SmokeViz: Using Pseudo-Labels to Develop a Human-Labeled Deep Learning Dataset of Wildfire Smoke Plumes in Satellite Imagery

Anonymous Author(s)
Affiliation
Address
email

Abstract

The global increase in the frequency and intensity of wildfires in the last decade underscores the need for advancements in fire monitoring techniques. In order to investigate deep learning approaches for detecting and tracking wildfires and the related human health impacts, we present SmokeViz, a large scale machine learning dataset of smoke plumes in satellite imagery. To build the dataset, we refine a set of expert-generated smoke annotations created by analysts at the National Oceanic and Atmospheric Administration. Each annotation gives a general temporal and geographical approximation of smoke plumes but at variable and, primarily, low temporal resolution. Along with the resulting SmokeViz dataset, we present pseudo-label dimension reduction (PLDR), a generalizable method that leverages the semi-supervised method, pseudo-labeling to refine datasets with mismatching temporal and/or spatial resolutions. Unlike typical pseudo-labeling applications that aim to increase the number of labeled samples, the objective of PLDR is to use intermediary pseudo-labels (IPLs) to refine an existing set of annotations so that each annotation is matched to the best candidate input data. For SmokeViz, we train a parent model to generate IPLs that are used to pinpoint the singular, most representative, satellite image within the analyst specified temporal range to match the smoke plume annotation. By identifying the most representative satellite image of a smoke plume for a given smoke plume annotation, this study produces a succinct and relevant deep learning dataset consisting of over correct number 180,000 manual annotations. The resulting SmokeViz dataset is anticipated to be an instrumental tool in developing further deep learning models for studying wildfires and is publicly available at [aws download link].

1 Introduction

In part, due to public policy, the average levels of fine particulate matter ($PM_{2.5}$) in the US have generally been declining over the past few decades [2]. Despite those improvements, the contribution of wildfire smoke to $PM_{2.5}$ concentrations in the US has more than doubled between 2010 to 2020, accounting for up to half of the overall $PM_{2.5}$ exposure in Western regions [7]. Increases in $PM_{2.5}$ due to wildfire smoke are concerning since ambient $PM_{2.5}$ exposure is a leading environmental risk factor for adverse health effects and premature mortality [13]. These risks underscore the necessity for efficient and effective monitoring methods to mitigate the adverse health impacts associated with wildfire smoke.

Without satellites, wildfire monitoring has relied on ground-based methods, such as forest service patrols, manned lookout towers, and aviation surveillance [3]. While these methods provide valuable localized insights, they are constrained by geographical and logistical limitations, often failing to

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

deliver timely and comprehensive data, especially over large and remote areas. In contrast, satellite imagery provides continuous, wide-area coverage and real-time streaming information crucial for assessing and responding to the health risks posed by wildfire smoke.

Satellite imagery, equipped with state-of-the-art sensors, such as the Advanced Baseline Imager (ABI) on the Geostationary Operational Environmental Satellites (GOES) [14], have revolutionized environmental monitoring. Unlike polar orbiting satellites such as the Suomi or Sentinel satellites, geostationary satellites maintain constant observation over an area that is necessary to capture the dynamic behavior of wildfire smoke plumes. In turn, GOES capabilities can provide critical insights into the concentration and movement of smoke particulates, facilitating real-time assessments of air quality.

Integrating satellite imagery into wildfire smoke monitoring provides real-time data that can improve the timeliness of public health planning and response. By mapping the spread and density of smoke, health authorities can issue prompt warnings, implement evacuation protocols, and deploy resources effectively to mitigate health risks. Furthermore, long-term data gathered from satellite observations can aid in understanding the broader impacts of wildfire smoke on public health, influencing policy decisions and preventive measures.

In addition, numerical models for real-time smoke dispersion currently have no smoke analysis product available for data assimilation [17, 1]. This results in delayed start up times for the smoke to begin being modeled and can result in further down-the-line errors. Providing a real-time data assimilation smoke product solely dependent on incoming satellite imagery has the potential to improve existing smoke dispersion models.

2 Related Work

2.1 Numerical

Currently, multi-channel thresholding is a popular method to distinguish smoke pixels from pixels containing dust, clouds or other phenomenon with similar signatures [37]. Thresholds are determined by using historical, labeled data to extract optimal radiance values for each channel that corresponds with the labeled class. These methods are tuned to particular biogeographies and often have issues with generalization to new locations with varying fuel types [27].

