive GSV database for large-scale operational crop type labelling (Fig. 3). The method involves four main steps: collecting target GSV panorama imagery, extracting geotagged roadside images, removing outlier (non-agricultural) roadside images, and enhancing cropland field-view images. These four steps ensure the relevance and quality of the street view images used in the subsequent crop type identification process.

Firstly, the metadata of all available GSV panoramic images are collected within the study area. Given the Latitude and Longitude information, each panoramic image is mapped with a specific GSV point. A set of filtering algorithms (i.e., non-agricultural land filter, primary road filter, road junction filter, and off-season filter) are sequentially utilized to filter out the GSV points that are unlikely to link with clear agricultural landscapes. The non-agricultural land filter is employed to remove the GSV points along the non-agricultural fields. This is achieved by overlaying the GSV points with the WorldCover land cover map. For each GSV point, a circular buffer with a radius of 100-m is generated. If this buffer does not contain any cropland pixels of WorldCover, the corresponding GSV point is deemed irrelevant and subsequently removed. The primary road filter and road junction filter are employed to target the GSV images with high potential capturing clearer cropland views using OpenStreetMap. The GSV images on the primary roads are excluded due to the restricted field view caused by relatively large distance from the roads to the roadside fields. The GSV images from road junctions are excluded due to challenges in determining the direction of

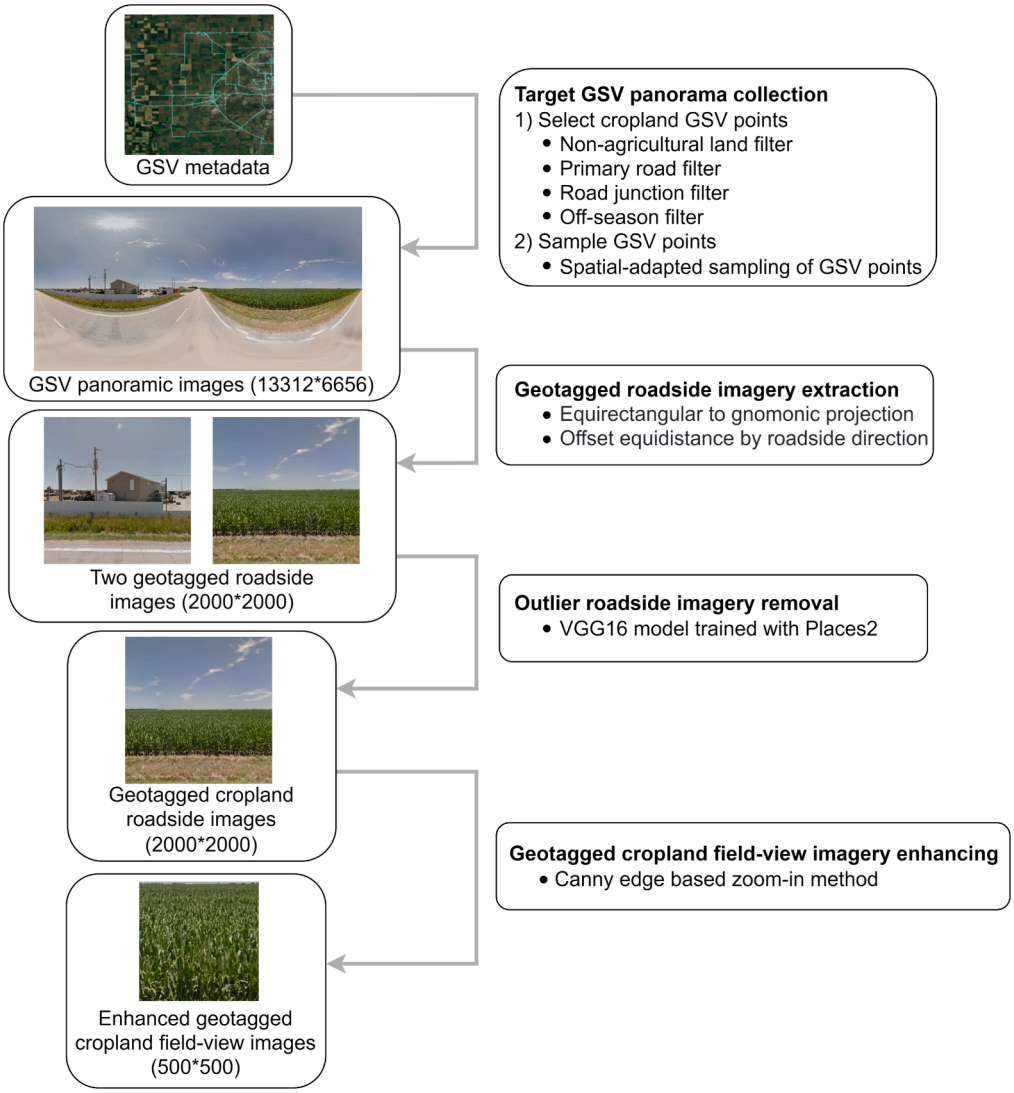

Fig. 3. Workflow of operational cropland field-view imagery collection method.

croplands surrounding these intersections. The off-season filter is utilized to filter out GSV points that are collected outside of the local crop growing season. The specific time windows of crop growing season are determined by local CPRs. This step is crucial to prevent the incorporation of GSV images that might depict barren landscapes during the agricultural off-season. These four filters ensure that only the most relevant and valuable GSV images are retained for crop type identification. Subsequently, a spatial-adapted sampling strategy is implemented to sample representative GSV points over large scales. This strategy dynamically selects GSV points from the filtered GSV points based on the required number of samples and the geographical scope of the study areas. For each administrative region (i.e., county), a Fishnet grid of uniform cells is established, covering the entire area. The cell sizes are iteratively adjusted until the number of cells with GSV points matches the desired sample size, which is calculated by dividing the total required number of samples by the number of administrative regions with filtered GSV points. In our study, the total required number of samples is set at 10,000 for each study area to ensure the prepared training data is adequately representative. Following the optimization of the Fishnet grid for each region, one GSV point is randomly selected from each cell. While we have set the county level as the default administrative level, this can be adjusted to better fit the study area or the total required number of sampling points. This spatial-adapted sampling approach guarantees that the ground truth collected from the available target GSV images is both geographically representative and widely distributed in our study areas. Based on the sampled GSV points, all the corresponding target GSV panoramic images (with the resolution of 13,312 pixels by 6656 pixels) are downloaded.

Secondly, the gathered GSV panoramic images are transformed into a set of two roadside images through a projection transformation. Specifically, the relative direction of the roadside view in relation to its corresponding GSV point is calculated using the vehicle's Heading metadata. These panoramic images are then converted from the Equirectangular projection to the Gnomonic projection, allowing for the extraction of right and left roadside images (with the resolution of 2000 pixels by 2000 pixels) based on the calculated direction. Simultaneously, the coordinates of the cropland in the roadside image are inferred by offsetting its corresponding GSV point by 50 m in the calculated direction. Each extracted roadside image is then associated with its corresponding inferred cropland coordinates for subsequent boundary delineation.

Thirdly, the collection of geotagged roadside GSV images is further refined by filtering out potential non-agricultural images using a pretrained VGG16 model. VGG16 is a deep convolutional neural network known for its effectiveness in image recognition, comprising 13 convolutional layers and 3 fully connected layers (Simonyan and Zisserman, 2015). This 16-layer setup enables the precise extraction and analysis of complex visual patterns. This step addresses the issue of potential misclassifications in the WorldCover land cover product, where non-agricultural areas might be incorrectly labeled as agricultural ones. The VGG16 model, pretrained on the Places2 dataset, is employed for the refined screening of the geotagged roadside images (Zhou et al., 2018). The Places2 dataset contains almost 10 million scene photos, labeled with 476 scene categories and attributes. Only the images belonging to the agriculture-related categories (i.e., field/cultivated, field/wild, farm, corn_field, rice_paddy, field_road, vineyard, wheat_field, orchard, tree_farm, botanical_garden, and forest/broadleaf) are retained for the follow-up classification. By removing the GSV images of residential, commercial, and tourist areas, this process reduces the complexity in crop type identification and improves the accuracy of the resulting crop type ground truth data.

Lastly, the cropland field-view images are enhanced from the remaining geotagged roadside GSV images to facilitate the crop type identification. A canny edge-based zoom-in method is proposed to automatically clip patches of the field from the GSV images. The canny edge detection algorithm locates the boundary between the field and the sky by detecting the edges present in the GSV images. After identifying the boundary, the enhanced cropland field-view images (with the resolution of 500 pixels by 500 pixels) are derived by trimming out the square patch located directly below the midpoint of the horizontal boundary line (Fig. S1). This technique has been rigorously tested across thousands of agricultural street view images, demonstrating robust performance. This strategy ensures that the primary focus is on the crop plant parts of the GSV images instead of weeds and other irrelevant elements, thereby enhancing the deep learning model's ability to accurately identify target crop types (Taesiri et al., 2023). Also, it prevents the potential loss of crucial plant visual features that can occur when a large image of 2000 pixels by 2000 pixels is downscaled to 500 pixels by 500 pixels as input into a deep learning model due to the computational resource constrains.

3.1.2. Uncertainty-aware crop type image classification model

To retrieve the crop type labels from the enhanced geotagged cropland field-view images, we propose the uncertainty-aware crop type image classification model: UncertainFusionNet (Fig. 4). This model is distinctively crafted with feature fusion mechanisms and uncertainty-aware prediction. Specifically, UncertainFusionNet contains two key components: 1) The feature fusion module, which concatenates local and global salient visual features learned by ViT-B16 and ResNet-50 for field-view image classification. 2) The Bayesian classification module, which quantifies the uncertainty associated with each prediction using Monte Carlo (MC) dropout sampling to filter out low-confidence classifications. The model architecture is detailed in the following sections.

The feature fusion module has two major branches. The first branch is the ResNet-50, a pioneering convolutional neural network for image classification (He et al., 2016). It leverages a hierarchy of residual blocks to methodically extract complex image features, thereby enriching the model's pattern discernment at various depths. Complementarily, ViT-B16, as the second branch, segments imagery into patches and employing self-attention mechanisms for sequential analysis (Dosovitskiy et al., 2021). It enables a comprehensive examination of the imagery, capturing extensive spatial relationships and contextual details across the field of view. By combining the unique and complementary strengths of these two architectures, the fusion strategy equips UncertainFusionNet with the ability to analyze both intricate local and salient global crop features in field-view images, facilitating crop type identification.

Specifically, the first branch (ResNet-50) incorporates skip connections, which create direct pathways between shallower and deeper layers in the network to overcome the vanishing gradient problem that commonly arises in deep neural networks. This enables ResNet-50 to learn more complex and discriminative crop features from images. ResNet-50 begins with a layer normalization layer (Layer Norm), followed by stage 1 including convolutional layer (CONV), batch normalization layer, ReLU activation layer, and max pooling layer. Progressing through stages 2 to 5, ResNet-50 alternates between convolutional (Conv) and identity (ID) blocks, each incorporating skip connections. These blocks ensure that input and output dimensions match, maintaining uniform feature processing as the network advances into deeper stages. Within each stage, ID blocks enable the unmodified transfer of features by directly applying skip connections. In contrast, Conv blocks introduce convolutional adjustments in their skip pathways, strategically tailoring features to match the network's changing dimensional requirements. This approach is pivotal for enabling hierarchical feature learning, allowing the network to efficiently abstract complex patterns at various depths. This architectural design allows ResNet-50 to progress from extracting local crop features to synthesizing these into comprehensive global representations. The second branch (ViT-B16) is a transformer-based model that employs a self-attention mechanism to model relationships between different regions of an image. ViT-B16 has been shown to be highly effective in computer vision tasks (e.g., image classification, segmentation, action recognition) (Han et al., 2023). It

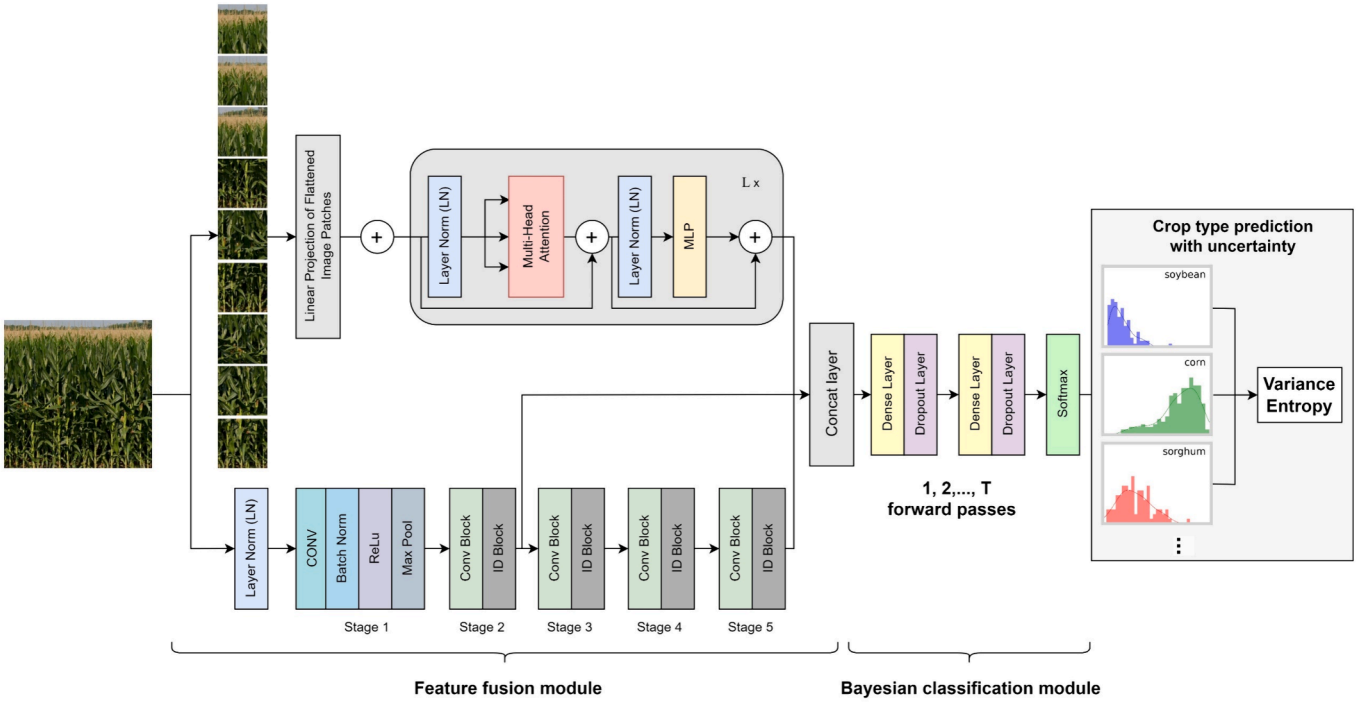

Fig. 4. Architecture of UncertainFusionNet model with feature fusion module and Bayesian classification module. The feature fusion module consists of dual branches: Vision Transformer (upper branch) and ResNet-50 (lower branch). The Bayesian classification module outputs the probability distribution of each crop type and the overall prediction uncertainty measures (i.e., variance and entropy).

divides the input image into small patches and flattens them into sequences, which are then processed by a transformer encoder to learn global features. The transformer encoder is composed of 'L' blocks. Each block starts with a Layer Norm layer for stabilizing inputs, and is followed by a Multi-head Self-Attention (MHSA) module. This module enables the model to focus on and integrate different parts of the image simultaneously, thus improving crop feature extraction and contextual understanding. A skip connection is utilized to merge the raw input with the MHSA output. Subsequently, another Layer Norm layer is applied before the Multilayer Perceptron (MLP) module. The MLP module, with its multiple dense layers, is employed to learn complex crop visual patterns in the data. A second skip connection incorporates the MHSA output (before the second Layer Norm layer) into the MLP's output. The enhanced output is then fed into the next block. This recursive processing ensures a deep, layered comprehension of the visual data, allowing the ViT-B16 to extract features with increasing sophistication, thereby boosting its analytical capabilities for complex vision tasks.

The Bayesian classification module utilizes the fused features and Bayesian theory to retrieve crop type labels and estimate the prediction uncertainty. It consists of two dense layers (also known as fully connected layers), with each followed by a dropout layer. The dropout layers serve two purposes: 1) improving network's generalization performance by mitigating overfitting during the training phase, and 2) enabling the Bayesian inference by the MC dropout technique (Gal and Ghahramani, 2016).

