RC2: 'Comment on essd-2023-73', Anonymous Referee #2, 23 Aug 2023

Summary

Comments: This work constructs a 100000-level FASDD based on multi-source heterogeneous flame and smoke images, which provides a challenging benchmark to drive the continuous evolution of fire detection models. However, there are still some issues that need to be addressed to demonstrate the reliability of the dataset.

Response: Many thanks for your positive comments and valuable suggestions on our study, which help evidently improve our manuscript. We greatly appreciate your expertise and the time you have dedicated to evaluating our paper. According to your suggestions and comments, we carefully revised the manuscript. We sincerely hope that the revisions we have made adequately address your issues and demonstrate our commitment to providing a robust and valuable contribution to the field of fire detection. We have thoroughly considered each of your suggestions and have conducted comprehensive modifications and refinements to our experiments, aiming to further validate the reliability and universality of our dataset. Once again, we sincerely thank you for your valuable insights.

Major comments

1. Comments: In terms of sample annotation, what are the differences between satellite images with different spatial resolutions in the remote sensing field and RGB images in the CV field? Can different annotation formats be integrated or unified into one annotation format for the widespread use of FASDD?

Response: Thank you very much for bringing up these comments. In terms of sample annotation, both satellite images with different spatial resolutions in the remote sensing field and RGB images in the CV field were annotated using a consistent manual visual labeling approach facilitated by the LabelImg software. The main difference is that the manual annotation process can be directly applied to CV images, while remote sensing images require a series of preprocessing steps, including true/false color synthesis and bit-depth conversion, before annotation can take place. Due to the significantly large volume of satellite images, manually selecting images containing fire-related data would have been exceedingly labor-intensive. Consequently, we employed a fire detection model trained on CV data to conduct an initial screening of the extensive satellite image dataset, identifying potential images with fire-related information. Subsequently, the selected remote sensing images, which may contain fire-related objects, underwent manual annotation.

Furthermore, in line with your suggestion, all images, both from the remote sensing and CV fields, have indeed been integrated and standardized into a uniform annotation format to facilitate the widespread use of FASDD. In fact, we have not only unified the label format across all images but have also provided four distinct unified formats to cater to the requirements of various deep learning frameworks. Among these, the label files in TXT format are tailored for YOLO series models, the XML format serves models oriented towards VOC datasets, and the JSON format is compatible with models tailored for COCO datasets or those adhering to the TDML specification.

Changes in manuscript: We have made refinements to the detailed description of the data annotation process in the paper to ensure greater clarity. For specific changes, please refer to Section 3.3.

2. Comments: How to consider the difference in spatial resolution between Sentinel-2 L1C and Landsat-8 TOA? How to solve the detection accuracy of targets with different sizes? What are the differences in detection accuracy between large and small targets during sample annotation, model training, and inference stages?

Response: Thank you for your valuable insights. In consideration of the spatial resolution difference between satellite data products, we employed different image sizes to ensure that most of the fire and smoke targets within the images fall within a reasonable size range. Specifically, for Landsat data with a spatial resolution of 30 meters, we set the image size to approximately 1000×1000 pixels, while for Sentinel-2 data with a spatial resolution of 10 meters, we configured the image size to be around 2000×2000 pixels. This approach ensured that the actual spatial coverage of fire images ranged from approximately 20×20 km to 30 × 30 km. Visual inspection revealed that this spatial range allowed for better consideration of both large and small fire targets. It is worth noting that due to anomalous reflectance values of smoke objects in Landsat-8 atmospheric correction data, we have discontinued the use of previous Landsat-8 data. In the currently available remote sensing data, we have retained only Sentinel-2 satellite imagery as sample data.

Furthermore, based on the dataset proposed in this study and its various domain-specific sub-datasets, we conducted inference experiments using the Swin Transformer model to compare the differences in detection accuracy between targets of different scales. Following the segmentation criteria of the COCO dataset (Lin et al., 2014), objects with bounding box dimensions smaller than 32×32 pixels were classified as small targets. Subsequently, we compared the accuracy differences between large and small targets during both the model training and inference phases, ss shown in Table A1. The experiments indicated that the Swin Transformer model exhibited superior detection performance on large targets but had room for improvement in detecting small targets. This may be attributed to the Swin Transformer's proficiency in capturing global contextual relationships while having relatively weaker local attention. To address the issue of low accuracy in small target detection, some studies have proposed using pyramid or multi-scale approaches to enhance the model's detection accuracy for targets of various sizes.