2.2 Analyst

In contrast to the numerical thresholding approach, human visual inspection of satellite imagery is another commonly used method for smoke identification. Trained analyst inspect satellite imagery and label the smoke by hand. An example of hand-labeled annotations is the National Oceanic and Atmospheric Administration (NOAA) Hazard Mapping System (HMS) fire and smoke product [23, 30]. For the HMS smoke product, trained satellite analysts use movement characteristics to help identify smoke by scanning through a time series of satellite imagery. When visual inspection indicates smoke, the analyst will draw a polygon that corresponds to the geolocation and density of smoke. By design of the product, the HMS annotations have varying time resolution and are released on a rolling but undefined schedule ranging from one to multiple times a day as observation conditions permit. If expanded beyond the current North American boundary, this method will not be as scalable as an automated approach and is limited by the availability of analysts and their time.

The HMS program, managed by NOAA, consists of an operational system that uses an aggregation of satellite data to generate active fire and smoke data. To train our model, we develop a supervised learning framework that uses the HMS analyst smoke product as truth labels during the model training process. HMS smoke analysis data gives the coordinates of the smoke perimeter as a polygon and classifies the smoke by density within a given time window. The time windows can range from instantaneous (same start/end time) to lengths of 22 hours. While the true bounds of the smoke can change within the larger time spans, the analyst is making an approximation that should reflect the smoke coverage over the duration of the time window. The density information is qualitatively determined by each analyst based on the apparent smoke opacity in the satellite imagery and categorized as either light, medium or heavy as seen in figure 1a [25].

2.3 Deep Learning

To address the challenges associated with thresholding and manual labels, we can look towards innovative approaches and recent technological advancements in computer vision. Machine learning methods have shown potential in improving the accuracy and efficiency of satellite-based wildfire smoke detection and monitoring. For instance, SmokeNet, uses a convolutional neural network (CNN) based framework to determine if a scene of MODIS satellite imagery contains smoke [4]. Another study, that looked at a singular wildfire event, also used a CNN to identify smoke on a pixel-wise basis using imagery from Himiwari-8 [20]. Additionally, Wen et al. developed a CNN architecture that takes GOES-East imagery as input and the HMS-generated annotations for the target labels during training [35]. Contrary to previous GOES-based machine learning datasets that include data from only one of the two operational GOES-series satellites, most commonly opting for GOES-East [35, 26, 22, 24], the annotations in SmokeViz are matched with either GOES-East or GOES-West based on which satellite has optimal coverage of the event.

The success of deep learning methods, such as CNNs, relies heavily on the availability of a large, representative dataset [33]. As laid out in table 1, prior studies use relatively small numbers of samples, where one sample represents a satellite image with a singular time and geolocation. In contrast, benchmark datasets for computer vision applications contain tens of thousands (CIFAR-10 [19] and MNIST [10]) to millions (CIFAR-100 and ImageNet [9]) of data samples compare to large remote sensing datasets. Keeping in mind the correlation between both the quality and quantity of data with model performance, we introduce the largest known smoke plume dataset, SmokeViz, containing over 180,000 samples.

Table 1: Comparison of satellite smoke plume datasets detailing the number of samples containing smoke, observational satellite, number of bands/channels included, how the labels were generated and whether the labels are for image scene classification (SC) or semantic segmentation (SS)and if the dataset is publicly available.

reference	# samples	satellite	# bands	label	task	avail.
[4]	1016	MODIS	5	students	SC	no
[35]	4095	GOES-East	5	HMS analysts	SS	no
[20]	975	Himiwari-8	7	algorithm	SS	no
[6]	125	Sentinel-2	3	crowdsourced	SC	yes
SmokeViz	183,672	GOES-East/West	3	HMS analysts	SS	yes

Semi-supervised learning is an approach that can be used to increase the number of labeled samples in a dataset. This is perfomed by leveraging a labeled dataset to generate new labels for an often larger, but unlabeled, dataset. Pseudo-labeling, a form of semi-supervised learning, is an iterative process that uses labeled data to train an initial parent model. The parent model is applied to unlabeled data to create pseudo-labels (PLs) that are then used to train a child model, if performed iteratively the child model will then create another round of PLs to train the next child model [21]. We introduce a non-iterative variation of pseudo-labeling, not to increase the size, but to reduce the noise and thus dimensionality of our dataset. This is done by generating intermediary PLs (IPLs) that are not used to train models but will serve as a tool to select the optimal satellite image out of the given time-window to best represent each smoke plume analyst annotation. The final resulting dataset, SmokeViz, is comprised expert smoke-specific analyst annotations of wildfire smoke plumes and corresponding GOES satellite imagery.

Although the SmokeViz dataset is anticipated to be primarily used for solving various wildfire smoke applications, the dataset has the potential to serve as a uniquely insightful test case for remote sensing semantic segmentation. Many existing remote sensing datasets have removed imagery with clouds [32, 16] and/or are focused on objects with sharp contrasts such as crops [16], human infrastructure [38], or clouds over oceans [18, 31], smoke has indistinct boundaries that often fade both spatially and temporally. Adding difficult to currate, high quality remote sensing specific datasets like SmokeViz to large-scale multi-category pre-training datasets such as SatlasPretrain [6] has the potential to advance generalizable remote sensing computer vision.