In a Bayesian convolutional neural network, the model output is described as a predictive distribution $p(y^*|x^*, X, Y)$ (Eq. (1)), rather than the deterministic point estimate (i.e., $y^* = f(x^*)$) in the conventional neural networks. This predictive distribution integrates the model likelihood (i.e., $p(y^*|x^*, \theta)$) over the posterior distribution of the model weights θ (i.e., $p(\theta|X, Y)$):

$$p(y^*|x^*, X, Y) = p\left((y_1^*, y_2^*, \cdots, y_c^*)|x^*, X, Y\right) = \int p\left((y_1^*, y_2^*, \cdots, y_c^*)|x^*, \theta\right) p(\theta|X, Y) d\theta \quad (1)$$

where x^* denotes an input image, and y^* denotes the corresponding output of the neural network model constructed with the training data X and Y. y^* is a vector comprising elements $y_1^*, y_2^*, \cdots, y_c^*$, with y_c^* in this vector signifying the predicted probability of class c, obtained through a softmax function. θ represents the set of weight parameters of the trained neural network model.

The posterior distribution $p(\theta|X, Y)$ is typically intractable for direct computation. To address this, the MC dropout is employed as a practical approximation method. The MC dropout follows a Bernoulli distribution, introducing a form of variability akin to training multiple sub-networks. Each sub-network is formulated with a different set of model weights. This method effectively mimics sampling from the posterior distribution of model weights for the Bayesian inference. During the inference, the MC dropout is applied to generate multiple sets of weights $\{\theta_1, \theta_2, \cdots, \theta_t\}$ by activating dropout. Each set of weights θ_t is used to make a prediction y^*, resulting in a total of T forward pass predictions. These T predictions are then aggregated to estimate the final predictive distribution. The expected value of the predictive distribution is estimated as the average of all predictions (Eq. (2)):

$$\mathbb{E}\left(p\left((y_1^*, y_2^*, \cdots, y_c^*)|x^*, X, Y\right)\right) \approx \frac{1}{T}\sum_{t=1}^{T} p\left((y_1^*, y_2^*, \cdots, y_c^*)|x^*, \theta_t\right) \quad (2)$$

By implementing the MC dropout for the Bayesian inference, UncertainFusionNet produces probabilistic predictions with uncertainty estimates, enabling the identification and filtering of low confidence crop type predictions. Uncertainty is quantified using two metrics: entropy (H) (Eq. (3)) and variance (σ^2) (Eq. (4)). Entropy is a measure of the disorder in the model's predictions, with higher entropy indicating a

larger degree of unpredictability in the model's outputs. Variance is a measure of the spread in the model's predictive probability of each class. A high predictive variance indicates that the model's predictions given multiple sets of weights $\{\theta_1, \theta_2, \cdots, \theta_t\}$ exhibit considerable variability within each class, implying higher uncertainty. Conversely, a low predictive variance indicates that the model's predictions are concentrated for each class, implying higher confidence.

$$H(p((y_1^*, y_2^*, \cdots, y_c^*)|x^*, X, Y)) = -\sum_{c=1}^{C}((\frac{1}{T}\sum_{t=1}^{T}p(y_c^*|x^*, \theta_t))\log(\frac{1}{T}\sum_{t=1}^{T}p(y_c^*|x^*, \theta_t) \tag{3}$$

$$\sigma^2(p((y_1^*, y_2^*, \cdots, y_c^*)|x^*, X, Y)) = \sum_{c=1}^{C}(\frac{1}{T}\sum_{t=1}^{T}(p(y_c^*|x^*, \theta_t) - \left(\frac{1}{T}\sum_{t'=1}^{T}p(y_c^*|x^*, \theta_{t'})\right)^2)) \tag{4}$$

where C denotes the number of crop type classes, while T indicates the number of forward passes.

To effectively filter out predictions with high uncertainty, UncertainFusionNet takes into account both entropy and variance for quantifying uncertainty with predefined thresholds set according to literatures (Abdar et al., 2021a; Gour and Jain, 2022; Arco et al., 2023). These thresholds are determined by analyzing the intersection points of the uncertainty metrics' density distributions (i.e., variance or entropy) for both correctly and wrongly identified samples among all classes during the UncertainFusionNet training (Fig. 5). Setting the thresholds at these intersections facilitates an optimal tradeoff between classification precision and amount of field-view images for ground truth collection. The filtering rule is expressed with an indicator function r_{x^*} (Eq. (5)).

$$r_{x^*} = 1_{x^*}(H, \sigma^2) = \begin{cases} 1, & if\ H < H_{thr}, \sigma^2 < \sigma^2{}_{thr} \\ 0, & otherwise \end{cases} \tag{5}$$

where an output of 1 signifies that a prediction has relatively high confidence and should be retained. An output of 0 indicates that a prediction has relatively high uncertainty and should be removed.

During the training process, an uncertainty-aware loss function (Eq. (6)) is employed to optimize the parameters of the model. The uncertainty-aware loss function not only includes the conventional cross-entropy (CE) loss function that is used to improve the predictive probability of the ground truth classes, but also the predictive uncertainty (i.e., entropy) to strengthen the differentiation of uncertainty measures between correct and wrong predictions (Shamsi et al., 2023). Compared to the conventional image classification loss function based only on cross-entropy, the uncertainty-aware loss function can minimize the overlay between the uncertainty distributions of correctly classified and wrongly classified samples (Fig. 5), while maintaining the overall classification performance of UncertainFusionNet. Consequently, this leads to more reliable and confident crop type labelling with consideration of both crop prediction probability as well as associated uncertainty.

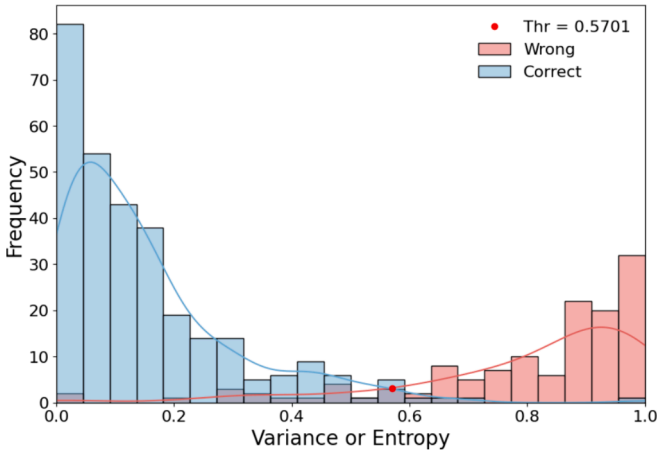

Fig. 5. Uncertainty distribution for correctly (blue) and incorrectly (red) classified samples by the UncertainFusionNet model, with the uncertainty threshold determined at the intersection of the two density curves. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(-\sum_{c=1}^{C}t_{ic}log(y_{ic}^*) + (-\sum_{c=1}^{C}y_{ic}^*log(y_{ic}^*))) \tag{6}$$

where t_{ic} is 1 when c is the index of correct class for the i th field-view image, otherwise it is 0. y_{ic}^* is the model's predicted probability that the i th field-view image belongs to the c^{th} class. N represents the number of all field-view images.

For each of our study areas, we train the UncertainFusionNet model using the corresponding CropGSV dataset, which is collected by the cropland field-view image collection method (Section 3.1.1). Fig. 6 displays all target crop-specific field-view images from the CropGSV dataset across four distinct study areas. From the 10,000 images sampled in each study area, each field-view image is manually labeled by a plant taxonomy expert with locally dominant crop types and an 'others' category. Images that do not represent the locally target crop types or those that are hard to identify due to factors such as poor lighting, blurring, or obstruction are categorized as 'others.' This focused approach facilities us to concentrate our analysis on crop species that consistently dominate local farming practices (Table S1). To ensure a balanced distribution of high-quality images for each class across all study areas, the surplus of low-quality 'others' GSV images from each study area is removed. The final dataset includes 4,508 images from study area A, 5,638 images from study area B, 5,074 images from study area C, and 6,005 images from study area D. Considering the similar visual attributes of rice and winter wheat, field-view images of 'rice' and 'winter wheat' at study area B are jointly classified as the 'cereal' category. In each study area, the CropGSV dataset is randomly divided into 60 %, 20 %, and 20 % for training, validation, and testing of the UncertainFusionNet model, respectively.

Within UncertainFusionNet, the ResNet-50 and ViT-B16 of the feature fusion module are initialized with the respective parameters pretrained on ImageNet dataset. This initialization helps equip the UncertainFusionNet with a strong visual feature extraction ability, which is beneficial to its adaptation to new tasks with fine-tuning. The remaining layers are initialized with random parameters. The hyperparameters for UncertainFusionNet (e.g., network architecture, learning rate, number of epochs, etc.) are determined by optimizing the model performance using the validation dataset, ensuring the model is both robust and generalizable. With a range of experiments in reference to previous studies (Gupta et al., 2021; Gour and Jain, 2022), stochastic gradient descent (SGD) is selected as the optimizer with learning rate of 0.001 and momentum of 0.9. The batch size is 16, and the number of

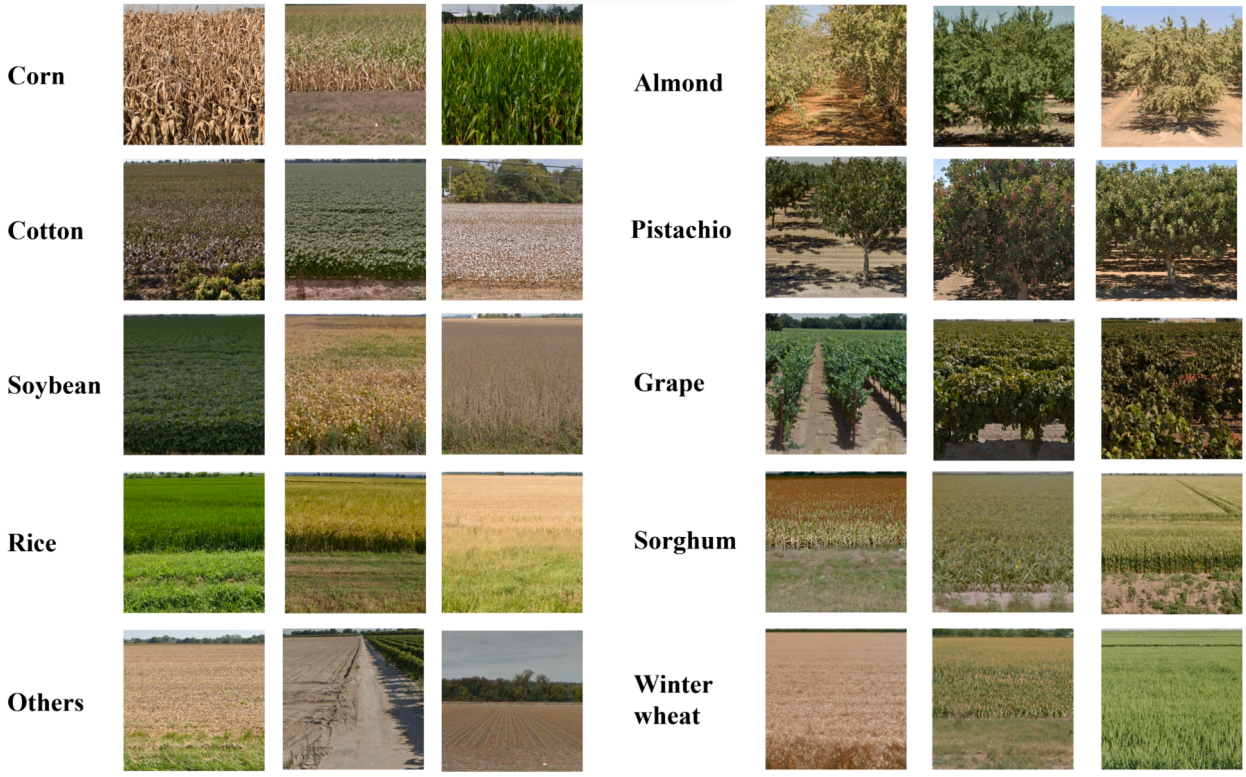

Fig. 6. Samples of CropGSV dataset showcasing field-view images of various crop types across four study areas.

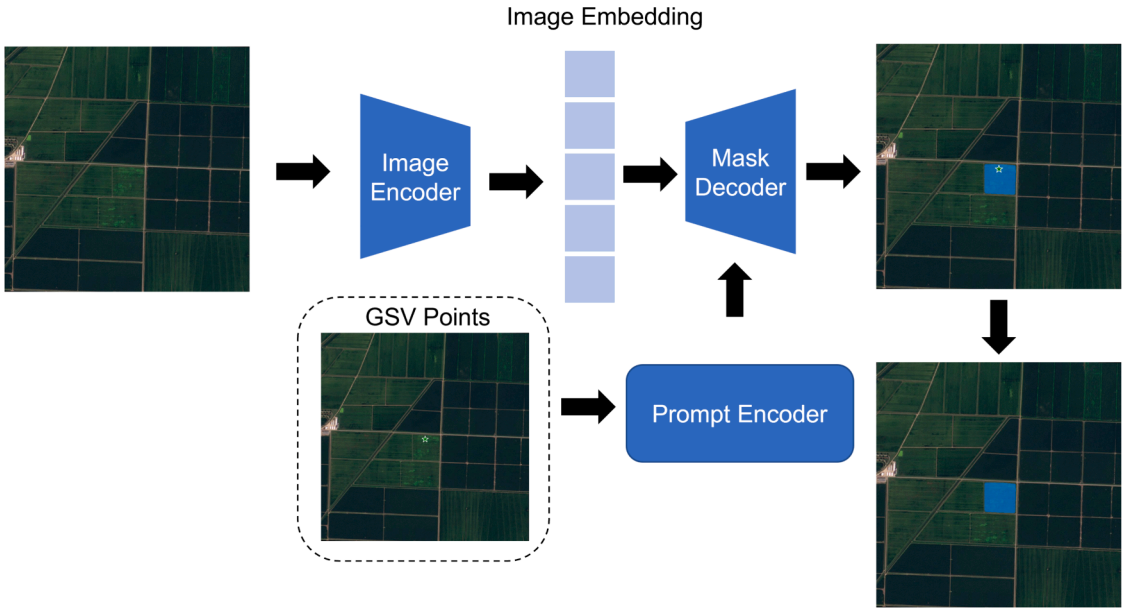

Fig. 7. Structure of SAM model.

epochs is 150. Early stopping based on the validation dataset is used in the fine-tuning process to prevent the network from overfitting when the validation loss stops decreasing.

3.1.3. *CropGSV-based cropland boundary delineation model*

With the crop type labels retrieved from the geotagged cropland field-view images, SAM is employed and fine-tuned to delineate the cropland boundaries associated with these crop type images. SAM is a highly effective state-of-the-art model designed to segment an object of interest in an image given certain prompts provided by a user (Kirillov et al., 2023). SAM diverges from traditional segmentation models by introducing a pioneering promptable segmentation strategy. Utilizing satellite imagery as the primary input and coordinates derived from geotagged field-view images as point prompts, SAM model emerges as an optimally tailored solution for extracting the cropland boundary corresponding to each field-view image in CropSight (Fig. 7). Specifically, it comprises three main components: an image encoder, a flexible prompt encoder, and a fast mask decoder: (1) The image encoder is

grounded on the architecture of a typical ViT. It is utilized to extract visual features from satellite images and convert them into image embeddings; (2) The prompt encoder is distinctively designed to embed user interactions (i.e., prompts) into an embedding vector effectively. It supports four types of prompts (i.e., points, boxes, texts, and masks), allowing for flexibility and adaptability to various user intervention. In our study, we utilize points inferred from the geotagged CropGSV images as the prompt; (3) The mask decoder is a modified Transformer decoder block, which can generate segmentation results along with confidence scores (i.e., estimated Interaction over Union (IoU)) based on the image embedding and prompt embedding. It uses two-way cross-attention, one for prompt-to-image embedding and the other for image-to-prompt embedding to learn the interaction between the prompt and image embedding for the mask (i.e., cropland boundary) generating.

