Dear reviewers and editors,

Thank you very much for your helpful and constructive feedback. In response, we have made clarifications and improvements to the manuscript, both in contents and in visual presentation, which we believe enhance the clarity and quality of the study. Detailed, point-by-point responses are shown below, with reference to the specific line numbers of tracked changes ("All-Markup" version) made in the updated manuscript.

Reviewer 1's Comments

This paper proposed an object-based crop type ground truth dataset from 2013 to 2023, called "CropSight-US", using street view and Sentinel-2 images. This dataset covered 17 major crop types across 294 Agricultural Statistics Districts (ASDs) in CONUS. Specifically, crop type labels were extracted from Google street view images and field boundary information was derived from Sentinel-2 imagery. CONUS-UncertainFusionNet was developed to generate this dataset with uncertainty information. In the experimental results, the performance of several deep learning-based networks was compared. This paper is well-organized and the reviewer has the following detailed comments:

Response: We sincerely thank Reviewer 1 for the thorough evaluation of our manuscript and for the clear summary of our work. We appreciate the recognition of the study's organization and the constructive feedback provided. Below, we address each of the reviewer's specific comments point by point and describe the corresponding revisions made in the manuscript.

Specific Comments:

1. This paper included that "CropSight-US, the first national scale, object-based crop type ground truth dataset". The authors were encouraged to review the existing products to highlight this contribution.

Response: Thank you for the opportunity to clarify the position of our contribution. We have revised the Introduction to better situate CropSight-US within the landscape of existing national agricultural datasets. Specifically, we have reviewed current crop type ground truth datasets in [Lines 46–70], categorizing them into traditional field survey-based and crop product-derived ground truth datasets.

We further clarify the technical limitations of existing datasets [Lines 71-92]. Most available datasets are primarily pixel-level, and the relatively few object-based products are typically generated through one of two approaches: 1) aligning pixel-level classifications with external field boundary datasets, or 2) clustering classified pixels into polygons (i.e., objects) based on spatial and spectral similarity, such as the Crop Sequence Boundary (CSB). Both approaches introduce additional post-processing steps that can affect scalability and boundary accuracy, thereby influencing the quality and consistency of object-based products at large spatial scales.

Finally, to avoid potential overstatement, we have replaced "the first" with "novel" in both the Abstract [Lines 19–21] and Introduction [Line 95]. This adjustment more accurately reflects the unique scope and

methodological contribution of our dataset without disregarding the existence of smaller-scale or differently processed object-based products.

2. This article contains multiple instances of the full name and abbreviation for CONUS. It is recommended to provide the full name only upon first mention, using the abbreviation thereafter.

Response: Thank you for this helpful suggestion. We have updated the manuscript so that the full name of CONUS is provided only at its first occurrence (excluding the abstract or the captions of tables and figures), which appears at [Line 96] in the Introduction section of the manuscript. The abbreviation is used consistently throughout the remainder of the manuscript.

3. For line 138 and 655, these links were invalid. Please double check all the links in the manuscript.

Response: Thank you for checking the links and bringing these issues to our attention. We appreciate your careful review and have corrected the invalid links as noted.

Regarding Line 138 (now [Line 145]): The documentation of all Street View-related APIs has been consolidated into a single page. Previously, we used an older link to showcase the endpoint API street view imagery downloading. Apologies for the confusion. We have now updated the link to point directly to the Google Street View Static API documentation page at:
https://developers.google.com/maps/documentation/streetview/overview

Regarding the link to our CropSight-US GEE application (now [Line 675]): We believe the issue occurred during the conversion from Word to PDF, where the hyphen ("-") in the URL was omitted (the GEE app is managed under the google cloud project [ee-azzhou249], as the same link is also provided in [Line 735] for the Data Availability section). The correct full address should be:
https://ee-azzhou249.projects.earthengine.app/view/cropsight-us

We have verified and corrected these links in the updated manuscript.

4. What are the advantages of using Sentinel-2 over NAIP imagery?

Response: Thank you for raising this question. The advantages of using Sentinel-2 over NAIP imagery in our workflow mainly relate to timeliness and transferability. We have revised the manuscript to explicitly clarify this rationale and to describe how Sentinel-2 availability and NAIP supplementation were handled in our data processing workflow [Lines 162-164, Line 167 and Lines 276-278].