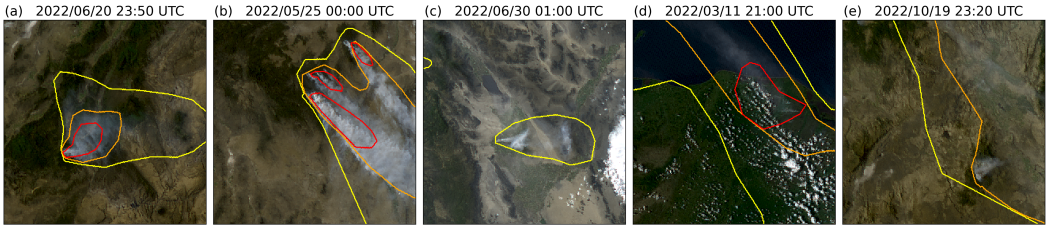

Figure 1: HMS smoke annotations overlaid on GOES imagery, the yellow, orange and red contours indicate the extent of light, medium and heavy density smoke. (a) and (b) show a canonical examples of a smoke plumes. (c)-(e) show observable variations in the density labels.

3 Methods

3.1 Datasets

The smoke plumes in our datasets are observed using the latest GOES operational satellites, GOES-16 (East), 17 and 18 (West) that each carry the ABI, that measure 16 bands between the visible and infrared wavelengths (λs) collected every 10 minutes for full disk imagery. Using PyTroll, a Python framework for processing satellite data [28], we input bands 1-3 (Table **??**) to a GOES-specific true color composite algorithm [5] to develop a, 1km resolution, true color image representation, similar to the imagery seen by HMS analysts. Discussed in further detail in the next section, we only include the first 3 out of 16 available bands due the smaller λs of light corresponding to the highest signal-to-noise ratio (SNR).

To take into account movement characteristics to help identify smoke, analysts use multi-frame animations of the satellite imagery. The resulting annotations primarily have time windows over multiple hours, with an average of 3 hours of imagery represents one smoke plume annotation. Since the goal of HMS smoke annotations is to show the general coverage over that time span, as shown in figure 2, the smoke boundaries don't often match up with the satellite imagery over the entire time window. One approach to this problem would be to use all the satellite images the analysts used as input. Since the timespans are non-uniform, this would vary the length in imagery inputs into the model, which would be difficult with a CNN architecture. Moreover, this would result in a smoke plume detection model that ingest a longer sequence of start-up imagery that requires additional memory and computational resources than a single input image. Instead of using the original analysts' many satellite image inputs to one annotated output, we develop a one-to-one input-to-output by identifying the optimal singular satellite image input to represent each analyst annotation.

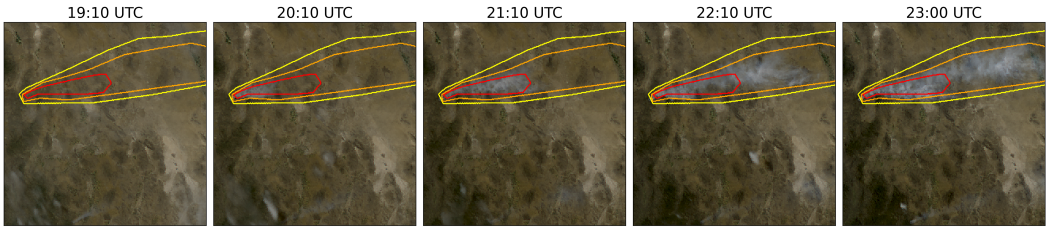

Figure 2: True color GOES-East imagery from May 5th, 2022, Southeast New Mexico (31.38°N, 107.87°W) during the start of the Foster Fire. The red, orange and yellow lines represent the heavy, medium and low density HMS smoke annotations that span 19:10–23:00 UTC.