Changes in manuscript: We have updated the data generation process diagram (Figure 1) and revised the associated description of data sources in Section 3.1. Furthermore, to investigate differences in detection accuracy between large and small targets in images from the Computer Vision (CV), Unmanned Aerial Vehicle (UAV), and Remote Sensing (RS) domains, we conducted accuracy comparisons using the Swin Transformer (ST) model on both the validation and test sets. Since our main objective was to introduce a large-scale benchmark dataset, rather than optimizing a state-of-the-art deep learning model, we have included comparison experiments in Appendix A.

3. Comment: How to consider the difference in radiometric resolution between RGB and satellite images? How robust is the detection algorithm for data with different radiation resolutions? Please confirm through ablation experiments.

Response: Thank you very much for comments and suggestions. Regarding the difference in radiometric resolution between RGB and satellite images, we first harmonized the satellite images through a uniform preprocessing pipeline to create RGB images. Subsequently, we transformed the 16-bit radiometric resolution RGB images into 8-bit radiometric resolution RGB images. This transformation enabled the integration of remote sensing data with standard 8-bit computer vision images for neural network training.

Furthermore, in order to ascertain the robustness of training models using 8-bit data, in line with your suggestion, we conducted additional ablation experiments to explore the impact of the difference in radiometric resolution between RGB and satellite images. In this experiment, we compared the model performance between those trained using 16-bit and 8-bit radiometric resolution images, and the experimental results are presented in Table B1 in the manuscript's appendix. The results indicate that models trained on 16-bit radiometric resolution remote sensing data perform comparably but with a slight performance gap when compared to models trained on 8-bit radiometric resolution data. This observation may be attributed to the possibility that 8-bit images may filter out some of the complex redundant features present in 16-bit images, thereby making our research objectives more distinct and easier to identify. Consequently, it can be inferred that 8-bit remote sensing images effectively capture the characteristic information of fire-affected areas and background features, thereby yielding satisfactory model performance.

Changes in manuscript: To investigate the difference in radiometric resolution, we have included supplementary experiments in the appendix. Please refer to Appendix B for specific details regarding these modifications.

4. Comments: The FASDD holds rich variations in image size, resolution, illumination, scenario, image range, viewing angle, platform, and data source. How to consider data at different scales for various deep learning models? Please choose the latest model in the CV field to verify the reliability of FASDD and the superiority of

Transformer-based models, such as DETR, etc.

Response: Thank you for your valuable suggestion. The inclusion of a variety of image sizes in our dataset was deliberate, as we cannot guarantee uniform image dimensions captured by all sensors. This diversity in data sizes enhances the dataset's generalizability, making it more suitable for deployment on different sensors. During model training, we ensure consistency by resizing all images to a uniform size (e.g. 1333×800) to align with the neural network's training requirements.

Furthermore, following your recommendation, we have removed previous outdated models and conducted comparative experiments using the latest models, such as transformer-based models (DETR and Swin Transformer) and CNN-based models (YOLOv5x). The experimental results, as presented in Table D1, reveal that DETR exhibits model performance similar to YOLOv5x, while the more advanced Swin Transformer significantly outperforms YOLOv5x in terms of model accuracy.

Changes in manuscript: In order to provide a clearer description of the data processing workflow for data at different scales, we have made adjustments to the content in the Experiment Setup section. Please refer to Section 4.1 for specific details regarding these modifications. Additionally, we have included supplementary performance comparison experiments between transformer and CNN architecture models. Specific changes related to this can be found in Appendix D.

5. Comments: What may be the reason for the poor validation performance of the transformer-based model compared to YOLOv5x?

Response: Thank you for your question. In light of a reminder from another reviewer, we detected a minor issue within our previous image dataset. Then we revised the dataset by removing the problematic data, and retrained the models and conducted an evaluation of their performance using the updated dataset. Table D1 presents the performance evaluation results for various models. The results indicate that, whether on the validation set or the test set, the Swin Transformer consistently exhibited superior model performance compared to YOLOv5x.

Changes in manuscript: Please refer to Appendix D for specific details regarding these modifications.

6. Comments: The annotation and partition ratio of samples directly affect the accuracy of deep learning models. Please supplement ablation experiments to demonstrate the reliability and universality of FASDD datasets under different training sample ratios.

Response: We sincerely appreciate your concern regarding the accuracy of deep learning models under different training sample ratios. We conducted model training using common sample partition ratios of 8:1:1, 7:2:1, 6:2:2, and 1/2:1/3:1/6 for the datasets (Xia et al., 2018; Wang et al., 2023), as presented in Table C1. The experiments demonstrated that the FASDD dataset maintains excellent reliability and generalizability under various training sample ratios.