In CropSight, SAM is further fine-tuned to improve its performance in delineating cropland boundaries from PlanetScope satellite images. We prioritize PlanetScope images that are free from cloud cover using the quality control layers, and select those captured in the same month as that of the corresponding GSV imagery collection for delineating cropland boundaries. SAM is built on an unprecedentedly large segmentation dataset, including over 1 billion ground-truth segmentation masks on 11 million natural images (Kirillov et al., 2023). The good accuracy of SAM in zero-shot applications indicates its potential to deliver reliable segmentation results of natural images without the need for re-training or fine-tuning on new, unseen datasets or segmentation tasks (Kirillov et al., 2023). Yet, the ability of SAM to conduct satellite image segmentation cannot be guaranteed due to the notable disparities between natural images and satellite images. Several studies have shown that the performance of SAM may degrade in challenging scenarios where the targets have weak boundaries (e.g., medical image) (Osco et al., 2023), as SAM's training set mainly contains natural images where the objects usually have strong and characteristic edges. Hence, our study employs a fine-tuning approach to adapt SAM specifically for satellite image segmentation, with a focus on accurately delineating cropland boundaries.

For each study area, we fine-tune the mask decoder of the SAM model while freezing the image encoder and prompt encoder using the manually collected CropBoundary dataset (Fig. 8). Fine-tuning only the mask decoder is demonstrated as the most computationally efficient adaptation method, effectively balancing resource use while ensuring satisfactory performance improvement of the base SAM model on new segmentation tasks (Li et al., 2023). The CropBoundary dataset consists of 200 high-resolution PlanetScope image tiles (1024 pixels by 1024 pixels) and associated digitized cropland boundaries with 50 tiles per each of the four study areas, covering diverse agricultural fields of varying sizes and patterns. Each tile captures detailed views of agricultural land during the growing season alongside accurately digitized cropland boundaries. In the process of building up the CropBoundary dataset, we initially employ the base SAM model to roughly delineate the boundary of each cropland using PlanetScope imagery with cropland points determined by visual interpretation as inputs. Manual refinements are then applied to all the fields to generate the cropland boundary reference. We focus on productive boundaries, which refer to the demarcations between different crop types, even when no physical boundaries exist within the same field. This approach ensures accurate delineation based on crop type rather than visible physical separations. In cases where PlanetScope imagery shows unclear boundaries for field parcels, potentially due to multiple crops planted in a single field, we exam images from different dates throughout the season to determine the precise parcel boundaries. The boundary data collection strategy reduces the labor intensity of manual labeling by leveraging the base SAM model's boundary identification capabilities.

For each study area, we split the corresponding 50 annotated images into training, validation, and testing datasets using the ratio 60 %, 20 %, and 20 %, respectively. These tailored datasets ensure the boundary delineation model is fine tuned to perform optimally under the unique agricultural conditions specific to each area. With the collected CropBoundary dataset, the base SAM model is fine-tuned using a loss function that combines the dice score loss and the confidence score loss (Eq. (9)). In each batch, the dice score loss, calculated for all cropland boundary masks, is the average of the difference between 1 and the dice score for each individual mask. The dice score measures the degree of overlap between the predicted and ground truth cropland boundary masks. It is calculated by dividing the twice of the area of overlap (Intersection) between the predicted and ground truth masks by the total area of prediction and ground truth (Eq. (8)). In each batch, the confidence score loss, calculated for all cropland boundary masks, is the

Fig. 8. Samples of CropBoundary dataset showcasing PlanetScope imagery and corresponding digitized cropland boundaries across four study areas A-D.

mean-square-error (MSE) between SAM's estimated IoU and the actual IoU between the predicted and ground truth masks (Eq. (9)). IoU quantifies the extent of overlap between the predicted mask and the ground truth mask. It is computed by dividing the area of overlap (Intersection) between the predicted and ground truth masks by the combined area of both (Eq. (7)). The estimated IoU is the output of SAM model. With a range of experiments in reference to previous studies (Li et al., 2023), Adam is selected as the optimizer with learning rate set as 0.00001 and weight decay as 0.9. The batch size is 2, and the number of epochs is 50. Early stopping based on the validation dataset is used in the fine-tuning process to prevent the network from overfitting when the validation loss stops decreasing.

$$IoU = \frac{Intersection}{Prediction + GroundTruth - Intersection} \tag{7}$$

$$Dice = \frac{2{*}Intersection}{Prediction + GroundTruth} \tag{8}$$

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(estimatedIoU_i - IoU_i)^2 + \frac{1}{N}\sum_{i=1}^{N}(1 - Dice_i) \tag{9}$$

where N represents the number of ground truth cropland boundaries.

3.2. CropSight framework evaluation

3.2.1. Crop type retrieval

To gain insights into UncertainFusionNet's efficacy in classifying crop types from field-view images, we compare its performance with that of FusionNet, a variant of UncertainFusionNet that maintains the same modeling architecture yet does not accommodate uncertainty. FusionNet utilizes the cross-entropy loss function to optimize its parameters and no MC dropout is utilized. Furthermore, we compare UncertainFusionNet's performance with that of two advanced benchmark models in UncertainFusionNet (i.e., ResNet-50 and ViT-B16). For a fair comparison, these two benchmark models are initialized with ImageNet pre-trained parameters and fine-tuned on the same dataset utilized by UncertainFusionNet. To analyze the capability of these models in feature learning and differentiation, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization on features from each model's penultimate layer. This layer is chosen for the visualization because it typically contains the most refined and high-level feature representations learned by the model, just before the final classification layer, thus providing a comprehensive insight into the model's ability to differentiate among various crop type classes. T-SNE is able to project and visualize high-dimensional data into a more interpretable two-dimensional space. It operates by modeling the probability distributions of data points in the high-dimensional space and then seeking a low-dimensional representation that preserves these distributions. Specifically, t-SNE calculates the similarities between points in the high-dimensional space and then optimizes the low-dimensional embedding such that similar points remain close together while dissimilar points are positioned far apart. The t-SNE analysis of the penultimate layer features aids in understanding the distribution and separation of the features learned by models. The t-SNE results are further analyzed using the Silhouette score, a key metric for evaluating class separability and cluster distinction.

Additionally, we utilize the Gradient-weighted Class Activation Mapping (Grad-CAM) visualization technique to visualize the significant features extracted by two distinct branches (i.e., ResNet-50 and ViT-B16) of the fine-tuned UncertainFusionNet model when processing field-view images. Grad-CAM highlights important areas in input images by using the gradients of a specific class being detected, as they flow into the final convolutional layer. These gradients are captured and projected back onto the input image, creating a heatmap that indicates which regions are most influential in the model's decision-making process.

Precision, recall, F1, and overall accuracy are calculated to assess models' performance on the CropGSV testing dataset. Precision (Eq. (10)) is defined as the ratio of true positive (TP) predictions to the sum of true positives and false positives (FP). This metric focuses on the model's ability to correctly identify the field-view images of specific crop type without mislabeling other types as that crop. Recall (Eq. (11)) is calculated as the ratio of true positives to the sum of true positives and false negatives (FN). This metric quantifies the model's ability in correctly capturing all relevant field-view images of specific crop type. Precision and Recall are individually calculated for each crop type, as well as aggregated across all types, providing both specific and holistic insights into the model's performance in identifying different crops. F1 score (Eq. (12)) is calculated as the harmonic mean of precision and recall for each crop type, reflecting the model's balanced performance in correctly identifying and capturing all field-view images of specific crop type. We average the F1 scores of all crop types to provide an overall view of the model's performance for crop type image classification. Overall accuracy (Eq. (13)) is a widely used metric that calculates the total proportion of correct predictions (both true positives and true negatives (TN)) across all types of crops for evaluating the model's general effectiveness across all crop types. Together, these metrics provide a comprehensive evaluation of the models' performance in retrieving crop types from GSV images.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2{*}Precision{*}Recall}{Precision + Recall} \tag{12}$$

$$Overall\ accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

3.2.2. Cropland boundary delineation

To assess the fine-tuned SAM model's ability of delineating cropland boundaries from PlanetScope images, we compare its performance with that of the base SAM and the Mask Region-based Convolutional Neural Network (Mask-RCNN). Mask-RCNN is one of the state-of-the-art model for instance segmentation that combines object detection and semantic segmentation (Waldner and Diakogiannis, 2020; Jong et al., 2022; Wang et al., 2022). It utilizes ResNet-50 as its backbone for feature extraction. Building on this, Mask-RCNN then employs a Region Proposal Network to detect objects and create corresponding bounding boxes, followed by a separate branch for generating precise segmentation masks within each identified bounding box (He et al., 2018). To ensure an equitable comparison and align with the promptable design of SAM, we refine the output of Mask-RCNN, pinpointing the final cropland boundaries by identifying the detected masks that encompass the relevant GSV image coordinates. We fine-tune the Mask-RCNN model initialized with ResNet-50 pretrained on ImageNet with our collected CropBoundary training dataset (Mei et al., 2022). All the parameters of Mask-RCNN are fine-tuned with the same CropBoundary training dataset as used by the SAM model.

With the base SAM, fine-tuned SAM and fine-tuned Mask-RCNN models, we assess their performance using the CropBoundary testing dataset. We evaluate the models' performance in boundary delineation at the object level using precision (Eq. (10)), recall (Eq. (11)), and F1 (Eq. (12)) to align with our goal of retrieving object-based crop type ground truth. The IoU threshold of 0.50 is utilized to determine the accuracy of cropland boundary delineation (G. Braga et al., 2020; Jong et al., 2022; Li et al., 2021; Mei et al., 2022). The cropland boundary is deemed correctly delineated (TP) if its IoU value exceeds 0.5. A false positive (FP) is recorded when the IoU value lies between 0 and 0.5, indicating a partial but incorrect overlap between the predicted and ground truth field boundaries. A false negative (FN) is noted when the

IoU value is 0, signifying that the model entirely misses the cropland boundary present in the image.

3.2.3. Evaluation of CropSight and benchmark crop type product

To assess the reliability of CropSight's collected object-based crop type information, we evaluate the quality of crop type information retrieved by CropSight along with the corresponding benchmark CDL crop information using a more extensive object-based crop type ground truth data (i.e., both crop type and associated field boundary) collected from GSV images and satellite images. This dataset includes a wider array of crop type samples across different CDL confidence levels. The varying levels of CDL confidence reflect diverse crop characteristics on the ground, facilitating a more thorough comparison and assessment. Specifically, we employ the cropland field-view collection method (section 3.1.1) to locate relevant GSV points and sample them across confidence levels of CDL. We then collect corresponding object-based crop type ground truth from GSV and PlanetScope images by visual interpretation. Any ambiguous samples are excluded from this ground truth dataset. At every study area, we collect 200 object-based ground truth samples for each study crop type.

Given that CDL is the pixel-based crop type classification product, we evaluate the performance of CDL and CropSight in retrieving crop type information at both pixel- and aggregated object-levels. For a comprehensive comparison, we calculate overall accuracy as well as precision

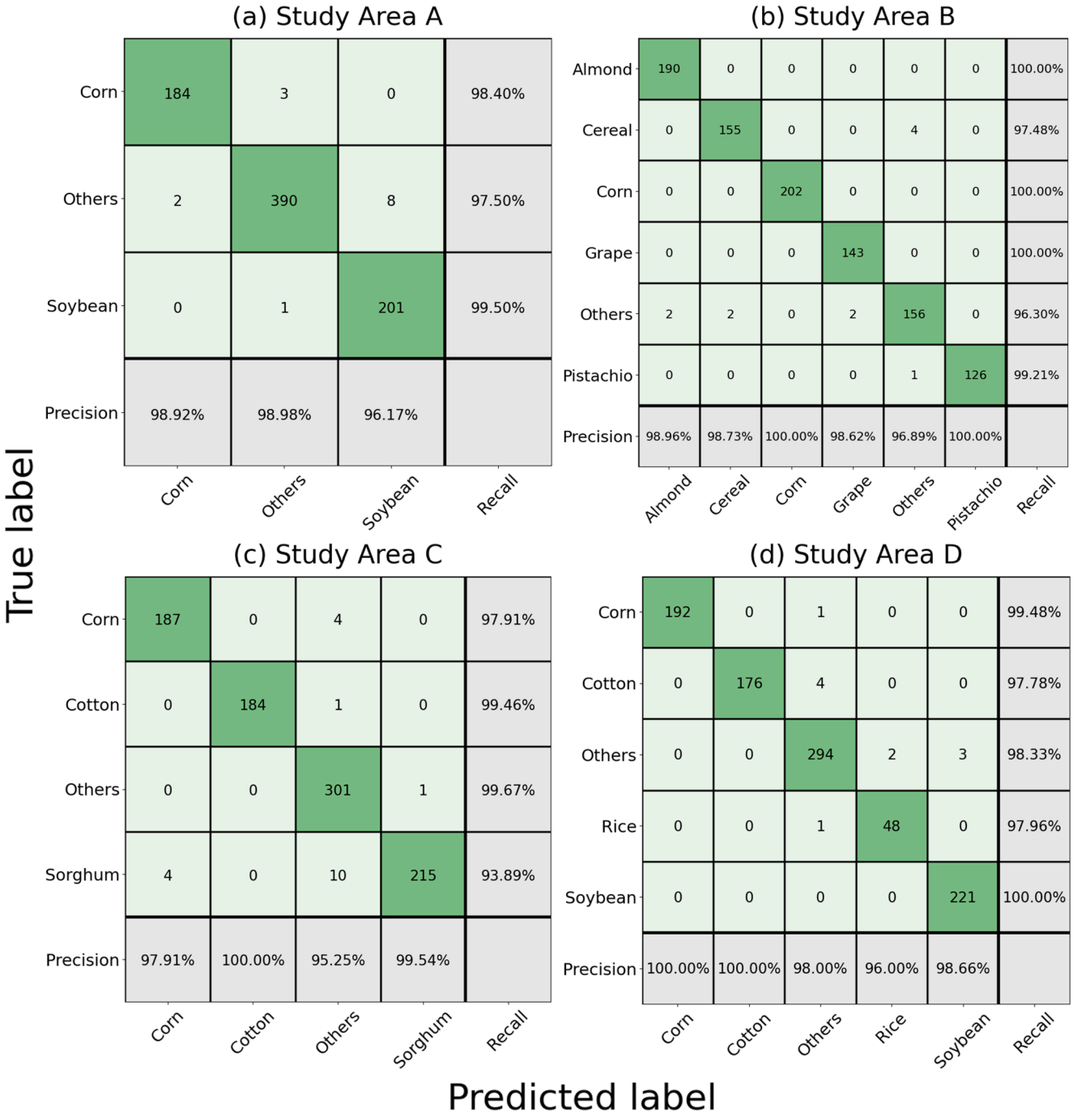

Fig. 9. Confusion matrix showing the performance of the UncertainFusionNet model (with incorporating uncertainty) in identifying the predominant crop species across four study areas A-D. The results are based on field-view image classification from the corresponding CropGSV test datasets.

for each crop species at both levels. The overall accuracy indicates the proportion of correctly identified crop types. Precision for each crop type indicates the proportion of correctly identified instances of that crop type out of all instances classified as that type, and is selected due to its importance in evaluating the quality of ground truthing for each crop type. At the pixel-level, we overlay the crop type label information from both CDL and CropSight with ground truth cropland boundaries. CropSight's pixel-level labels are consistent within a single field due to its unique GSV-based retrieved method. CDL's pixel-level labels are directly retrieved from the CDL layer. To evaluate pixel-level accuracy, we compare the estimated crop type labels of the pixels within the boundaries against the ground truth labels for those pixels. At the object-level, we overlay the crop type label information from both CDL and CropSight with the cropland boundaries detected by CropSight to ensure a fair comparison. CDL's object-level labels are derived from the most common crop type among its pixels within field boundaries. To assess object-level accuracy, we examine whether the estimated object-level crop type information aligns with the ground truth. Correctly retrieved samples at this level are defined as those with both matching crop type labels and accurately delineated cropland boundaries with corresponding IoU values exceeding 0.5 (section 3.2.2).