As mentioned prior, rather than using one satellite or the cumulative data from both GOES-West and GOES-East images, we select between one or the other satellite based on the solar zenith angle (SZA). For smoke identification, this approach can achieve a much higher SNR than imaging the earth's surface from an arbitrary angle. While the atmosphere is composed primarily of molecules with size $<$1nm, smoke particles are larger in comparison, varying from 100 nm – 10 μm in diameter, d. When the λ of light $\lambda < d$, the elastic scattering of light against matter is modeled through Mie rather than Rayleigh scattering (figure 3) which contributes to two critical consequences: (1) the

forward scattering light that travels colinearly from sun to smoke, to satellite will have the highest SNR as seen in subfigures 5(a) vs 5(b) that show GOES-East providing a higher SNR near sunset compared to GOES-West. (2) The smaller the λ in relationship to d, the higher the smoke SNR, as seen in subfigures 5(c)-5(e). In order to balance the need for a representative, information dense dataset that remains a manageable size to host online, we use the Mie scattering physics principles to narrow down our dataset. (1) we choose imagery from GOES-East or GOES-West that would be able to observe the highest smoke related forward scattering. (2) we choose the three ABI bands with the smallest available λ, 0.47μm, 0.64μm and 0.865 μm (C01-C03).

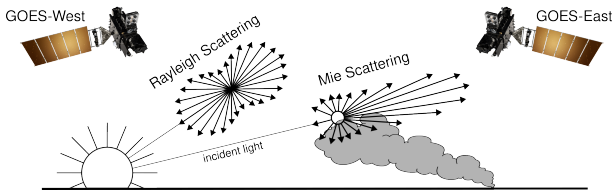

Figure 3: If the particle size is $< \frac{1}{10}$ the λ of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than the λ of light. This schematic demonstrates that when the sun is setting in the West, the Mie scattering will predominately forward scatter towards GOES-East.

For the existing dataset of HMS smoke plume annotations and all corresponding satellite imagery, $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, for each sample, a label $y_i \in \mathcal{Y}$ has a set of satellite imagery $[x_{(i,t_0)}, ..., x_{(i,t_N)}] \in \mathcal{X}$ over the analyst defined time window t. To develop a one-to-one data-to-label dataset, we generate IPLs to develop a subset of $\mathcal{D}$, denoted as $\mathcal{D}_p$, that has a one-to-one ratio such that $|\mathcal{X}_p| = |\mathcal{Y}|$, where we aim to find the satellite image that has the maximum overlap between the geolocation of smoke in the imagery and the analyst annotation using the PLDR method shown in figure 4.

To train an initial parent model, $f_\circ$, we developed a method of leveraging Mie scattering physics to select $x_{(i,t_M)} \in \mathcal{X}$ (figure 4(b)) so that $x_{(i,t_M)}$ has a higher chance than random selection out of the set of imagery $[x_{(i,t_0)}, ..., x_{(i,t_N)}]$ to be representative of y_i. This Mie-Derived Dataset, $\mathcal{D}_M$ produces a training set so that if there is smoke present during the entire time window, the selected timestamp t_M would give the highest smoke SNR.

More importantly than finding the timestamp for maximum the SNR, we want to determine which image actually has smoke present within the smoke label boundaries. We use $\mathcal{D}_M$ to train $f_\circ$ (figure 4(d)), to identify smoke in satellite imagery, and then use that $f_\circ$ to create IPLs of each satellite image in a given annotation's time-window (figure 4(e)). From those results, the optimal satellite image is chosen based on which image's IPLs has the greatest overlap with the analyst annotation (figure 4(f)-(h)).

3.1.1 Mie-Derived Dataset $\mathcal{D}_M$

We apply a physics-based approach to select the Mie-derived dataset, $\mathcal{D}_M$, from the full dataset $\mathcal{D}$, so that $\mathcal{D}_M \subset \mathcal{D}$ and so that $|\mathcal{X}_M| = |\mathcal{Y}|$ so that we can efficiently and effectively train f_o. Based on the criteria to optimize for maximum observation of Mie forward scattering, the trivial strategy would be to pull imagery from GOES-West right after sunrise and from GOES-East right before sunset when the SZA is 90°. However, the time when the SZA approaches 90° coincides with when there are maximized atmospheric interactions that cause an increase in noise and artifacts [29]. With the goal to reduce large SZA related noise, when there are multiple frames to choose from, we choose the image with the largest SZA that is $< 88^\circ$.

The Mie-derived image selection process accounts for atmospheric properties and light scattering physics to calculate which singular satellite image within the analyst time-window could give the highest possible smoke SNR if smoke was equally present throughout the time-window. The resulting Mie-derived dataset, $\mathcal{D}_M = \{\mathcal{X}_M, \mathcal{Y}\}$, contains over 200,000 samples and was then used to train a parent model, $f_\circ$.