Each Google Street View imagery captures crop conditions during a specific month of a given year. To correctly delineate the corresponding field boundary for that same period, we need satellite imagery that is both temporally aligned and cloud-free. Sentinel-2 provides frequent revisit cycles (5 days at mid-latitudes) with Sentinel-2A launched in 2015 and Sentinel-2B in 2017, which greatly increases the likelihood of obtaining a cloud-free or minimally cloudy scene close to the date of the Street View image. In contrast, NAIP imagery is typically collected only once every 1–3 years, which makes it difficult to match the exact year and growing season represented in the Street View photos. Timely alignment is

essential for ensuring that the field boundary accurately reflects the field configuration at the time of observation.

Additionally, while NAIP offers very high spatial resolution, it is available only for the United States. For applications beyond the U.S., using NAIP would not be feasible. Sentinel-2, on the other hand, provides global coverage with consistent acquisition schedules and harmonized data quality. This makes the workflow more flexible and transferable, enabling extension of the CropSight framework to other countries or regions without relying on region-specific aerial programs.

In CropSight-US, each object-based ground truth unit includes both the crop-type label identified from Street View imagery and its corresponding field boundary derived from the closest suitable Sentinel-2 scene. This ensures that the final dataset is both temporally consistent and scalable to broader geographical contexts. Only for years prior to full Sentinel-2 availability (before 2017) or when suitable cloud-free imagery was not available, NAIP imagery was used as a supplementary source.

5. What is the experimental environment of ViT-B16 and ResNet-50 training, such as GPU and CPU? The authors are suggested to offer more details about the setting of the proposed network.

Response: Thank you for asking this question. All model training was conducted on the Illinois Campus Cluster Program (ICCP) at the University of Illinois Urbana-Champaign. The computational infrastructure consists of dual AMD EPYC 7763 processors, 512 GB of system memory, and four NVIDIA A100 GPUs with 80 GB of GPU memory each, connected through 25G high-speed networking nodes. We have updated the relevant section of the manuscript to clearly document these specifications for both benchmark models (ViT-B16 and ResNet-50) as well as our proposed CONUS-UncertainFusionNet, ensuring that the experimental setup is transparent and reproducible [Lines 377-380].

6. For all the tables, it is recommended to display the best performance in bold to help readers better follow.

Response: Thank you for this suggestion. We have updated the manuscript to highlight the tabulated rows of best performance in bold to enhance the readability.

7. SAM is trained on high-resolution natural images, and Sentinel-2 has a lower resolution. How to overcome this resolution shift?

Response: Thank you and we appreciate the opportunity to clarify this point and updated the manuscript to provide further explanation [Lines 392-396]. To address the resolution shift between SAM's pre-training domain (high-resolution natural images) and Sentinel-2 imagery, we fine-tune the mask decoder of SAM using the training split of the field-boundary component of the CropGSV-Ref dataset, while keeping both the image encoder and prompt encoder frozen (Section 3.1.4). This lightweight fine-tuning strategy has been shown to be an effective and computationally efficient way to adapt SAM to remote sensing tasks, including cropland field-boundary segmentation, without requiring full model retraining. Prior studies have demonstrated that decoder-only adaptation substantially improves SAM's performance on delineating agricultural boundaries while maintaining low computational cost (Liu et al., 2024; Pu et al., 2025). We

follow this established approach to ensure that SAM can better capture field-level boundaries from Sentinel-2 imagery despite its lower spatial resolution.

Reference:

Liu, Y., Diao, C., Mei, W., and Zhang, C.: CropSight: Towards a large-scale operational framework for object-based crop type ground truth retrieval using street view and PlanetScope satellite imagery, ISPRS J. Photogramm. Remote Sens., 216, 66–89, https://doi.org/10.1016/j.isprsjprs.2024.07.025, 2024.

Pu, X., Jia, H., Zheng, L., Wang, F., and Xu, F.: Classwise-sam-adapter: Parameter efficient fine-tuning adapts Segment Anything to SAR domain for semantic segmentation, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., https://doi.org/10.1109/JSTARS.2025.3532690, 2025.

8. For Figure 6, there are some image patches after image embedding. Do they refer to image embeddings? The authors are suggested to check this as image patches are not equal to image embeddings.

Response: Thank you for highlighting this issue. As you noted, image embeddings are high-level feature representations derived from the model and are not visually equivalent to raw image patches. To address this, we have revised Figure 6 to reflect the proper representation of image embeddings and updated the corresponding description "image embedding" [386] to avoid confusion.