To evaluate crop label availability for downstream mapping tasks, we employ the CropSight framework to gather crop type labels in four study areas and create density maps. These maps display the distribution of crop type labels across 50 km by 50 km grids, providing insights into label density crucial for crop type mapping. To extend the evaluation of CropSight's applicability in regions with diverse agricultural practices, we further assess the accuracy of the object-based crop type ground truth data retrieved by CropSight in Brazil by following the same procedure as the extensive object-based crop type ground truth data collected in US. Our study area covers Paraná and São Paulo, which are key agricultural regions in Brazil and are predominantly cultivated with key crops such as soybeans, corn, sugarcane, and wheat. This area exhibits a mix of advanced agricultural technologies and traditional practices (e.g., manual labor), in contrast to the extensively mechanized agriculture predominant in the US (Bolfe et al., 2020). Additionally, the fields in Brazil are more fragmented with irregular shapes and varying sizes compared to those in the US (Fig. S4).

4. Results

4.1. Crop type identification

4.1.1. Performance of UncertainFusionNet

Fig. 9 shows the confusion matrix of crop type identification results of UncertainFusionNet on CropGSV testing dataset across four study areas. At study area A, UncertainFusionNet demonstrates strong discriminatory ability, effectively distinguishing between 'corn' and 'soybean' with minimal misclassification. It achieves precision rates of 98.92 % for 'corn' field-view images and 96.17 % for 'soybean'field-view images. Additionally, the model exhibits high recall for these two categories, each exceeding 98.4 %, suggesting its robustness in accurately identifying true instances of each crop type. Within study area B, UncertainFusionNet excels with an accurate classification of 'corn' field-view images, reaching 100 % precision and recall. This exceptional performance is likely attributed to the distinct characteristics of corn, such as its long leaves and stalks, which are unique compared to other local crop types. Additionally, the model precisely identifies 'almond', 'cereal', 'grape', and 'pistachio', with each category achieving a precision rate exceeding 98.6 %. UncertainFusionNet effectively distinguishes between 'almond' and 'pistachio', despite both being tree crops with similar-looking leaves. As for study area C, UncertainFusionNet shows high accuracy in classifying 'cotton' field-view images, achieving a precision rate of 100 % and a recall rate of 99.46 %. In the classification of 'corn' and 'sorghum' field-view images, it maintains high precision, reaching 97.91 % for 'corn' and 99.54 % for 'sorghum'. At study area D, UncertainFusionNet exhibits good classification precision for field-view images of 'corn', 'cotton', and 'soybean', consistently exceeding 98.6 %. UncertainFusionNet's recall accuracy is consistently higher than 97.7 % across all local species. Overall, these results highlight UncertainFusionNet's strong capability in accurately identifying crop types from field-view images across diverse agricultural landscapes.

Table 1
Performance evaluation of UncertainFusionNet and FusionNet in crop type identification across study areas A-D. Evaluation metrics (i.e., precision, recall, F1, and overall accuracy) are derived from the CropGSV test datasets for each study area.

Method	Study area	Precision	Recall	F1	Overall accuracy
FusionNet	A	0.9551	0.9654	0.9600	0.9589
	B	0.9658	0.9650	0.9653	0.9627
	C	0.9627	0.9625	0.9629	0.9576
	D	0.9377	0.9182	0.9270	0.9385
UncertainFusionNet	A	0.9803	0.9847	0.9824	0.9823
	B	0.9887	0.9883	0.9885	0.9888
	C	0.9887	0.9883	0.9885	0.9779
	D	0.9853	0.9871	0.9862	0.9883

4.1.2. Impact of uncertainty

Table 1 presents a comparative analysis between FusionNet and UncertainFusionNet in terms of their performance in identifying crop types from field-view images across all study areas using the CropGSV test dataset. Across four study areas, UncertainFusionNet consistently surpasses FusionNet, exhibiting superior precision, recall, F1 scores, and overall accuracy with improvements in each metric ranging between 0.02 and 0.06. UncertainFusionNet achieves these accuracy metrics of around 0.98 for all the study areas (Table 1). In particular at study area D, UncertainFusionNet shows a significant improvement in field-view crop type classification over FusionNet, with overall accuracy of 0.9883 versus 0.9385. This improvement is likely attributed to UncertainFusionNet's enhanced ability in handling the field-view images of other plants (e.g., grass and weed) with visual similarities to target crops (e.g., rice). With the Bayesian design, UncertainFusionNet can assess the uncertainty level of its prediction. This uncertainty measure enables the model to identify and exclude ambiguous samples, particularly those with similar visual features in planting structures, appearances, and colors to target crops. These results highlight the value of integrating uncertainty information into predictions, thereby improving the accuracy of determined crop type labels.

During the process of crop type labelling, UncertainFusionNet leverages uncertainty information (i.e., variance and entropy) associated with each field-view image prediction to help enhance the confidence and reliability of collected crop type labels (Fig. 10). These uncertainty measures are calculated based on the probability distribution of all the classes. Field-view images exhibiting high variance or entropy are discarded by UncertainFusionNet (where $r_{x^*} = 1$ indicates retaining, and $r_{x^*} = 0$ indicates discarding). Specifically in Fig. 10, (a) displays two 'corn' field-view images at study area A. The left image, characterized by high uncertainty and confusion between harvested corn and soybean of similar stage, is discarded. In contrast, the right image with distinct corn leaves and silking is retained with its low uncertainty; (b) presents two 'grape' field-view images at study area B. Despite correct identification, the left image exhibits elevated uncertainty, potentially due to weed interference and empty trellises. The right image with orderly rows of grapevines with lush green leaves above trellises is classified with more confidence; In (c), two 'sorghum' field-view images with different confidence at study area C are shown. The left image is contaminated by weed, resulting in confusion among the classes 'sorghum', 'corn', and 'others' and high uncertainty in the identification process. The sorghum

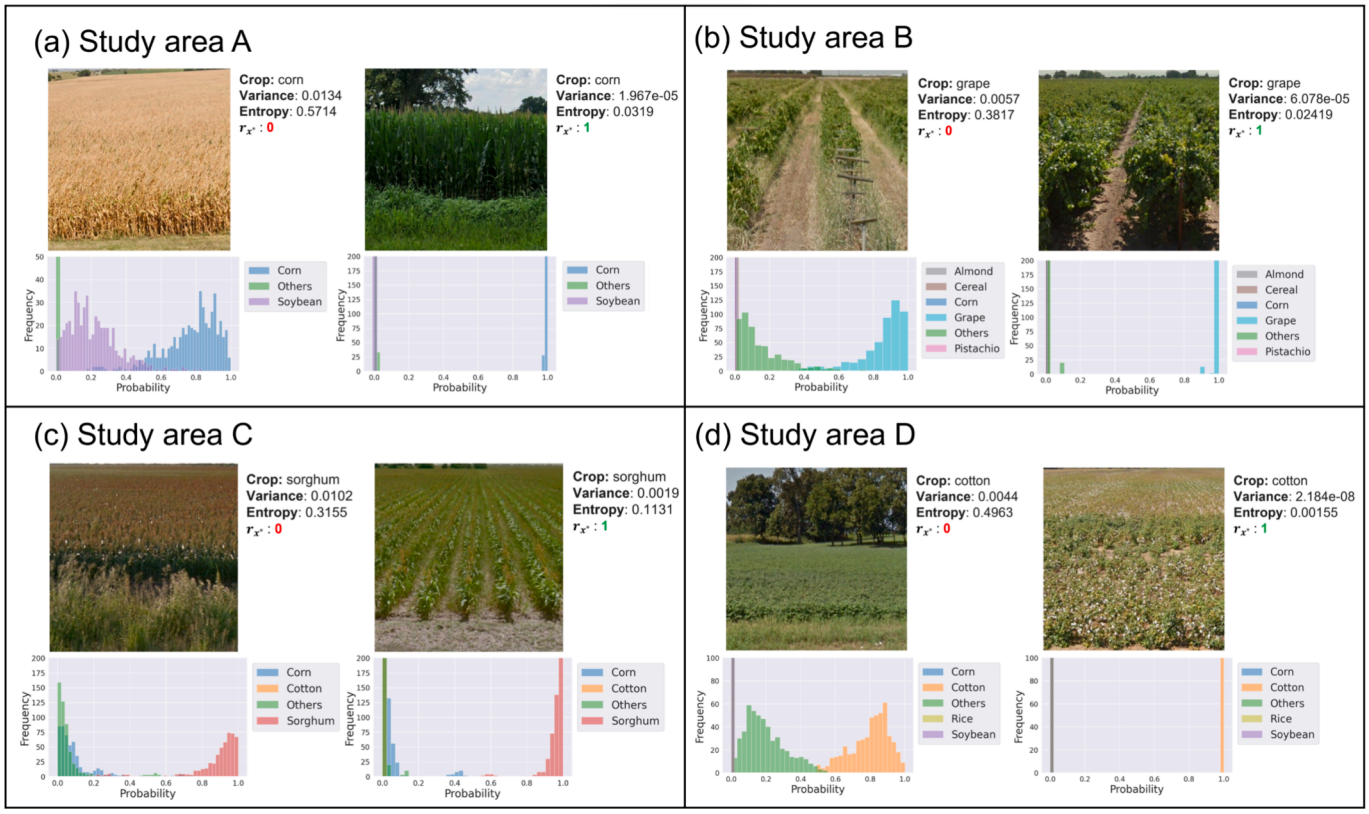

Fig. 10. Visualization of crop type labelling processes of the UncertainFusionNet model across four study areas A-D. Each field-view image is accompanied by uncertainty information (i.e., variance and entropy) and confidence indicator (r_{x^*}), where $r_{x^*} = 1$ denotes high confidence and $r_{x^*} = 0$ signifies low confidence. (a) displays the corn field-view images at study area A; (b) presents the grape field-view images at study area B; (c) shows the sorghum field-view images at study area C; (d) shows the cotton field-view images at study area D.

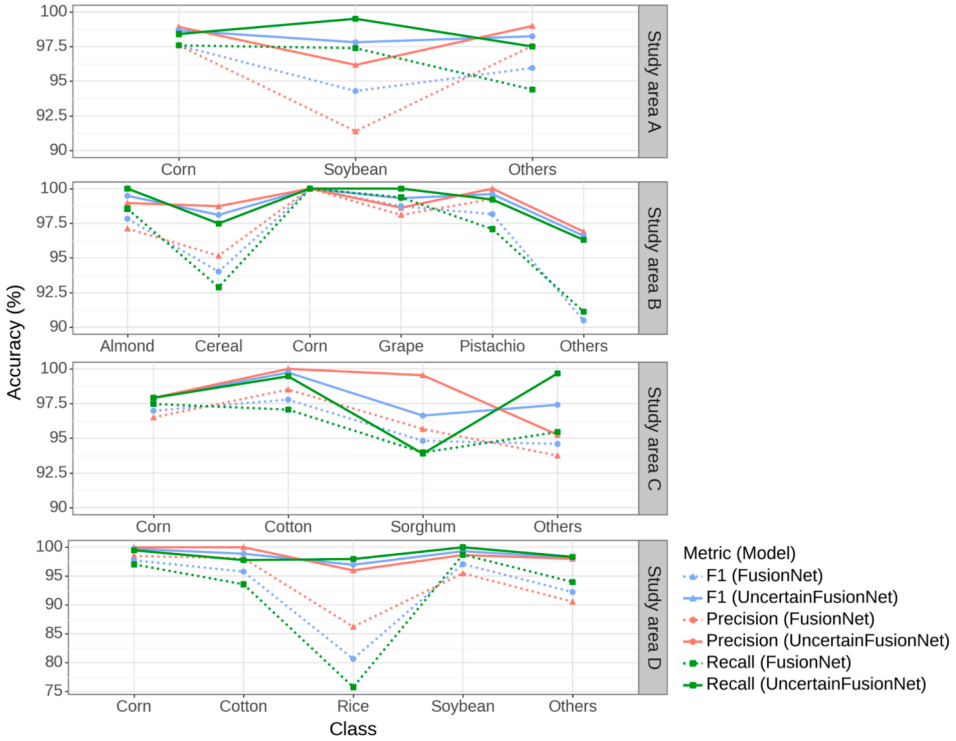

Fig. 11. Comparative performance analysis (i.e., F1, precision, and recall) of UncertainFusionNet and FusionNet (without incorporating uncertainty) in identifying the dominant crop species across four study areas A-D. The evaluation is based on field-view image classification results from the corresponding CropGSV test datasets.

plants are neatly arranged in the right image with vibrant green leaves and unique inflorescence structures, forming compact clusters in spike shapes, which is confidently classified with low uncertainty; (d) shows predictions for two 'cotton' field-view images at study area D. The right image, with visible white blooms, is identified as 'cotton' with high confidence, whereas the left image, though labeled as 'cotton', exhibits significant uncertainty, possibly due to background trees and foreground weed interference. Across all study areas, images with more distinct features of crops (e.g., leaf shape, canopy morphology, and flowering) tend to yield lower variance and entropy derived from corresponding probability distributions. However, in situations where images share common features among different local plants and target crops, the model may produce inaccurate classifications characterized by high variance and entropy. This underscores that the integration of these uncertainty measures within UncertainFusionNet serves as a vital filter, bolstering both the robustness and reliability of collected crop type labels across varied field conditions.

By accounting for the uncertainty in each prediction, UncertainFusionNet effectively diminishes misclassifications, yielding improvements in precision, recall and F1 score for each crop species throughout all study areas, in comparison to FusionNet (Fig. 11). At study area A, UncertainFusionNet demonstrates an advantage over FusionNet, achieving a 1–4 % increase in accuracy across recall, precision, and F1 scores, with all metrics exceeding 96 % for the classification of each crop type. The precision of classifying 'soybean' field-view images increases the most by approximately 4 %. At study area B, UncertainFusionNet achieves a 2 % enhancement in all accuracy metrics for classifying 'almond', 'grape', and 'pistachio' field-view images over FusionNet, and an even more pronounced 5 % increase for 'cereal' field-view images. These improvements are attributed to UncertainFusionNet's estimation of prediction uncertainty, which effectively reduces misclassifications arising from visually similar plants/crops at various growth stages. This is especially apparent in distinguishing between 'pistachio' and 'almond', as well as between 'cereal' and 'others' (Fig. 9 and Fig. S2). At study area C, UncertainFusionNet outperforms FusionNet in classifying 'corn' and 'cotton' field-view images, surpassing it by approximately 1–2 % across all accuracy metrics and consistently maintaining over 97.5 %. It shows strong performance in 'cotton' classification, achieving 100 % precision. Furthermore, UncertainFusionNet exhibits a 3 % improvement in precision for 'sorghum' compared to FusionNet. This enhancement stems from a reduction in the misclassification of non-sorghum images as 'sorghum' (Fig. 9 and Fig. S2), with its precision approaching 100 %. At study area D, UncertainFusionNet demonstrates substantial improvements over FusionNet, with increases in precision, recall, and F1 for field-view images of each crop ranging from 2 % to 20 %. Particularly noteworthy are the 'rice' field-view images, where precision is boosted from 86.21 % (FusionNet) to 96.00 %, and recall is elevated from 75.76 % (FusionNet) to 97.96 %. Overall, UncertainFusionNet, through its novel integration of uncertainty information, provides CropSight with a robust and reliable model for crop type identification from street view images. This ensures the high confidence of the labels collected across a wide range of scenarios.

Table 2
Performance evaluation of UncertainFusionNet and benchmark models (i.e., ResNet-50 and ViT-B16) in crop type identification across study areas A-D. Evaluation metrics (i.e., precision, recall, F1, and overall accuracy) are derived from the CropGSV test datasets for each area. Bolded numbers indicate the highest evaluation metric values among three models at each study area.