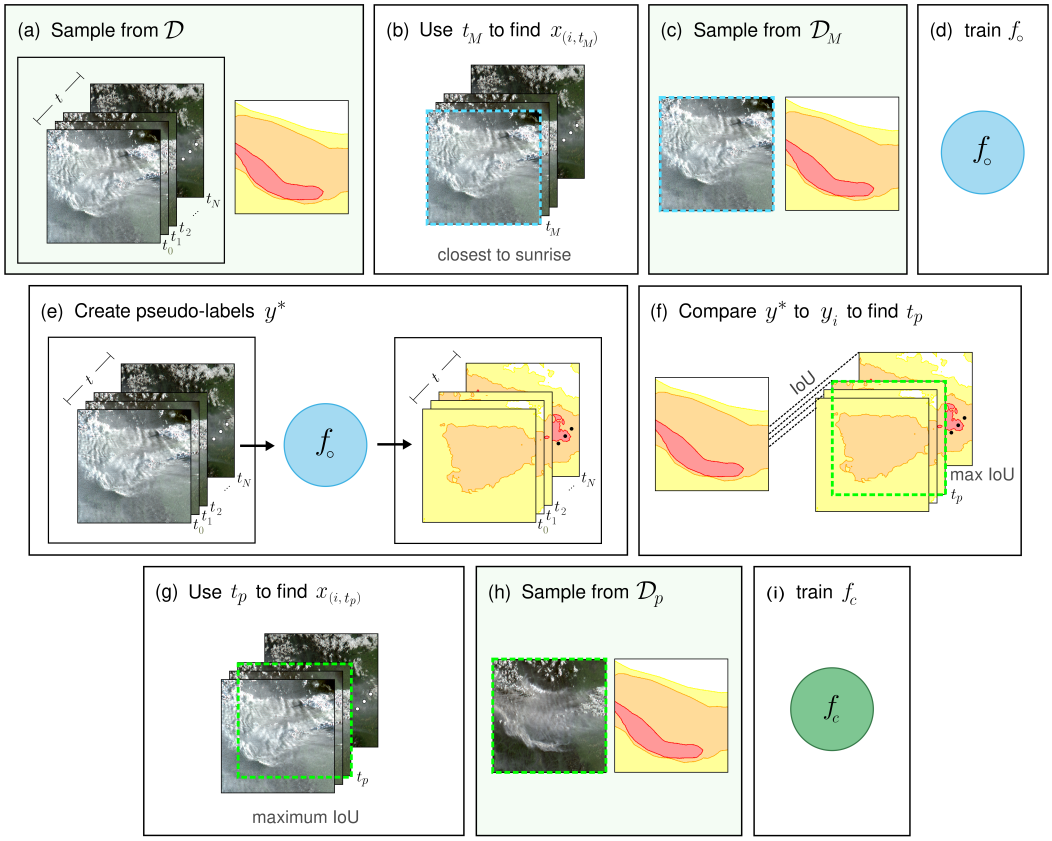

Figure 4: PLDR applied to create the SmokeViz dataset, the green boxes represent the dataset stages, note that the HMS analyst label y remains the same, the only change is for the satellite image(s) x. (a) for a sample in the original full dataset $\mathcal{D}$, each analyst annoation y_i corresponds to N satellite images that cover time window t so that $([x_{(i,t_0)}, ..., x_{(i,t_N)}], y_i) \in \mathcal{D}$ (b) using Mie scattering physics we find the satellite image, $x_{(i,t_M)}$ that correlates with the time, t_M, that would produce the highest possible SNR if smoke was present (c) the resulting $\mathcal{D}_M$ has a one-to-one ratio of image-to-annotation $(x_{(i,t_M)}, y_i) \in \mathcal{D}_M$ (d) use $\mathcal{D}_M$ to train the parent model, $f_\circ(x_{(i,t_M)}) = y_i$ (e) apply a greedy algorithm $f_\circ([x_{(i,t_0)}, ..., x_{(i,t_N)}]) = [y^*_{(i,t_0)}, ..., y^*_{(i,t_N)}]$ to create IPLs y^* for each candidate image (f) use the intersection over union (IoU) metric to compare y^* to the analyst annotation, y_i, and identify the time, t_p, where the IPL and analyst annotation have the maximum IoU (g) use the time of highest overlap between y^* and y_i to identify the sample, $x_{(i,t_p)}$, that should best match the analyst annotation (h) $\mathcal{D}_p$ has a one-to-one ratio of image-to-annotation $(x_{(i,t_p)}, y_i) \in \mathcal{D}_p$ (i) use the SmokeViz dataset $\mathcal{D}_p$ to train deep learning models $f_c(x_{(i,t_p)}) = y_i$ to detect and classify the density of wildfire smoke plumes in GOES imagery.

3.1.2 PLDR Dataset $\mathcal{D}_p$

To build the deep learning architecture for f_o we use the high level API package Segmenation Models Pytoch [15] to implement EfficientNetV2 [34] as the encoder with random initialized weights connected to the semantic segmentation classifier from the PSPNet model [36]. The data input is 256x256x3 snapshots of 1km resolution true color GOES imagery that contain smoke and output a 256x256x3 classification map that predicts if a pixel contains smoke and if so, what the categorical density of that smoke is. Since the smoke density labels are ordinal, we apply the thermometer encoding shown in table 2 to encode the smoke densities and apply binary cross entropy as the loss function per density of smoke.