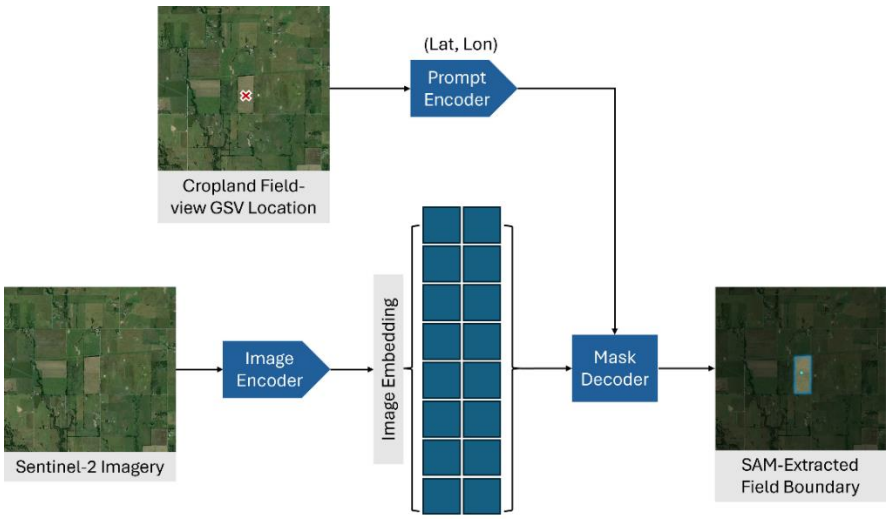

Figure 1: Structure of the SAM with cropland field-view GSV location coordinate as the point prompt for cropland field boundary delineation from Sentinel-2 imagery (© European Union/ESA/Copernicus, processed via Google Earth Engine).

9. For line 353, there is an error regarding reference "(Error! Reference source not found.)."

Response: Thank you for pointing out this issue. We have corrected the broken reference (now at [Line 365]) and updated it to properly point to [Fig. 3], where the corresponding crop types we have identified through street view imagery across CONUS is shown. The manuscript has been checked accordingly to ensure all references are correctly linked.

10. How to choose reasonable thresholds for entropy and variance?

Response: Thank you for this helpful question. In our method, we select the uncertainty (i.e., entropy and variance) thresholds to balance classification confidence with sufficient sample retention for ground-truth collection. We expanded our explanations of the process in the updated manuscript [Lines 370-377].

Specifically, to identify reasonable thresholds for achieving this trade-off, we analyze the empirical distributions of entropy and variance for correctly versus incorrectly classified samples during model training. By examining where the density curves of these two groups intersect (Fig. 5), we select threshold values that best separate confident predictions from uncertain ones. This intersection-based criterion (Abdar et al., 2021; Gour and Jain, 2022; Arco et al., 2023) provides a balance between retaining as many field-view samples as possible and ensuring that the retained samples have sufficiently high confidence. These thresholds provide a consistent confidence standard across all crop types, ensuring that a given entropy or variance value represents the same level of model ambiguity regardless of the crop, which facilitates uniform data filtering for downstream applications