Study area	Model	Precision	Recall	F1	Overall accuracy
A	ResNet-50	0.9567	0.9567	0.9567	0.9566
	ViT-B16	0.9221	0.9294	0.9254	0.9256
	UncertainFusionNet	**0.9803**	**0.9847**	**0.9824**	**0.9823**
B	ResNet-50	0.9490	0.9484	0.9486	0.9458
	ViT-B16	0.9489	0.9504	0.9491	0.9467
	UncertainFusionNet	**0.9887**	**0.9883**	**0.9885**	**0.9888**
C	ResNet-50	0.9644	0.9656	0.9650	0.9625
	ViT-B16	0.9484	0.9550	0.9513	0.9487
	UncertainFusionNet	**0.9887**	**0.9883**	**0.9885**	**0.9779**
D	ResNet-50	0.9042	0.9251	0.9131	0.9212
	ViT-B16	0.9155	0.8669	0.8833	0.9021
	UncertainFusionNet	**0.9853**	**0.9871**	**0.9862**	**0.9883**

4.1.3. Comparison of UncertainFusionNet and benchmark models

Table 2 shows the accuracy metrics of crop type identification results from field-view images using UncertainFusionNet and two benchmark models (i.e., ResNet-50 and ViT-B16) across four study areas. All these three models are fine-tuned and evaluated using the CropGSV dataset. UncertainFusionNet consistently shows better performance in the field-view image classification than the two benchmark models with precision, recall, F1, and overall accuracy all larger than 0.98 across four study areas. The performance of ResNet-50 and ViT-B16 is more up to the study areas. At study areas A and C, ResNet-50 yields higher accuracy across all metrics than ViT-B16. At study area B, these two benchmark models show comparable performance. At study area D, ViT-B16 demonstrates comparatively superior precision performance compared to ResNet-50, while it exhibits slightly lower recall, F1 score, and overall accuracy, with a difference of 0.02.

Closely related to user accuracy, precision is vital in assessing the reliability of the final collected crop type labels. For all target crop species, UncertainFusionNet outperforms the two benchmark models in crop type identification across all four study areas, consistently achieving a precision exceeding 0.95 (Fig. 12). At study area A, UncertainFusionNet exhibits better performance in identifying both 'corn' and 'soybean' field-view images compared to ResNet-50 and ViT-B16. The two benchmarks achieve similar performance in crop type identification, with ResNet-50 excelling in 'corn' and ViT-B16 in 'soybean'. At study area B, UncertainFusionNet outperforms both Resnet-50 and ViT-B16 in identifying 'cereal' (0.98 vs. 0.91 vs. 0.90), 'almond' (0.98 vs. 0.96 vs. 0.94), and 'pistachio' (0.99 vs. 0.96 vs. 0.95) from field-view images. At study area C, UncertainFusionNet excels in identifying field-view images of local dominant crop species (i.e., corn, cotton, and sorghum), and ResNet-50 performs better than ViT-B16. At study area D, UncertainFusionNet consistently outperforms the second-best model, achieving an improved precision by approximately 0.04 in identifying the four local crop species. ResNet-50 slightly outperforms ViT-B16 in identifying 'corn', 'cotton', and 'soybean' field-view images. Yet, ResNet-50 lags behind ViT-B16 in 'rice' field-view image identifying, achieving a precision near 0.7.

As depicted through the t-SNE visualization in Fig. 13, UncertainFusionNet consistently produces clearer and cohesively clustered features than the two benchmark models across four study areas, as evidenced by higher Silhouette scores. This is achieved even without leveraging uncertainty information to exclude predictions with high uncertainty. This enhanced discriminative ability potentially explains UncertainFusionNet's superior crop type identification performance compared to the two benchmark models. Additionally, uncertainty information associated with each prediction could further help mitigate misclassifications that occur between crops with similar visual characteristics, such as leaf shape and canopy morphology (Fig. S3).

Further investigations into the feature learning of UncertainFusionNet are conducted using the Grad-CAM visualization technique. It examines the contributions of two modeling components (i.e., ResNet-50 and ViT-B16) to field-view image classification within UncertainFusionNet. Under the trained UncertainFusionNet, the first branch (i.e., ResNet-50) gradually shifts focus from local features towards global features as the layers go deeper, primarily highlighting the central

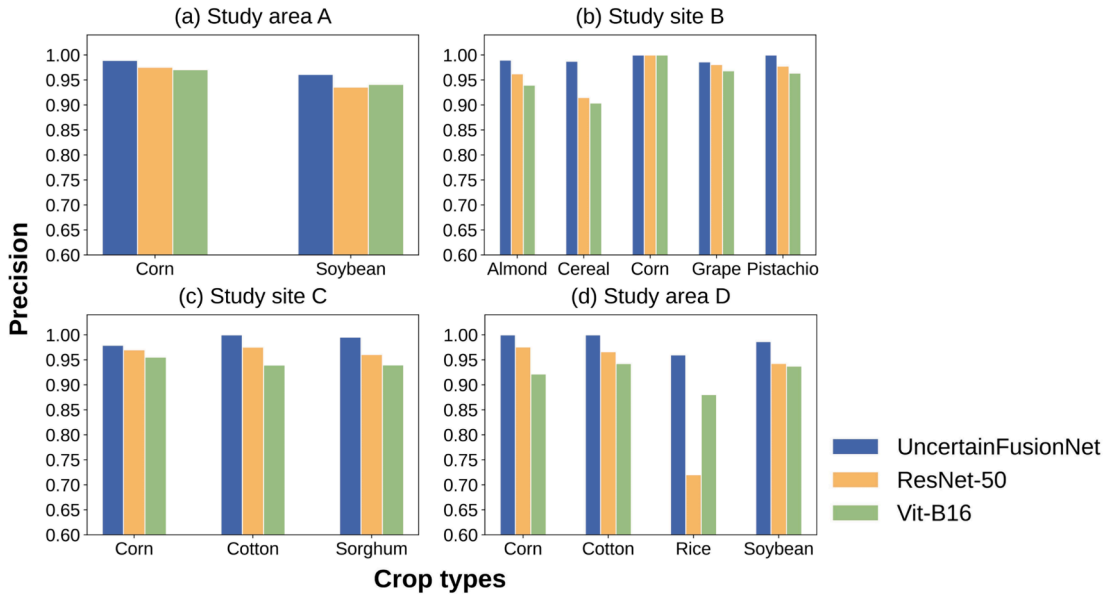

Fig. 12. Precision of crop type identification among UncertainFusionNet and two benchmark models across the four study areas A-D. These precision metrics are derived from predictions on the CropGSV test datasets of the respective study areas.

elements of corn images (Fig. 14). By contrast, the second branch (i.e., ViT-B16) showcases a consistent emphasis on global features right from the initial layers, with the significance of these features (e.g., the leaves and stalks of the corn) becoming increasingly pronounced in deeper layers. This result indicates that the feature fusion design within UncertainFusionNet integrates the complementary strengths of both architectures (i.e., ResNet-50 and ViT-B16), ensuring comprehensive learning of both local and global feature representations in field-view images. To further assess the performance of the model that integrates features from both ViT-B16 and ResNet-50, we conduct a comparison between FusionNet (without considering uncertainty information) and these two benchmark models. FusionNet shows relatively better performance in crop type identification from field-view images than the benchmark models over four study areas (Table S2). Specifically, FusionNet demonstrates superior performance in terms of precision, recall, F1 score, and overall accuracy, exhibiting improvements ranging from 0.5 % to 3 % in study areas A, B, and D. Overall, UncertainFusionNet shows superiority in providing high-confidence crop type label predictions across a variety of study areas and crop species, attributed to its feature fusion module and uncertainty-aware prediction. This superiority ensures the reliability of the crop type ground truth collected by CropSight.

4.2. Cropland boundary delineation

As shown in Table 3, fine-tuned SAM shows the best performance in delineating the cropland boundaries compared to the two benchmarks with consistently highest F1 score across four study areas. At study area A, fine-tuned SAM achieves a precision of 0.8414, which surpasses SAM (0.7470) and is comparable with that of Mask-RCNN (0.8581). Both fine-tuned SAM and SAM achieve a recall of 1, indicating their exceptional capability in identifying all relevant cropland areas without omissions, while Mask-RCNN's recall is about 0.8917. The F1 score for fine-tuned SAM is about 0.9138, outperforming both Mask-RCNN (0.8746) and SAM (0.8552). At study area B, fine-tuned SAM achieves the highest F1 score of 0.9455, followed by Mask-RCNN's 0.8870 and SAM's 0.8803. Fine-tuned SAM and SAM maintain a recall of 1, yet fine-tuned SAM's precision (0.8967) markedly exceeds SAM (0.7862). At study areas C and D, the cropland boundary delineation performance of these three models maintains a consistent pattern as that of study areas A and B. Overall, fine-tuned SAM outperforms SAM, achieving higher precision and F1 scores, which suggests enhanced boundary delineation ability of SAM on PlanetScope imagery after fine-tuning. Compared to Mask-RCNN, fine-tuned SAM shows superior performance in ensuring adequate boundary delineation of all target GSV images, reflected in its higher recall and F1 score. This is attributed to SAM's design of the prompt that ensures no omission of agricultural boundary delineation corresponding to each GSV image. Mask-RCNN, despite slightly higher precision, has much larger false negative values than fine-tuned SAM and may miss the GSV's cropland boundaries entirely. The comparative analysis underscores fine-tuned SAM's consistent and robust performance in delineating cropland boundaries with points as prompts across diverse agricultural landscapes. The enhanced performance of fine-tuned SAM is also observed in overall IoU and Dice scores among three models (Table S3).

Fig. 15 presents visualization of typical boundary delineation results from these models across four study areas. At study area A, fine-tuned SAM effectively delineates cropland boundaries, whereas SAM occasionally produces some irregular boundaries for certain fields. Mask-RCNN fails to detect the fields in the right-lower corner of the viewing area. At study area B, fine-tuned SAM more accurately captures field boundaries in terms of completeness and correctness. SAM sometimes struggles with fields that share similar colors or patterns, leading to irregular boundaries. Mask-RCNN generally achieves comparable completeness as fine-tuned SAM, though it may overlook the delineation of some fields. At study area C, which features similar color tones and less distinct boundaries, fine-tuned SAM offers the most accurate boundary delineation, while SAM might occasionally group distinct croplands as a single entity. Mask-RCNN faces challenges in accurately capturing boundaries in this context. At study area D, the fine-tuned SAM maintains its high level of accuracy in identifying cropland boundaries. In contrast, SAM sometimes inaccurately merges multiple fields into a single entity. Meanwhile, Mask-RCNN generally delineates boundaries with precision but notably fails to identify one field located in the central left portion of the viewing area. In summary, both the accuracy metrics and visual analysis consistently demonstrate the superior capability of the fine-tuned SAM and the base SAM in capturing cropland boundaries corresponding to GSV images, outperforming Mask-RCNN in this aspect. Additionally, the fine-tuning process further elevates SAM's precision in delineating cropland boundaries from

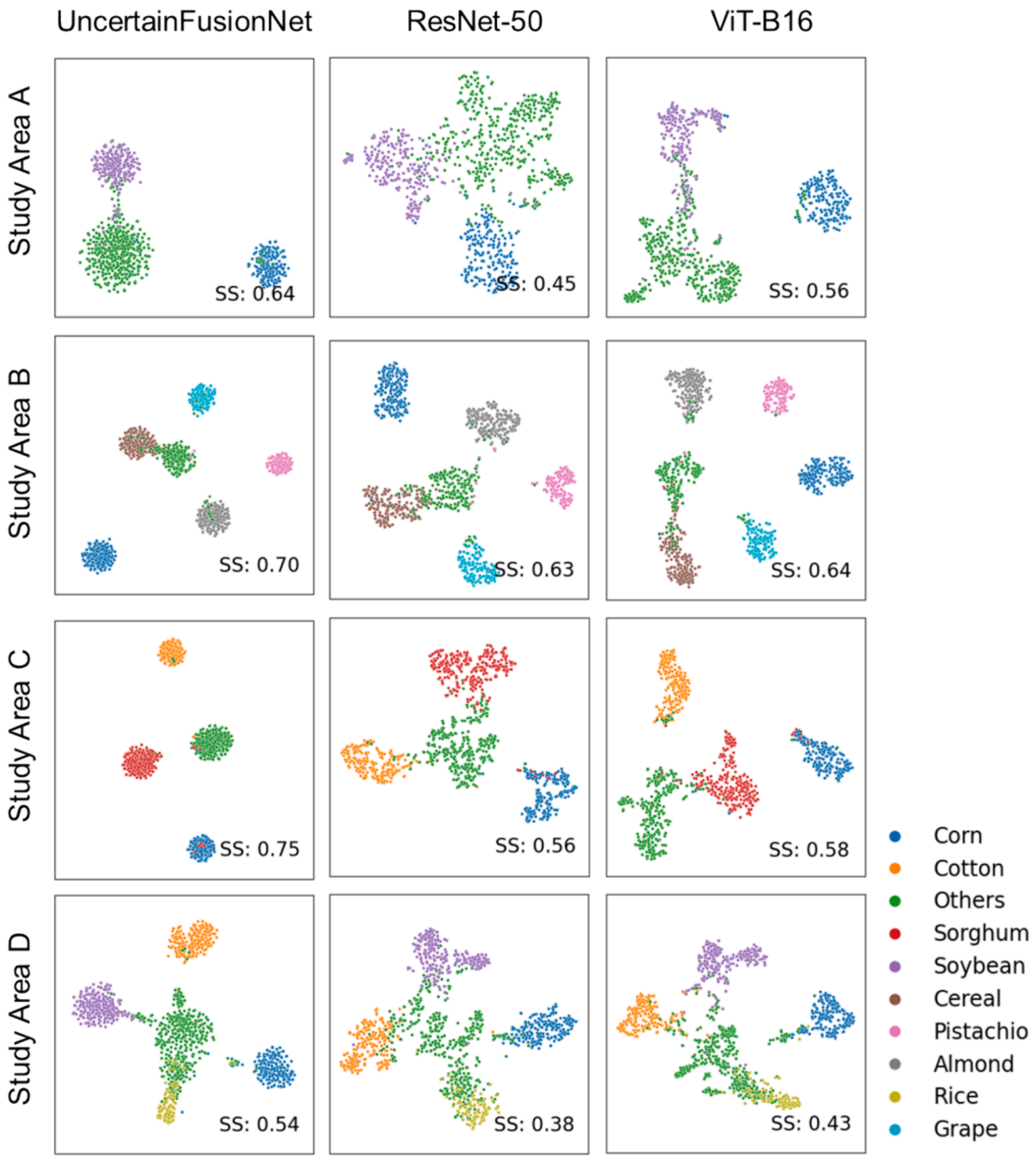

Fig. 13. t-SNE visualization illustrating the feature separation achieved by UncertainFusionNet (without filtering out predictions with high uncertainty) in comparison to two benchmark models across the four study areas A-D. The t-SNE maps are generated using the features from the penultimate layer of each model on the CropGSV test datasets of the respective study areas. SS refers to Silhouette score.

PlanetScope imagery.

4.3. Evaluation of CropSight and benchmark CDL product

As shown in Table 4, CropSight consistently outperforms CDL on both evaluation levels across four study areas, maintaining an average overall accuracy above 0.95. At study area A, CropSight achieves an overall accuracy of 1.0 at both pixel-level and 0. 9732 at object-level. CDL offers a comparable object-level accuracy of 0. 9695, but its pixel-level accuracy drops to approximately 0.9332. At study area B, CropSight achieves an overall accuracy of around 0.9436 at pixel-level and 0. 9944 at object-level. However, CDL underperforms with an object-level accuracy of 0.8276 and a significantly lower pixel-level accuracy of 0.7586. The low overall accuracy of CDL is likely attributed to the complex agricultural landscape of study area B (i.e., California) with diverse local crop species. This complexity is reflected in the CDL reported accuracies, with grape identification around 0.8 and winter wheat close to 0.7 for both user and producer accuracies, highlighting the difficulty in accurately labeling varied crop species from satellite data. At study areas C and D, CropSight's accuracy remains above 0.97 at pixel level and above 0.93 at object-level, whereas CDL achieves a decent object-level accuracy of approximately 0.9 but a lower pixel-level accuracy of around 0.84.

Closely linked to user accuracy, precision is crucial for evaluating the reliability of the final gathered crop type labels. Across all the four study areas, CropSight outperforms CDL in precison at both pixel- and object-levels across four study areas (Fig. 16). CropSight maintains a precision accuracy of over 95 % for most crop species at these regions. By contrast, CDL's precision accuracy falls behind by more than 5 %, particularly for crops like 'almond' and 'cereal' in study area B, and 'corn', 'cotton' and 'soybean' in D. The better performance of CropSight across all study areas highlights its effectiveness in handling diverse crop species. This success can be largely credited to its method of analyzing field-view images to identify crop types, focusing on distinct visual features.