To determine which image out of the relevant imagery for the given time window best represents the analyst annotation, we implement a greedy algorithm (figure 4(e)) $f_\circ([x_{(i,t_0)}, ..., x_{(i,t_N)}]) = [y^*_{(i,t_0)}, ..., y^*_{(i,t_N)}]$. The outputs of $f_\circ$, $[y^*_{(i,t_0)}, ..., y^*_{(i,t_N)}]$ give N predictions on the location and

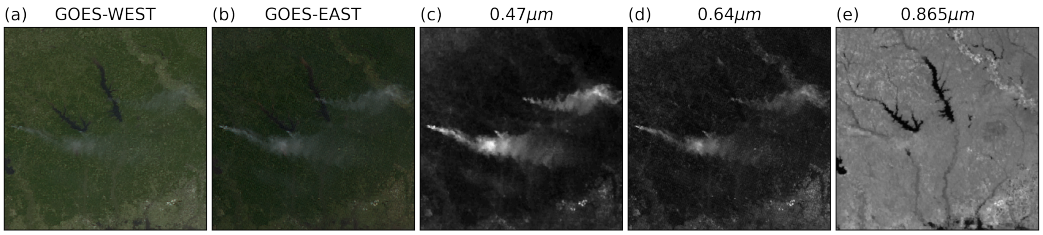

Figure 5: True color (a) GOES-WEST and (b) GOES-EAST imagery from March 23^{rd}, 2022 centered at $(31.1°, -93.8°)$ in Texas, USA taken at 23:20 UTC. The GOES-EAST raw band imagery for (c) blue, (d) red and (e) veggie bands show variations in the SNR for smoke detection in relation to the λ of light being measured.

Table 2: A comparison of how smoke density would be represented by one-hot encoding commonly used for categorical data to thermometer encoding often used for ordinal data.

density	one-hot	thermometer
none	[0 0 0]	[0 0 0]
light	[0 0 1]	[0 0 1]
medium	[0 1 0]	[0 1 1]
heavy	[1 0 0]	[1 1 1]

Table 3: Dataset split for $\mathcal{D}_M$ and $\mathcal{D}_p$, samples for 2024 go up to November 1st do 2018-2022 instead. We use an entire year of data for both validation and testing sets to capture year-long wildfire trends.

dataset	$\mathcal{D}_M$	$\mathcal{D}_p$	years
training	165,609	144,225	2018-22
validation	20,056	19,223	2023
testing	21,541	20,224	2022

desnity of smoke in the imagery over the analyst defined timewindow t. Each y^* serve as a semantic segmenation IPL of smoke and are compared to the corresponding analyst annotation, y_i. To compare each y^* to y_i (figure refPLDR(f)), we calculate the intersection over union (IoU) using the total set of pixels for y^* at that density of smoke and the entire set of pixels for y_i for a particular smoke density in each image as shown in equation 1. The timestamp, t_p associated with the IPL with the highest IoU score, $y^*_{(i,t_p)}$ is matched to the corresponding image, $x_{(i,t_p)}$ (figure 4(g)), that is predicted to best represent the analyst smoke annotation, y_i. A stardard implementation detail for PLs [12], a confidence threshold value is defined to determine if the corresponding sample $(x_{(i,t_p)}, y_i)$ should to be included in the dataset $\mathcal{D}_p$. Based on subjective visual inspection by the authors of when the satellite imagery became consistently unidentifiable to its label, we chose a confidence threshold that would include the sample in $\mathcal{D}_p$ if the maximum overall IoU was over 0.01. say size

$$IoU_{\text{overall}} = \frac{\sum_{j=\text{light}}^{\text{heavy}} |y_j \cap y_j^*|}{\sum_{j=\text{light}}^{\text{heavy}} |y_j| \cup |y_j^*|} \tag{1}$$

The same dataset split choices and model training setup that were used for $\mathcal{D}_M$ and $f_\circ$ were implemented for $\mathcal{D}_p$ and f_c. As depicted in figure 4(e), we use $\mathcal{D}_p$ to train a child model f_c. To assess if training with $\mathcal{D}_p$ can produce a more robust semantic segmentation model compared to training on $\mathcal{D}_M$ we run f_c on the $\mathcal{D}_p$ and $\mathcal{D}_M$ test sets.