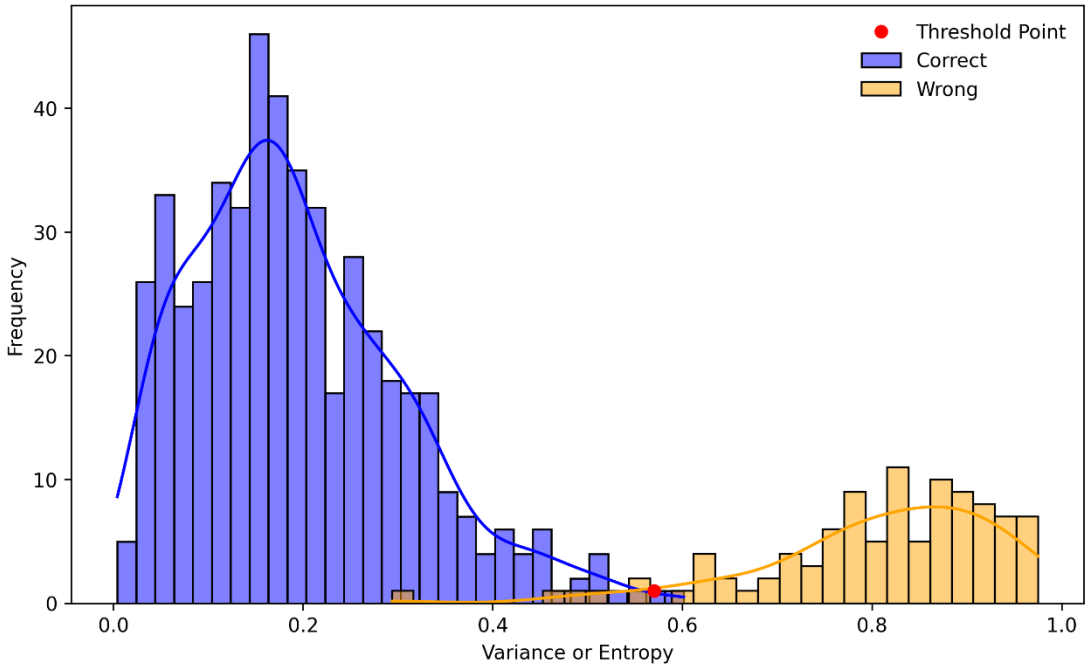

Figure 2: Distributions of uncertainty scores for correctly (blue) and incorrectly (orange) classified samples produced by the CONUS-UncertainFusionNet model. The uncertainty threshold is determined at the intersection point of the two density curves, representing the optimal separation between high- and low-confidence predictions.

Reference

Abdar, M., Fahami, M. A., Chakrabarti, S., Khosravi, A., Pławiak, P., Acharya, U. R., Tadeusiewicz, R., and Nahavandi, S.: BARF: A new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification, Inf. Sci., 577, 353–378, doi:10.1016/j.ins.2021.07.047, 2021.

Arco, J. E., Ortiz, A., Ramírez, J., Martínez-Murcia, F. J., Zhang, Y.-D., and Górriz, J. M.: 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, 2023.
Gour, M., and Jain, S.: Uncertainty-aware Convolutional Neural Network for COVID-19 X-ray Images Classification, Comput. Biol. Med., 140, Article 105047, https://doi.org/10.1016/j.compbiomed.2021.105047, 2022.

Reviewer 2's Comments

This manuscript introduces CropSight-US, a national-scale, object-based crop type ground truth dataset for the contiguous United States (2013–2023), derived from Google Street View imagery and Sentinel-2–based field boundary delineation. The dataset is novel in its integration of street-level imagery at national scale and in its object-level design with uncertainty metrics. However, several fundamental aspects of dataset representativeness, temporal consistency, and uncertainty interpretation remain insufficiently documented, limiting confidence in its general applicability.

Response: We thank Reviewer 2 for the careful evaluation and for highlighting both the novelty of CropSight-US and the areas requiring clarification. We have carefully revised the manuscript to provide additional documentation on dataset representativeness, temporal consistency, and the interpretation of uncertainty metrics.

Specific Comments:

1. All samples are derived from road-accessible GSV imagery, yet the manuscript provides no quantitative assessment of resulting spatial bias. Without statistics on distance-to-road coverage or cropland representativeness at the ASD level, it is difficult to evaluate how well CropSight-US captures agricultural landscapes away from road networks.

Response: We appreciate the reviewer's comment regarding potential spatial bias associated with the use of road-accessible GSV imagery. To address this, we conducted an ASD-level analysis of road-field-proximity. Using the 2023 CSB dataset as a proxy for the total agricultural landscape, we sampled 5% of all fields in each ASD and calculated the distance from field boundaries to the nearest road network. To ensure these distances reflect functional management units rather than classification artifacts, we dissolved adjacent CSB polygons sharing the same 2023 crop type label prior to sampling. Across the sampled fields in ASDs, the average distance from field boundaries to roads was 112.32 ± 74.66 m at the ASD level. As for our CropSight-US dataset, the field-to-road distances range from under 10 m up to 150 m. Therefore, CropSight-US could mostly capture conditions representative of typical agricultural field configurations in CONUS.

Additionally, much of the U.S. cropland landscape, particularly in the Midwest, is organized under the Public Land Survey System (PLSS), which establishes a regular one-mile grid of section lines along which rural roads are commonly constructed (Abreu et al., 2017). This grid-based infrastructure results in a high proportion of agricultural management units being road-adjacent, thereby enhancing the spatial coverage and practical representativeness of roadside imagery relative to regions with more irregular road networks. However, we admit that reliance on road-accessible GSV imagery may constrain the spatial representativeness of the GSV-derived ground-truth dataset, particularly in heterogenous agricultural regions

characterized by sparse road networks. We have revised the manuscript to explicitly address these considerations and clarify their implications for spatial representativeness [Lines 486-493, Lines 625-628, Lines 749-756].