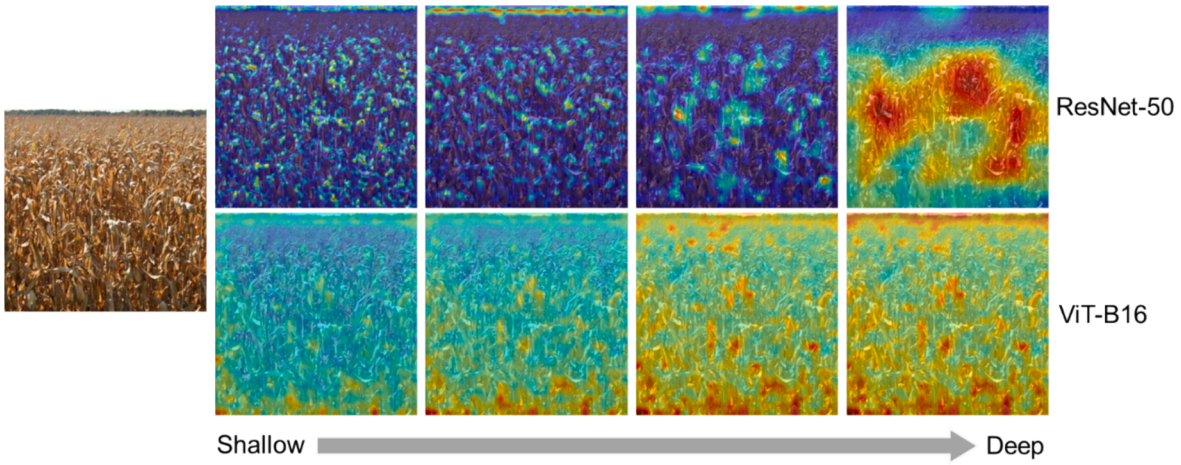

Fig. 14. Grad-CAM visualization of the first branch (i.e., ResNet-50) and the second branch (i.e., ViT-B16) of the fine-tuned UncertainFusionNet model, illustrating the feature maps from shallow to deep layers.

Table 3
Performance evaluation of fine-tuned SAM and benchmark models (i.e., base SAM and Mask-RCNN) in cropland boundary delineation across study areas A-D. Evaluation metrics (i.e., TP, FP, FN, precision, recall, and F1) are derived from the CropBoundary test datasets for each area. Bolded numbers indicate the highest evaluation metric values among three models at each study area.

Study area	Model	TP	FP	FN	Precision	Recall	F1
A	Mask-RCNN	387	64	**47**	**0.8581**	0.8917	0.8746
	SAM	372	**126**	0	0.7470	**1.0000**	0.8552
	Fine-tuned SAM	**419**	79	0	0.8414	**1.0000**	**0.9138**
B	Mask-RCNN	671	56	**115**	**0.9230**	0.8537	0.8870
	SAM	662	**180**	0	0.7862	**1.0000**	0.8803
	Fine-tuned SAM	**755**	87	0	0.8967	**1.0000**	**0.9455**
C	Mask-RCNN	291	38	**105**	**0.8845**	0.7348	0.8028
	SAM	349	**85**	0	0.8041	**1.0000**	0.8914
	Fine-tuned SAM	**372**	62	0	0.8571	**1.0000**	**0.9231**
D	Mask-RCNN	620	80	**66**	**0.8857**	0.9038	0.8947
	SAM	560	**206**	0	0.7311	**1.0000**	0.8446
	Fine-tuned SAM	**659**	107	0	0.8603	**1.0000**	**0.9249**

Furthermore, the result shows a large discrepancy in the precision of CDL's retrieved crop type labels between pixel-level and object-level assessments. The object-based crop type labels from CropSight exhibits higher consistency in representing crop types, as opposed to the pixel-based labels from CDL, which tends to include a mixture of crop types in agricultural fields (Fig. 17). Overall, these results demonstrate the ability of CropSight to provide reliable and consistent crop type ground truth across diverse crop species at the four study areas.

Using CropSight, we produce object-based crop type ground truth maps of four distinct sites with each situated in one of our four study areas in 2023 (Fig. 18). The collected crop type ground truth is notably dense, demonstrating CropSight's capability to potentially acquire a large number of crop type ground truth labels across those areas. In addition, there is a precise alignment of the collected object-based crop type data from fine-tuned SAM with the field boundaries of the background PlanetScope images (Fig. 18(b)), suggesting the effective boundary delineation ability of CropSight. It is worth noting that these maps are generated using the within-season street view and satellite images in 2023 when the corresponding CDL has not yet been produced. These mapping results further indicate the potential of CropSight in retrieving within-season object-based crop type ground truths at large scales when the corresponding street view and satellite imagery becomes publicly accessible online.

5. Discussion

In the study, we propose an innovative CropSight framework to efficiently retrieve object-based crop type ground truth using street view and satellite imagery. CropSight integrates the operational cropland field-view imagery collection method, the advanced uncertainty-aware crop type image classification model (UncertainFusionNet), and the cutting-edge promptable cropland boundary delineation model (SAM) to systematically obtain massive ground truth locations of a diversity of crop types and associated field boundaries. It streamlines the crop type data collection process, significantly reducing the time and labor traditionally required for field surveys, and provides a promising alternative to conventional crop type mapping products for crop type ground truthing with high accuracy. CropSight demonstrates significant potential for wall-to-wall crop type mapping through its collected crop type labels, which align well with the corresponding distributions in the CDL (Fig. S5). As the first framework to retrieve crop type ground truth at the object level, CropSight significantly expands pixel-level ground truth of existing studies (Pott et al., 2021; Yan and Ryu, 2021; Laguarta et al., 2024) with potentially more enriched agricultural field characteristics (e.g., within-object and between-object characteristics), facilitating more accurate crop distribution mapping and subsequent agricultural applications. Furthermore, CropSight's innovative Bayesian approach to estimating the uncertainty in deep learning model predictions introduces a measure of confidence for each identified crop type label. This approach greatly increases the accuracy of the final crop type labels retrieved from street view imagery, ensuring the quality of derived crop type ground truth. By integrating these two advanced components with a specially designed method for cropland field-view imagery collection, CropSight facilitates the large-scale, operational acquisition of high-quality, object-level crop type labels.

Through the combined use of a suite of devised street imagery filters and a spatially-adapted sampling strategy, the cropland field-view imagery collection method offers CropSight an operational means to

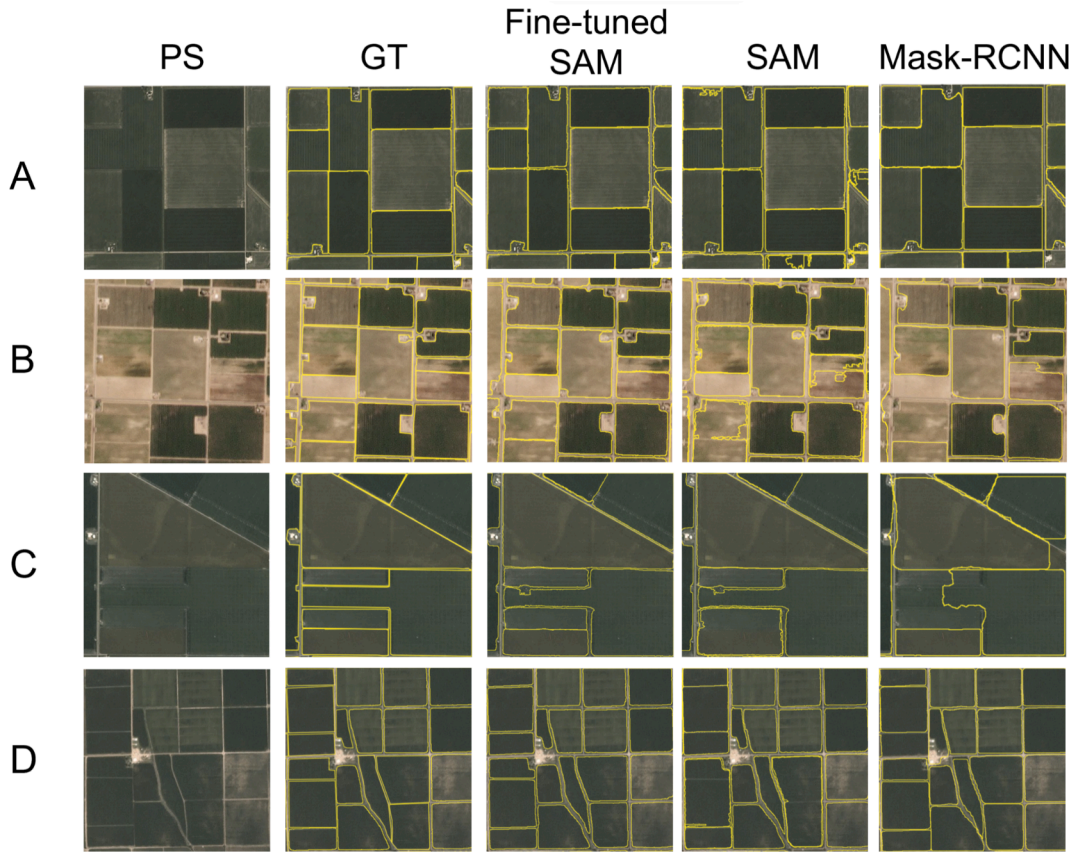

Fig. 15. Visualization of typical boundary delineation results from fine-tuned SAM, base SAM, and Mask-RCNN across four study areas A-D. These boundaries are derived from predictions on the CropBoundary test datasets of the respective study area. The columns, from left to right, represent original PlanetScope images (PS), ground truth boundaries (GT), boundaries derived from fine-tuned SAM, from base SAM, and from Mask-RCNN overlaid on PS.

Table 4
Overall accuracy of CDL's and CropSight's crop type labels at pixel- and object-levels across four study areas using object-based crop type ground truth as reference.

Study area	CDL		CropSight	
	Pixel-level	Object-level	Pixel-level	Object-level
A	0.9332	0.9695	1.00	0.9732
B	0.7586	0.8276	0.9944	0.9436
C	0.8433	0.9218	0.9942	0.9301
D	0.8426	0.8923	0.9701	0.9648

effectively collect representative high-quality field-view imagery of croplands at large scales. It enhances efficiency and reduces costs by eliminating the need to collect all available street view imagery within a study area for crop type labeling (Ringland et al., 2019; Wu et al., 2021; Yan and Ryu, 2021). Furthermore, the enhancement of street view imagery largely reduces uncertainties in identifying crop types. Traditional approaches frequently involve downsampling to fit large street view images into deep learning models, which can lead to resolution degradation and loss of detail (Ringland et al., 2019; Wu et al., 2021; Yan and Ryu, 2021; Laguarta et al., 2024). Our method reduces the size of imagery by dynamically pinpointing the cropland's central region. It preserves the essential visual features and overall integrity of crops in the imagery without compromising the spatial resolution, thereby ensuring the quality of street view imagery for subsequent crop type classification. While initially designed for GSV images, this operational collection method can be easily adapted for other street view image datasets (e.g., Baidu Total View, KartaView, Mapillary). With the rise of autonomous vehicles, an increasing number of cars are equipped with cameras. The cropland field-view imagery collection method can be also extended to handle vast image datasets captured by these vehicle-mounted cameras in the future.

The UncertainFusionNet model, an innovative Bayesian convolutional neural network, plays a crucial role in CropSight for precise crop type labelling in complex agricultural landscapes. By integrating the complementary strengths of two leading image classification architectures, UncertainFusionNet could capture both discriminant local and global visual features of crop plants from field-view imagery. This feature fusion design significantly enhances the model's ability to accurately identify crop types with similar structures and morphologies, a challenge that standalone convolutional neural network models (e.g., ResNet-50 and Inception v3) have encountered in previous research (Kang et al., 2018; Ringland et al., 2019; Wu et al., 2021; Yan and Ryu, 2021; Zou and Wang, 2021; Paliyam et al., 2021; d'Andrimont et al., 2022; Laguarta et al., 2024;). Additionally, its Bayesian approach allows for the estimation of uncertainty in each prediction, offering valuable insights into the confidence level of each crop type label. With entropy and variance as the uncertainty metrics, this approach quantifies the reliability of model's predictions and enhances decision-making processes by characterizing crop type labels of varying degrees of confidence to potentially guide further data collection or model refinement efforts. This novel uncertainty-aware method largely alleviates the common challenge faced in previous studies of labeling images of crop types with comparable visual features (e.g., leaf shape and canopy morphology) throughout their growth stages, such as images of grass and rice, or images of cotton and soybean (Fig. S3) (Ringland et al., 2019; Wu et al., 2021; Yan and Ryu, 2021; d'Andrimont et al., 2022; Laguarta et al., 2024).

SAM is employed in CropSight to extract the cropland boundary of

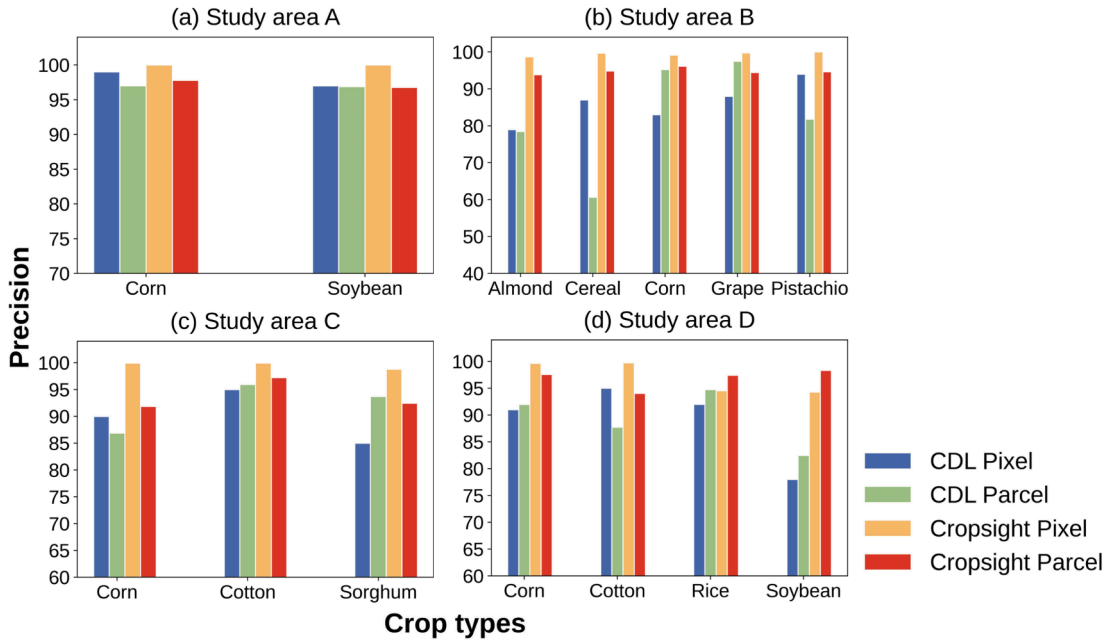

Fig. 16. Precision of CDL's and CropSight's crop type labels at pixel- and object-levels for each dominant crop species across four study areas A-D using the collected object-based crop type ground truth.

each geotagged field-view imagery. As a pioneering promptable segmentation model, SAM utilizes point prompts taken from field-view imagery locations to guide the boundary delineation of each roadside crop field from PlanetScope satellite images, achieving a recall value of 1. It effectively addresses the issue of missed cropland boundaries, a common shortfall in traditional boundary delineation models (Zhang et al., 2021; Jong et al., 2022; Mei et al., 2022; Cai et al., 2023). For example, although Mask-RCNN matches SAM in terms of precision, dice, and IoU metrics for detected cropland boundaries, its traditional non-prompt-based design, which employs a sequential process of object detection and boundary delineation, often results in overlooked cropland boundaries with lower recall and F1 scores. In addition, fine-tuning SAM with the high-quality cropland boundary dataset (i.e., CropBoundary) effectively resolves its zero-shot generalization challenges in segmenting agricultural fields from satellite imagery. The original SAM model, trained on a dataset primarily composed of natural imagery, faces performance degradation when applied to satellite images, particularly in complex scenarios where croplands display similar colors and canopy textures, resulting in obscure boundaries across fields. This fine-tuning process adapts SAM from its original focus on natural imagery to meet the unique needs of agricultural landscapes, enhancing its ability to precisely delineate cropland boundaries and differentiate fields of similar visual characteristics from satellite imagery. However, the fine-tuned SAM may still face challenges in accurately delineating productive boundaries for different crop types within a single field from satellite images, particularly for fields that exhibit uniform canopy features (e.g., color and texture) throughout the growing season in satellite imagery.