3.2 Benchmark Models

We benchmark the SmokeViz dataset $\mathcal{D}_p$ by varying the semantic segmentation classification heads. We train Linknet [8], PSPNet [36] and MANet [11] using the same encoder for f_c and $f_\circ$, EfficientNetV2. Each model is trained over 100 epochs using a batch size of 32 and the Adam optimizer on 8 Nvidia P100 GPUs allocating 100GB of memory over 12 hours of allotted training time. We choose these architectures because of their abilities to capture multi-scale objects such the varying spatial extents of smoke plumes.

Table 4: IoU results per density of smoke and over all densities using $f_\circ$ and f_c with $\mathcal{D}_M$ and $\mathcal{D}_p$.

	$f_\circ$		f_c	
	$\mathcal{D}_M$	$\mathcal{D}_p$	$\mathcal{D}_M$	$\mathcal{D}_p$
Heavy	0.278	0.368	0.218	0.411
Medium	0.310	0.417	0.319	0.484
Light	0.480	0.585	0.491	0.660
Overall	0.430	0.533	0.438	0.607

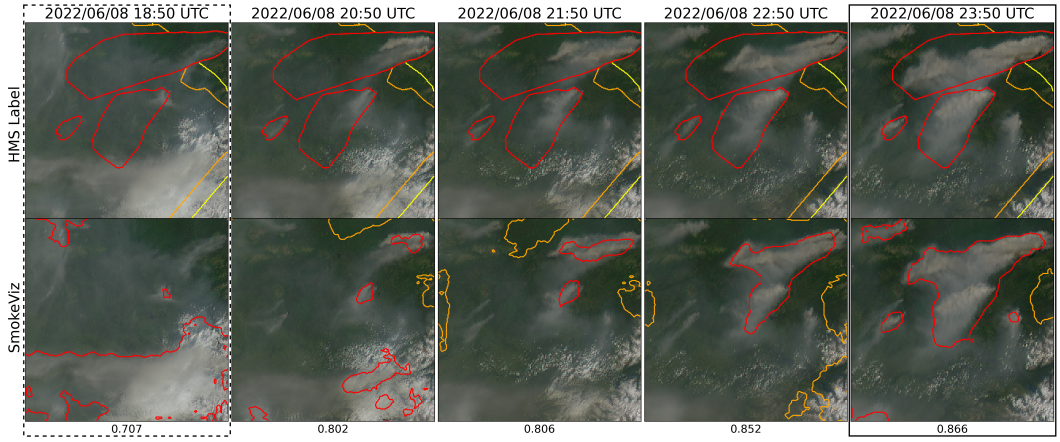

Figure 6: GOES-West imagery showing smoke on June 8th, 2022 in Alaska where, at this geolocation (61.06°N, 156.12°W), daylight was between 12:43-7:53 UTC. The HMS smoke annotations (top row) span from 18:50 to 23:50 UTC and are compared to the $f_\circ$ generated pseudo-labels (bottom row). The first column (dotted outline) would be the GOES imagery selected for $\mathcal{D}_M$ since it is closest to sunrise. The last column (solid outline) was selected for $\mathcal{D}_p$ since it had the highest IoU value between the pseudo-label and analyst annotation. The IoU score over all densities is reported at the bottom of each column.

4 Results

To interpret the performance of $f_\circ$, we report the IoU metrics in table 4 that were computed by running $f_\circ$ and f_c on $\mathcal{D}_M$ and $\mathcal{D}_p$. For each density, we calculate the IoU using the total set of pixels that $f_\circ$ predicts as that density of smoke and the entire set of pixels labeled by the analyst as a particular smoke density over all imagery contained in the testing dataset. Additionally, we compute the overall IoU for all densities by first computing the number of pixels that intersect their corresponding density and divide that by the total number of pixels that make up the union of model predicted and analyst labeled smoke in the testing dataset.

An illustration of a pseudo-label picked image better representing the analyst annotation when compared to the Mie-derived image selection is evident in Figure 6, where the heavy density smoke IoU increases from 0.01 to 0.59. The analyst annotation for these densities cover 5 hours of imagery, the Mie-derived selection optimizes for the image closest to sunrise while the pseudo-label image selection chooses the image with the highest overlap between the pseudo-label and the analyst annotation. The figure also illustrates how using a deep learning model can provide higher time resolution and give a dynamic representation of smoke over time.

To get an idea on how f_c compares to the HMS analyst annotations we show a series of samples from $\mathcal{D}_p$ in figure 5. The examples give a qualitative representation of how the predictions from f_c can provide more detailed boundaries of smoke densities than the HMS annotations do.

The results for the benchmarking models (table 5) show similar performance across the models. DeepLabV3+ (f_c) gives the highest heavy density smoke IoU value, while PSPNet gives the highest overall IoU score.

Table 5: Comparison of semantic segmentation model IoU performance on $\mathcal{D}_p$.