Reference:

Abreu, A., Lawson, L. A., Hyman, M., Hardin, R., and Gerling, M.: Collecting data from a permanent grid sampling frame via a mobile mapping instrument, NASS Res. Rep., 322824, United States Department of Agriculture, National Agricultural Statistics Service, 2017. https://doi.org/10.22004/ag.econ.322824

2. The selection of Sentinel-2 (or NAIP) imagery "closest in time" to GSV acquisition is not quantified. The manuscript should report typical and maximum temporal offsets and discuss potential impacts on both crop labeling and boundary delineation.

Response: Thank you for the suggestion to clarify the temporal alignment of GSV and the associated satellite imagery for cropland field boundary delineation. To ensure temporal alignment between GSV observations for crop type labeling and satellite imagery for corresponding field-boundary delineation, we relied on the Month–Year metadata provided by GSV, which specifies the capture month of each panorama (no finer temporal scale information can be retrieved from GSV yet). For every labeled GSV image, we searched for the highest-quality, least-cloudy Sentinel-2 acquisition within a ±1-month window around the GSV capture month. This yields a typical temporal offset of 0–30 days, with a maximum allowable offset of 60 days when suitable scenes appear only at the edges of the search window. When Sentinel-2 scenes within this window were unavailable or had excessive cloud cover or seasonal mismatch (e.g., pre-greenup or post-harvest phenology), we used NAIP aerial imagery as a fallback, selecting the temporally closest NAIP acquisition to maintain temporal consistency. This ensures that SAM-based delineation accurately captures the productive field boundaries corresponding to the crop type visible from the GSV imagery. We have updated the descriptions in Section 3.1.2 to reflect this clarification in methods [Line 276-278].

3. The manuscript places strong emphasis on uncertainty as a key advantage of the dataset. However, it remains unclear how users are expected to interpret and use the provided uncertainty metrics in practice. In particular, it is not specified whether the reported entropy and variance values are directly comparable across crop types, or whether they should be interpreted as crop-specific relative measures. Without such clarification, the practical value of the uncertainty information for downstream applications is difficult to assess.

Response: Thank you for this important request for clarification. To ensure the dataset is practical for downstream applications, the reported entropy and variance values are directly comparable absolute measures across all crop types, rather than crop-specific relative measures. As shown in Section 3.1.3 and Figure 5, we achieved this by determining the high-confidence using the aggregated density distributions of the entire validation split, rather than calculating independent thresholds for each crop. We have expanded the explanation of uncertainty metrics in Section 3.1.3 of the manuscript [Line 371-377], and in our GEE interface as well.

4. The authors group wheat, rice, and other small-grain cereals into a single "cereal" class based on their visual similarity in street-level imagery. While this decision is understandable from an

engineering and labeling perspective, it introduces substantial risks at the application level that are not sufficiently discussed. In particular, rice paddies are typically characterized by surface flooding and distinct water management practices, whereas wheat is associated with fundamentally different hydrological and energy conditions. This raises the concern that water-related signals may be implicitly mixed into the "cereal" class, potentially degrading the reliability of rice-related samples. The manuscript should more explicitly discuss these implications and the limitations they impose on crop-specific analyses.

Response: We appreciate your insightful observation regarding the distinct hydrological and energy conditions associated with rice paddies compared to wheat and other small-grain cereals. We fully agree that grouping these crops into a single Cereal class introduces potential hydrological mixing, particularly the risk that water-related signals characteristic of flooded rice fields may be implicitly incorporated into samples labeled as cereals. We have expanded the Discussion to explicitly address this limitation.

In the revised manuscript, we now clarify that our current grouping is driven by the visual similarity of canopy architecture in roadside GSV imagery, where wheat, rice, rye, and triticale often exhibit comparable field-level appearances. Because the labeling relies on static, ground-level images, surface flooding and water-management structures common in rice systems are not always reliably visible from the road, which further complicates species-level separation. To mitigate potential hydrological mixing for downstream users, we note in the revision that researchers focusing on rice-specific applications should apply secondary water-related filters, such as Normalized Difference Water Index (NDWI) or other moisture indices derived from satellite imagery, when attempting to disaggregate the Cereal class into more specific crop categories. This allows users to screen for likely flooded or wet environments that correspond more closely to paddy rice conditions.

We have updated the manuscript in Section 6: Discussion and Conclusion to address this comment [Lines 742-749].

Technical Issues:

Sect. 3.1.3 contains an unresolved reference ("Error! Reference source not found.").

Response: Thank you for pointing out this issue. We have corrected the broken reference at Line 353 (now at [Line 365]) and updated it to properly point to Fig.3, where the corresponding crop types we have identified through street view imagery across CONUS is shown. The manuscript has been revised accordingly to ensure all references are correctly linked.