The CropSight framework exhibits strong applicability and generalizability in retrieving accurate crop type labels across various landscapes and species, as evidenced by the 2023 ground truth data collected in four study areas and Brazil (Fig. 18 and Fig. S4). Its consistently superior performance highlights its potential to complement (or as a promising alternative to) the conventional method of deriving crop type ground truth from satellite-based crop type products. Even for crops characterized by relatively low producer and user accuracies in the crop type product CDL of our study site (Liu et al., 2004), CropSight could achieve a consistent overall accuracy over 93 %. This disparity in performance is primarily attributed to the different data sources and methods employed for acquiring crop type labels. The satellite-based crop type products are typically generated by analyzing crop phenological patterns and temporal changes in crop spectral characteristics from satellite imagery time series. Dense satellite imagery ensures extensive coverage of satellite-based crop type products, enabling the retrieval of abundant ground truth data for various crop types. Yet the temporal and phenological patterns of crop species are affected by a combination of climatic and environmental factors, and may vary largely over space and time. Also the phenological patterns of certain crop types may be comparable, further affecting the accuracy and reliability of the ground truth data sourced from satellite-derived products. By contrast, CropSight leverages visual features (e.g., leaf shape, plant structure, canopy morphology, and flowering) learned from street view images for crop type classification. These high-resolution, close-up visualizations of the crops' physical characteristics enable the accurate identification of crops that may share similar phenological growth patterns but exhibit distinct morphological details, particularly in regions with complex agricultural landscapes and a diverse array of dominant crop species (Ringland et al., 2019; Wu et al., 2021; Yan and Ryu, 2021; Laguarta et al., 2024). Compared to satellite imagery, the visual features of the same crop type from street imagery are more consistent over space and time, facilitating more generalizable crop type labeling retrieval. Moreover, current crop type products are typically generated at the pixel level, with possibly mixed pixels in the same cropland area. By integrating crop type image classification and associated cropland boundary delineation, CropSight is pioneering in providing high-quality crop type ground truth at the object level. It eliminates the pixel mixture issue from different crop types within a field, as well as enriches the ground truth information of crop types with cropland boundary associated characteristics.

While the CropSight framework possesses its unique advantages and shows promising performance in retrieving object-based crop type ground truth, there still exist limitations. Large-scale applications of CropSight may be financially expensive due to the costs associated with using commercial data (i.e., GSV and PlanetScope). Within CropSight, UncertainFusionNet may need to be further adapted and refined to classify fields intercropped with various crop types using street view images. A potential adaptation is to enrich the training dataset with images of intercropping scenarios. This expansion will enable the model to differentiate between solitary plants and interacting plants as distinct

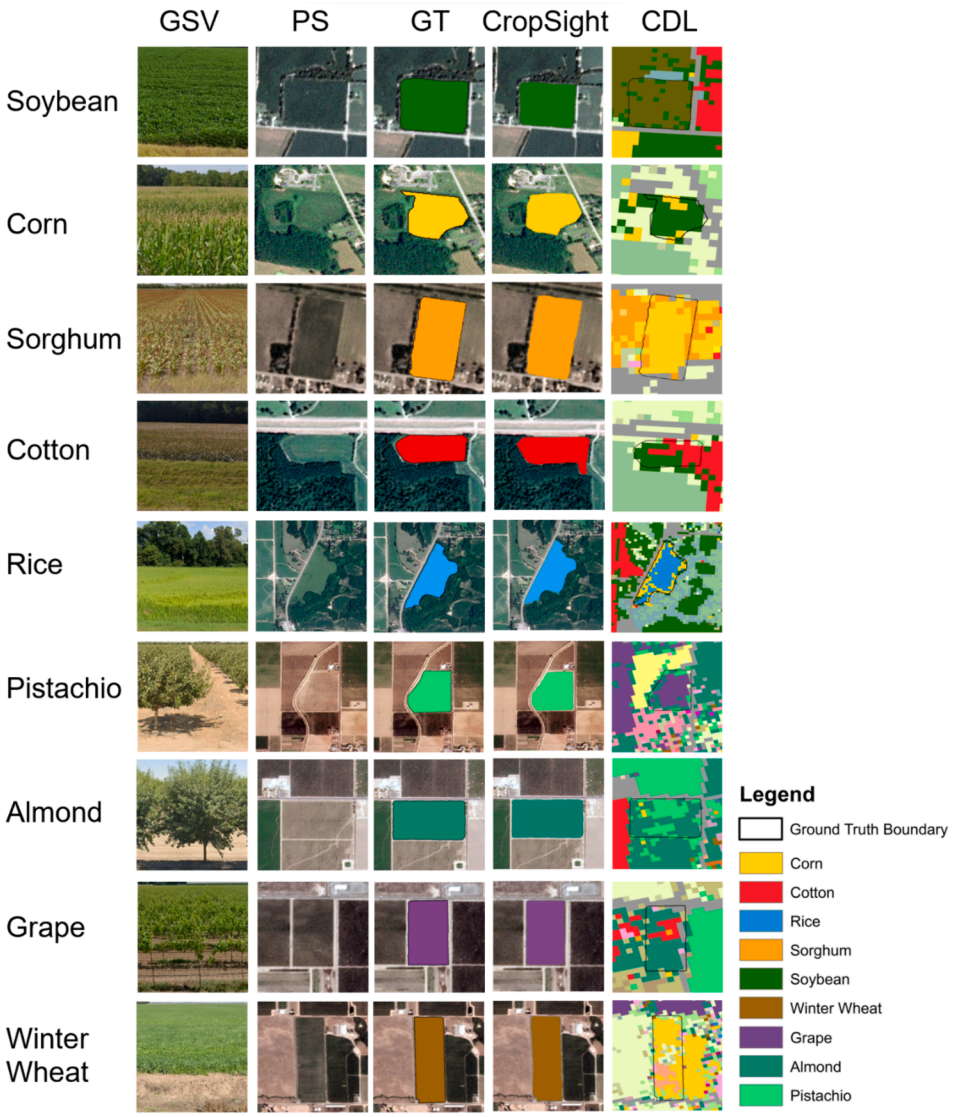

Fig. 17. Examples showcasing roadside field-view images (GSV), PlanetScope imagery (PS), crop type ground truth (GT), and crop type labels from CropSight and CDL, of all studied crop species.

categories, thereby refining the retrieval of crop type information for intercropping systems. Furthermore, the availability of street view imagery may be unevenly distributed, possibly more prevalent in urban areas than in rural regions. This imbalance may impact the distribution of the crop type ground truth data collected by CropSight (Fig. S5). Additionally, crop species from the same family, such as small grain cereals (e.g., rice and winter wheat), can exhibit relatively similar visual traits, including color, growth patterns, and morphology. These similarities pose challenges in accurately differentiating certain crop types within the same family from street view imagery. A potential solution lies in combining street view imagery with temporal remote sensing data. This integration takes into account both the rich visual crop features from street view imagery and the crop spectral and phenological characteristics from remote sensing time series (Diao, 2020; Diao et al., 2021), possibly facilitating more robust and improved identification of crop species within a family.

Expanding CropSight's application to national and even international scales will be beneficial for the creation of a more extensive and all-encompassing global dataset of object-based crop type ground truth. By expanding to include a variety of commercial street view imagery and satellite platforms, CropSight can significantly enhance its data collection capabilities worldwide. In regions without GSV, alternative street view image sources (e.g., Baidu Total View, KartaView, and Mapillary) could be utilized to expand the data sources for retrieving crop type labels. In countries with smallholder fields, higher spatial resolution satellite imagery (e.g., WorldView) can be employed to provide clearer canopy features for boundary delineation. This collected rich object-based crop type data can be instrumental in providing ground truth data for extensive crop distribution mapping, especially for regions without crop type mapping products. Furthermore, the CropSight framework shows capacity for advancing near-real-time crop type mapping, which is crucial for timely evaluations of weather impacts on agriculture, aiding in early detection of food security risks and rapid damage assessments (Gao and Zhang, 2021; Yang et al., 2023; Yang et al., 2024;). As autonomous vehicles become more prevalent, CropSight can uniquely capitalize on this trend by repurposing navigation images to acquire precise ground truth data on crop types. This approach enables rapid collection of detailed, object-based crop type ground truth, essential for timely and operational mapping of diverse crop species over large areas. Along with crop yield maps, crop type maps can further be leveraged to analyze drivers of yield gaps of each crop species and overall crop productivity (Zhang and Diao, 2023), especially for those regions with diverse agricultural practices and varying crop rotations. The assessment of the gaps between achieved and attainable

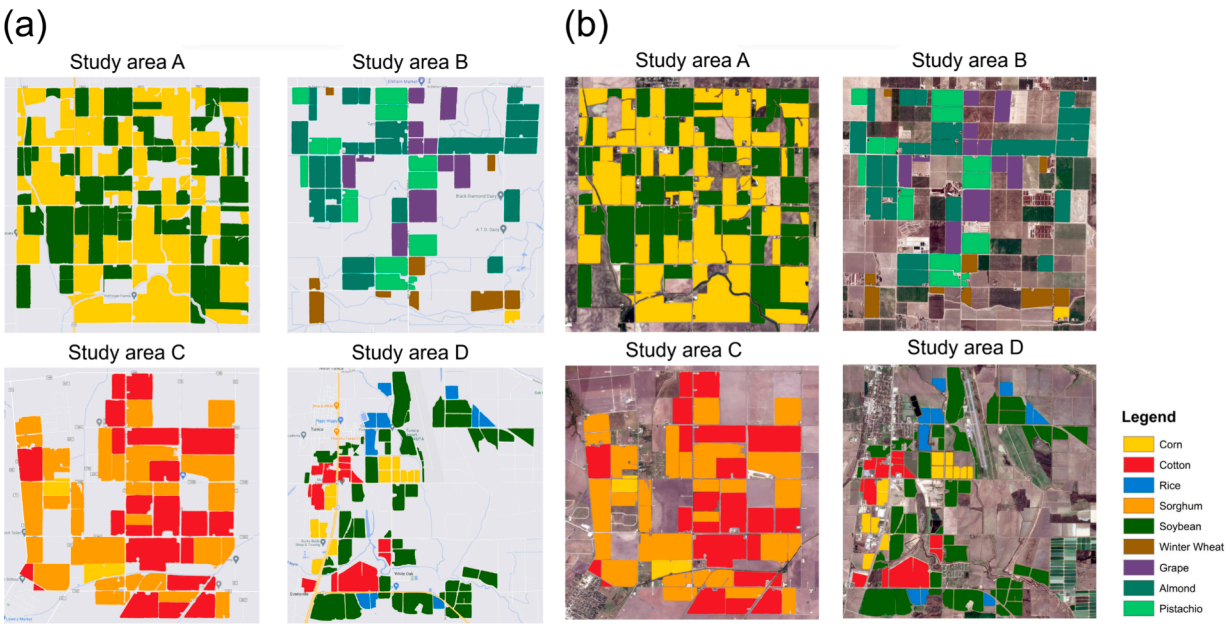

Fig. 18. Maps of object-based crop type ground truth produced by CropSight using latest images (2023). These maps represent four distinct study areas, each situated in one of our four study areas A-D. (a) Displays the overlay of crop type labels on Google Maps. (b) Displays the overlay of crop type labels on off-season PlanetScope images.

yields can enable more informed and strategic crop species planting decision, enhancing agricultural productivity and resilience, as well as driving more effective policy-making for addressing food insecurity.

6. Conclusion

In our study, we develop an innovative deep learning-based CropSight modeling framework to retrieve object-based crop type ground truth by synthesizing Google Street View and PlanetScope satellite images. CropSight comprises three key components: large-scale operational cropland field-view imagery collection method, uncertainty-aware crop type image classification model (UncertainFusionNet), and cropland boundary delineation model (SAM). Across four agriculturally dominated regions in the US, CropSight consistently achieves an overall accuracy of around 97 % in retrieving crop type label for multiple dominant crop species and an F1 score of approximately 92 % in delineating cropland boundaries. With the feature fusion and Bayesian classification module design, UncertainFusionNet surpasses benchmark image classification models (i.e., ResNet-50 and ViT-B16) in the identification of crop types from collected geotagged field-view images. The uncertainty quantification of CropSight further enhances the quality of retrieved crop type ground truth labels. With the promptable design, the fine-tuned SAM outperforms benchmark segmentation models (i.e., the base SAM and Mask-RCNN) in delineating boundaries corresponding to each geotagged field-view image with improved F1 and the recall being 1. CropSight also shows high potential in timely retrieving object-based crop type ground truth given the low latency of Google Street View and PlanetScope satellite images. Overall, the CropSight framework enables large-scale operational collection of crop type ground truth across various study areas and crop species at the object level, without requiring in-situ field observation. The retrieved ground truth is critical in advancing within-season crop type mapping and crop species-specific growth monitoring, aiding in timely decision-making for building more sustainable agricultural systems.

CRediT authorship contribution statement

Yin Liu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Chunyuan Diao:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Weiye Mei:** Methodology, Writing – review & editing. **Chishan Zhang:** Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the funding support partly by the National Science Foundation (2048068), partly by the National Aeronautics and Space Administration (80NSSC21K0946), and partly by the United States Department of Agriculture (2021-67021-33446).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2024.07.025.

References

Abdar, M., Fahami, M.A., Chakrabarti, S., Khosravi, A., Pławiak, P., Acharya, U.R., Tadeusiewicz, R., Nahavandi, S., 2021a. BARF: a new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification. Inf. Sci. 577, 353–378. https://doi.org/10.1016/j.ins.2021.07.024.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S., 2021b. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf. Fusion 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008.

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google street view: capturing the world at street level. Computer 43, 32–38. https://doi.org/10.1109/MC.2010.170.

Arco, J.E., Ortiz, A., Ramírez, J., Martínez-Murcia, F.J., Zhang, Y.-D., Górriz, J.M., 2023. Uncertainty-driven ensembles of multi-scale deep architectures for image classification. Inf. Fusion 89, 53–65. https://doi.org/10.1016/j.inffus.2022.08.010.

Belgiu, M., Csillik, O., 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. Remote Sens. Environ. 204, 509–523. https://doi.org/10.1016/j.rse.2017.10.005.

Bellvert, J., Adeline, K., Baram, S., Pierce, L., Sanden, B.L., Smart, D.R., 2018. Monitoring crop evapotranspiration and crop coefficients over an almond and pistachio orchard throughout remote sensing. Remote Sens. 10, 2001. https://doi.org/10.3390/rs10122001.

Bennett, A.J., Bending, G.D., Chandler, D., Hilton, S., Mills, P., 2012. Meeting the demand for crop production: the challenge of yield decline in crops grown in short rotations. Biol. Rev. 87, 52–71. https://doi.org/10.1111/j.1469-185X.2011.00184.x.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: a review. Landsc. Urban Plann. 215, 104217 https://doi.org/10.1016/j.landurbplan.2021.104217.

Blickensdörfer, L., Schwieder, M., Pflugmacher, D., Nendel, C., Erasmi, S., Hostert, P., 2022. Mapping of crop types and crop sequences with combined time series of Sentinel-1, Sentinel-2 and Landsat 8 data for Germany. Remote Sens. Environ. 269, 112831 https://doi.org/10.1016/j.rse.2021.112831.

Bolfe, É.L., Jorge, L.A.D.C., Sanches, I.D., Luchiari Júnior, A., Da Costa, C.C., Victoria, D. D.C., Inamasu, R.Y., Grego, C.R., Ferreira, V.R., Ramirez, A.R., 2020. Precision and digital agriculture: adoption of technologies and perception of Brazilian farmers. Agriculture 10, 653. https://doi.org/10.3390/agriculture10120653.

Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. Geocarto Int. 26, 341–358. https://doi.org/10.1080/10106049.2011.562309.

Braga, J.R.G., Peripato, V., Dalagnol, R., Ferreira, P.M., Tarabalka, Y.O.C., Aragão, L.E., De Campos Velho, H.F., Shiguemori, E.H., Wagner, F.H., 2020. Tree crown delineation algorithm based on a convolutional neural network. Remote Sens. 12, 1288. https://doi.org/10.3390/rs12081288.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., Li, Z., 2018. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sens. Environ. 210, 35–47. https://doi.org/10.1016/j.rse.2018.02.045.