Irrespective of the chosen partition ratio, the Swin Transformer model consistently achieved approximately 80% mAP@0.5 and over 91.5% recall on both the validation and test sets. Notably, under the 1/2:1/3:1/6 distribution ratio, the test set's mAP@0.5 was approximately 4% higher compared to the other three partition methods. This is attributed to the inclusion of a larger number of validation set samples in this partition, resulting in models with stronger generalization capabilities. Consequently, the open-source dataset provided in this study adopts the 1/2:1/3:1/6 partition ratio.

Changes in the manuscript: We have incorporated new comparative experiments based on different sample partition ratios. Please refer to Appendix C for specific details regarding these modifications.

7. Comments: In the case of small samples, the Transformer-based model has poor convergence. The linear expansion of the transformer leads to a sharp increase in parameter size and insufficient local feature extraction. How to solve the problem of insufficient local feature extraction in transformer and poor target detection performance in FASDD?

Response: Thank you for bringing up this issue. To address the issue of insufficient local feature extraction in the transformer, one approach is to incorporate convolutional modules that introduce localized attention. Alternatively, multi-scale feature extraction modules can be employed to capture image features at different scales, further improving the model's ability to perceive local context. These strategies can enhance the model's ability to perceive local context, consequently improving the performance of the transformer.

To tackle the challenges associated with poor target detection performance in FASDD, techniques such as random scaling and other data augmentation methods can be employed to capture finer details of small targets and provide the model with learning samples of various sizes. This can help mitigate the issue of low accuracy in small target detection. Additionally, pretraining models on large-scale relevant datasets can assist the model to learn more feature distributions from targets of different sizes, which can help alleviate the problem of difficult convergence for small target training.

Changes in the manuscript: We conducted comparative experiments using transfer learning to investigate the impact of pretrained models on detection performance, as shown in Table 3. The experimental results indicate that the "pretraining + fine-tuning" transfer learning approach leads to further improvements in model performance. It also demonstrates the valuable contribution of our open-access large-scale heterogeneous data to the performance improvement of fire detection models.

8. Comments: Is the low accuracy of the Transformer-based model caused by overfitting? Please provide the training accuracy and validation accuracy curve of the Transformer-based model.

Response: Thank you very much for your suggestions. The low accuracy of the Transformer-based model may be attributed to various factors, including but not limited to overfitting, and possibly issues within our previous dataset. Following your advice, we retrained the model on the revised dataset and generated training loss curves and validation accuracy curves to visualize the training process.

Changes in manuscript: We have included training process curves for various classic models in Figure D1 and conducted an analysis of the model fitting performance. Please refer to the relevant content in Appendix D for specific details regarding these modifications.

9. Comments: Please increase the types and quantity of deep learning algorithms (each model type contains at least two algorithms) to fully validate the universality and reliability of the dataset.

Response: Thank you very much for your valuable suggestions. We have heeded your advice and removed the older deep learning models we previously employed. Instead, we have opted for lateset models with superior performance. Specifically, we conducted training with two distinct architectural paradigms: models based on a CNN architecture (YOLOv5x and InternImage) and models based on a transformer architecture (DETR and Swin Transformer). Given the substantial size of our dataset, training one model on a single GPU with 48G of memory takes approximately 15-20 days. Therefore, with limited computational resources, we did not add more models for comparison experiments. Table D1 presents the results of comparative experiments involving various classical models. The various experiments presented in the manuscript effectively establish the universality and reliability of our dataset, demonstrating exceptional representativeness and persuasiveness. We eagerly anticipate that other researchers will leverage our dataset to explore more advanced models, thereby fostering further advancements in the field of fire detection.

Changes in manuscript: Considering the two fundamentally distinct architectural paradigms, CNN and Transformer, we have chosen four models (two based on CNN architecture and two based on Transformer architecture) from the past three years, known for their comparative excellence, to conduct accuracy comparison. Furthermore, we have provided their accuracy curves during the training process. For specific details regarding these modifications, please refer to Table D1 and Figure D1 in Appendix D.

Reference

Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., ..., and Zhang, L.: DOTA: A large-scale dataset for object detection in aerial images, In Proceedings of the IEEE conference on computer vision and pattern recognition, 3974-3983, 2018.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, in: Computer Vision - ECCV 2014, Cham, 740-755, https://doi.org/10.1007/978-3-319-10602-1_48, 2014.

Wang, J., Li, X., Jin, L., Li, J., Sun, Q., and Wang, H.: An air quality index prediction model based on CNN-ILSTM. Scientific Reports, 12(1), 8373, 2022