Cai, Z., Hu, Q., Zhang, X., Yang, J., Wei, H., Wang, J., Zeng, Y., Yin, G., Li, W., You, L., Xu, B., Shi, Z., 2023. Improving agricultural field parcel delineation with a dual branch spatiotemporal fusion network by integrating multimodal satellite data. ISPRS J. Photogramm. Remote Sens. 205, 34–49. https://doi.org/10.1016/j.isprsjprs.2023.09.021.

Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., Zhang, Q., Qiu, G., 2018. Integrating aerial and street view images for urban land use classification. Remote Sens. 10, 1553. https://doi.org/10.3390/rs10101553.

d'Andrimont, R., Yordanov, M., Lemoine, G., Yoong, J., Nikel, K., Van Der Velde, M., 2018. Crowdsourced street-level imagery as a potential source of in-situ data for crop monitoring. Land 7, 127. https://doi.org/10.3390/land7040127.

d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Van Der Velde, M., 2022. Monitoring crop phenology with street-level imagery using computer vision. Comput. Electron. Agric. 196, 106866 https://doi.org/10.1016/j.compag.2022.106866.

Dakir, A., Bachir Alami, O., Barramou, F., 2020. Crop type mapping using optical and radar images: a review, in: 2020 IEEE International Conference of Moroccan Geomatics (Morgeo). Presented at the 2020 IEEE International conference of Moroccan Geomatics (Morgeo), pp. 1–8. doi: 10.1109/Morgeo49228.2020.9121869.

Di Tommaso, S., Wang, S., Lobell, D.B., 2021. Combining GEDI and Sentinel-2 for wall-to-wall mapping of tall and short crops. Environ. Res. Lett. 16, 125002 https://doi.org/10.1088/1748-9326/ac358c.

Diao, C., 2020. Remote sensing phenological monitoring framework to characterize corn and soybean physiological growing stages. Remote Sens. Environ. 248, 111960 https://doi.org/10.1016/j.rse.2020.111960.

Diao, C., Yang, Z., Gao, F., Zhang, X., Yang, Z., 2021. Hybrid phenology matching model for robust crop phenological retrieval. ISPRS J. Photogramm. Remote Sens. 181, 308–326. https://doi.org/10.1016/j.isprsjprs.2021.09.011.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: transformers for image recognition at scale. doi: 10.48550/arXiv.2010.11929.

Fatchurrachman, R., Soh, N.C., Shah, R.M., Giap, S.G.E., Setiawan, B.I., Minasny, B., 2022. High-resolution mapping of paddy rice extent and growth stages across peninsular Malaysia using a fusion of sentinel-1 and 2 time series data in google earth engine. Remote Sens. 14, 1875. https://doi.org/10.3390/rs14081875.

Gal, Y., Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceedings of the 33rd International Conference on Machine Learning. *Proceedings of The 33rd International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 48:1050-1059 Available from https://proceedings.mlr.press/v48/gal16.html.

Gallo, I., Ranghetti, L., Landro, N., La Grassa, R., Boschetti, M., 2023. In-season and dynamic crop mapping using 3D convolution neural networks and sentinel-2 time series. ISPRS J. Photogramm. Remote Sens. 195, 335–352. https://doi.org/10.1016/j.isprsjprs.2022.12.005.

Gao, F., Zhang, X., 2021. Mapping crop phenology in near real-time using satellite remote sensing: challenges and opportunities. J. Remote Sens. 2021, 8379391. https://doi.org/10.34133/2021/8379391.

Goel, R., Garcia, L.M.T., Goodman, A., Johnson, R., Aldred, R., Murugesan, M., Brage, S., Bhalla, K., Woodcock, J., 2018. Estimating city-level travel patterns using street imagery: a case study of using Google Street View in Britain. PLoS One 13, e0196521. https://doi.org/10.1371/journal.pone.0196521.

Gour, M., Jain, S., 2022. Uncertainty-aware convolutional neural network for COVID-19 X-ray images classification. Comput. Biol. Med. 140, 105047 https://doi.org/10.1016/j.compbiomed.2021.105047.

Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. Remote Sens. Environ. 220, 135–151. https://doi.org/10.1016/j.rse.2018.10.031.

Gupta, A., Ramanath, R., Shi, J., Keerthi, S.S.. Adam vs. SGD: Closing the generalization gap on image classification. https://www.opt-ml.org/papers/2021/paper53.pdf.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D., 2023. A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45, 87–110. https://doi.org/10.1109/TPAMI.2022.3152247.

He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2018. Mask R-CNN. CoRR abs/1703.06870.. http://arxiv.org/abs/1703.06870.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image. Recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 770–778. https://doi.org/10.48550/arXiv.1512.03385.

Hou, Y., Biljecki, F., 2022. A comprehensive framework for evaluating the quality of street view imagery. Int. J. Appl. Earth Obs. Geoinf. 115, 103094 https://doi.org/10.1016/j.jag.2022.103094.

Hu, Y., Zeng, H., Tian, F., Zhang, M., Wu, B., Gilliams, S., Li, S., Li, Y., Lu, Y., Yang, H., 2022. An Interannual transfer learning approach for crop classification in the Hetao Irrigation District, China. Remote Sens. 14, 1208. https://doi.org/10.3390/rs14051208.

Jia, M., Wang, Z., Mao, D., Ren, C., Wang, C., Wang, Y., 2021. Rapid, robust, and automated mapping of tidal flats in China using time series Sentinel-2 images and Google Earth Engine. Remote Sens. Environ. 255, 112285 https://doi.org/10.1016/j.rse.2021.112285.

Jiang, C., Guan, K., Huang, Y., Jong, M., 2024. A vehicle imaging approach to acquire ground truth data for upscaling to satellite data: a case study for estimating harvesting dates. Remote Sens. Environ. 300, 113894 https://doi.org/10.1016/j.rse.2023.113894.

Johnson, D.M., Mueller, R., 2021. Pre- and within-season crop type classification trained with archival land cover information. Remote Sens. Environ. 264, 112576 https://doi.org/10.1016/j.rse.2021.112576.

Jong, M., Guan, K., Wang, S., Huang, Y., Peng, B., 2022. Improving field boundary delineation in ResUNets via adversarial deep learning. Int. J. Appl. Earth Obs. Geoinf. 112, 102877 https://doi.org/10.1016/j.jag.2022.102877.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS J. Photogramm. Remote Sens. 145, 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R. 2023. Segment Anything. arXiv Preprint arXiv:2304.02643. https://doi.org/10.48550/arXiv.2304.02643.

Kussul, N., Lemoine, G., Gallego, F.J., Skakun, S.V., Lavreniuk, M., Shelestov, A.Y., 2016. Parcel-based crop classification in Ukraine using Landsat-8 data and Sentinel-1A data. IEEE J. Selected Top. Appl. Earth Obs. Remote Sens. 9, 2500–2508. https://doi.org/10.1109/JSTARS.2016.2560141.

Li, Y., Hu, M., Yang, X., 2023. Polyp-SAM: Transfer SAM for Polyp Segmentation. In: Chen, W., Astley, S.M. (Eds.), Proceedings of Medical Imaging 2024: Computer-Aided Diagnosis 12927. https://doi.org/10.1117/12.3006809.

Laguarta, J., Friedel, T., Wang, S., 2024. Combining deep learning and street view imagery to map smallholder crop types.Proceedings of the AAAI Conference on Artificial Intelligence 38 (20). https://doi.org/10.1609/aaai.v38i20.30225.

Li, Y., Xu, W., Chen, H., Jiang, J., Li, X., 2021. A novel framework based on mask R-CNN and histogram thresholding for scalable segmentation of new and old rural buildings. Remote Sens. 13, 1070. https://doi.org/10.3390/rs13061070.

Lin, C., Zhong, L., Song, X.-P., Dong, J., Lobell, D.B., Jin, Z., 2022. Early- and in-season crop type mapping without current-year ground truth: Generating labels from historical information via a topology-based approach. Remote Sens. Environ.. 274, 112994 https://doi.org/10.1016/j.rse.2022.112994.

Liu, W., Gopal, S., Woodcock, C.E., 2004. Uncertainty and confidence in land cover classification using a hybrid classifier approach. Photogramm. Eng. Remote Sens. 70, 963–971. https://doi.org/10.14358/PERS.70.8.963.

Luo, Y., Zhang, Z., Zhang, L., Han, J., Cao, J., Zhang, J., 2022. Developing high-resolution crop maps for major crops in the European union based on transductive transfer learning and limited ground data. Remote Sens. 14, 1809. https://doi.org/10.3390/rs14081809.

Mei, W., Wang, H., Fouhey, D., Zhou, W., Hinks, I., Gray, J.M., Van Berkel, D., Jain, M., 2022. Using deep learning and very-high-resolution imagery to map smallholder field boundaries. Remote Sens. 14, 3046. https://doi.org/10.3390/rs14133046.

Ok, A.O., Akar, O., Gungor, O., 2012. Evaluation of random forest method for agricultural crop classification. Eur. J. Remote Sens. 45, 421–432. https://doi.org/10.5721/EuJRS20124535.

Oliphant, A.J., Thenkabail, P.S., Teluguntla, P., Xiong, J., Gumma, M.K., Congalton, R.G., Yadav, K., 2019. Mapping cropland extent of Southeast and Northeast Asia using multi-year time-series Landsat 30-m data using a random forest classifier on the Google Earth Engine Cloud. Int. J. Appl. Earth Obs. Geoinf. 81, 110–124. https://doi.org/10.1016/j.jag.2018.11.014.

Osco, L.P., Wu, Q., de Lemos, E.L., Gonçalves, W.N., Ramos, A.P.M., Li, J., Marcato, J., 2023. The segment anything model (SAM) for remote sensing applications: from zero to one shot. Int. J. Appl. Earth Obs. Geoinf. 124, 103540 https://doi.org/10.1016/j.jag.2023.103540.

Paliyam, M., Nakalembe, C., Liu, K., Nyiawung, R., Kerner, H., 2021. Street2Sat: A Machine Learning Pipeline for Generating Ground-truth Geo-referenced Labeled

Datasets from Street-Level Images. Proceedings of the 38th International Conference on Machine Learning.
Pott, L.P., Amado, T.J.C., Schwalbert, R.A., Corassa, G.M., Ciampitti, I.A., 2021. Satellite-based data fusion crop type classification and mapping in Rio Grande do Sul, Brazil. ISPRS J. Photogramm. Remote Sens. 176, 196–210. https://doi.org/10.1016/j.isprsjprs.2021.04.015.
Ringland, J., Bohm, M., Baek, S.-R., 2019. Characterization of food cultivation along roadside transects with Google Street View imagery and deep learning. Comput. Electron. Agric. 158, 36–50. https://doi.org/10.1016/j.compag.2019.01.014.
Schmedtmann, J., Campagnolo, M., 2015. Reliable crop identification with satellite imagery in the context of common agriculture policy subsidy control. Remote Sens. 7, 9325–9346. https://doi.org/10.3390/rs70709325.
Shamsi, A., Asgharnezhad, H., Tajally, A., Nahavandi, S., Leung, H., 2023. An Uncertainty-aware Loss Function for Training Neural Networks with Calibrated. arXiv Preprint arXiv:2110.03260. https://doi.org/10.48550/arXiv.2110.03260.
Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv Preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556.
Som-ard, J., Immitzer, M., Vuolo, F., Ninsawat, S., Atzberger, C., 2022. Mapping of crop types in 1989, 1999, 2009 and 2019 to assess major land cover trends of the Udon Thani Province, Thailand. Comput. Electron. Agric. 198, 107083 https://doi.org/10.1016/j.compag.2022.107083.
Taesiri, M.R., Nguyen, G., Habchi, S., Bezemer, C.-P., Nguyen, A. 2023. ImageNet-Hard: The Hardest Images Remaining from a Study of the Power of Zoom and Spatial Biases in Image Classification.
Tran, K.H., Zhang, H.K., McMaine, J.T., Zhang, X., Luo, D., 2022. 10 m crop type mapping using Sentinel-2 reflectance and 30 m cropland data layer product. Int. J. Appl. Earth Obs. Geoinf. 107, 102692 https://doi.org/10.1016/j.jag.2022.102692.
Vuolo, F., Neuwirth, M., Immitzer, M., Atzberger, C., Ng, W.-T., 2018. How much does multi-temporal Sentinel-2 data improve crop type classification? Int. J. Appl. Earth Obs. Geoinf. 72, 122–130. https://doi.org/10.1016/j.jag.2018.06.007.
Waldner, F., Diakogiannis, F.I., 2020. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. Remote Sens. Environ. 245, 111741 https://doi.org/10.1016/j.rse.2020.111741.
Wang, S., Azzari, G., Lobell, D.B., 2019. Crop type mapping without field-level labels: random forest transfer and unsupervised clustering techniques. Remote Sens. Environ. 222, 303–317. https://doi.org/10.1016/j.rse.2018.12.026.
Wang, S., Waldner, F., Lobell, D.B., 2022. Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. Remote Sens. 14, 5738. https://doi.org/10.3390/rs14225738.
Wu, F., Wu, B., Zhang, M., Zeng, H., Tian, F., 2021. Identification of crop type in crowdsourced road view photos with deep convolutional neural network. Sensors 21, 1165. https://doi.org/10.3390/s21041165.
Xu, X., Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., Luo, D., 2022. Associations between street-view perceptions and housing prices: subjective vs. objective measures using computer vision and machine learning techniques. Remote Sens. 14, 891. https://doi.org/10.3390/rs14040891.
Yan, Y., Ryu, Y., 2021. Exploring Google street view with deep learning for crop type mapping. ISPRS J. Photogramm. Remote Sens. 171, 278–296. https://doi.org/10.1016/j.isprsjprs.2020.11.022.
Yang, Z., Diao, C., Gao, F., 2023. Towards scalable within-season crop mapping with phenology normalization and deep learning. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 1390–1402. https://doi.org/10.1109/JSTARS.2023.3237500.
Yang, Z., Diao, C., Gao, F., Li, B., 2024. EMET: An emergence-based thermal phenological framework for near real-time crop type mapping. ISPRS Journal of Photogrammetry and Remote Sensing 215, 271–291. https://doi.org/10.1016/j.isprsjprs.2024.07.007.
Yin, L., Cheng, Q., Wang, Z., Shao, Z., 2015. 'Big data' for pedestrian volume: exploring the use of Google Street View images for pedestrian counts. Appl. Geogr. 63, 337–345. https://doi.org/10.1016/j.apgeog.2015.07.010.
Yordanov, M., d'Andrimont, R., Martinez-Sanchez, L., Lemoine, G., Fasbender, D., Van Der Velde, M., 2023. Crop identification using deep learning on LUCAS crop cover photos. Sensors 23, 6298. https://doi.org/10.3390/s23146298.
Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W. 2022. ESA WorldCover 10 m 2021 v200 (Version v200) [Data set]. https://doi.org/10.5281/zenodo.5571936.
Zhang, C., Di, L., Lin, L., Li, H., Guo, L., Yang, Z., Yu, E.G., Di, Y., Yang, A., 2022. Towards automation of in-season crop type mapping using spatiotemporal crop information and remote sensing data. Agric. Syst. 201, 103462 https://doi.org/10.1016/j.agsy.2022.103462.
Zhang, C., Diao, C., 2023. A Phenology-guided Bayesian-CNN (PB-CNN) framework for soybean yield estimation and uncertainty analysis. ISPRS J. Photogramm. Remote Sens. 205, 50–73. https://doi.org/10.1016/j.isprsjprs.2023.09.025.
Zhang, H., Liu, M., Wang, Y., Shang, J., Liu, X., Li, B., Song, A., Li, Q., 2021. Automated delineation of agricultural field boundaries from Sentinel-2 images using recurrent residual U-Net. Int. J. Appl. Earth Obs. Geoinf. 105, 102557 https://doi.org/10.1016/j.jag.2021.102557.
Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018. An object-based convolutional neural network (OCNN) for urban land use classification. Remote Sens. Environ. 216, 57–70. https://doi.org/10.1016/j.rse.2018.06.034.
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009.
Zou, S., Wang, L., 2021. Detecting individual abandoned houses from google street view: a hierarchical deep learning approach. ISPRS J. Photogramm. Remote Sens. 175, 298–310. https://doi.org/10.1016/j.isprsjprs.2021.03.020.