	DLV3+	MANet	PSPNet	Linknet
heavy	0.411	0.336	0.355	0.324
medium	0.484	0.487	0.502	0.456
light	0.662	0.675	0.690	0.662
overall	0.607	0.615	0.626	0.601

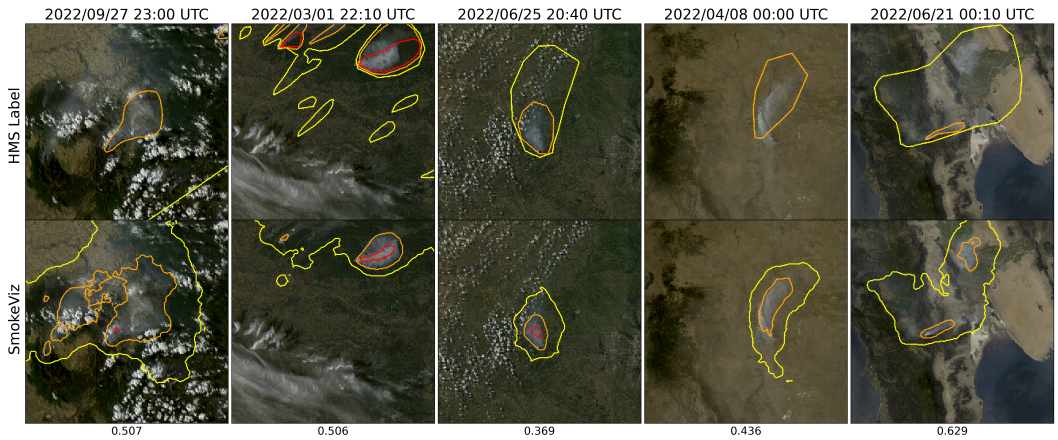

Figure 7: Examples of HMS annotations (top row) vs f_c output (bottom row) on $\mathcal{D}_p$ samples. The overall IoU score is reported at the bottom of each column.

5 Limitations

One of the concerns that comes with using pseudo-labeling methods is that you can perpetuate biases from the parent model into subsequent child models. Due to the increase in detectable forward scattered light off smoke particular matter, we expect the model to have a bias towards producing a higher success rate for smoke detection at larger solar zenith angles. The original HMS annotations do not distinguish by type of fire and include a large representation of controlled agricultural burns. This can be a limitation to consider if the dataset is being trained to target detection of large wildfires. All these limitations are discussed and analyzed further in the Supplementary Materials. Additional work should be done to analyze the performance of SmokeViz derived models on dust vs smoke.

6 Conclusion

In this study, we have refined an existing dataset originally curated by NOAA's HMS team, transforming it from a many-to-one imagery-to-annotation format to a more succinct, one-to-one satellite image-to-annotation dataset. The initial HMS dataset provided a general approximation of where smoke had been present for a given time window, though it did not guarantee the actual existence of smoke in the labeled pixels during the given times. Our goal was to create a dataset that could be used, along with additional applications, to train a model to detect wildfire smoke in real-time on an image-by-image level. The Mie-derived dataset selection process determined that if smoke was present, what timestamp within the analyst time window would the give the highest smoke signal-to-noise ratio. While optimizing for being able to detect smoke, if it is present, the Mie-dataset selection had no metric to determine if the smoke was effectually present in the selected image. Since many of the images within the HMS time-window either contained no smoke at all or the smoke was not contained within the geospatial bounds of the annotations, the Mie-derived dataset contained a large number of mislabeled samples. Discrepancies between data and labels can be detrimental towards the model's capacity to improve on feature representations in the target domain. During model training, the penalization of accurate predictions can inadvertently introduce biases towards misclassifying noise as meaningful signal.

To improve the dataset's capacity to accurately represent wildfire smoke plumes, we train a parent machine learning model, $f_\circ$, using the Mie-derived dataset, $\mathcal{D}_M$, and run it on the relevant satellite images within the time-frame. The image with the maximum IoU score between the model's smoke predictions, or pseudo-label, and the analyst smoke annotations are used to create the pseudo-label generated dataset, $\mathcal{D}_p$. We then train a child model, f_c, using $\mathcal{D}_p$ and test $f_\circ$ and f_c on both the 2022 testing sets from $\mathcal{D}_M$ and $\mathcal{D}_p$. The results reported in table 4 suggest that $\mathcal{D}_p$ was able to train a better performing model, f_c, that gave higher IoU metrics on both dataset's testing sets in comparison to the original parent model, $f_\circ$.

The result of this study is a representative dataset, SmokeViz, that can be used to train machine learning models for various wildfire smoke applications. A future goal is to produce a robust and reliable machine learning based approach for detecting wildfires using satellite imagery. That information can be used for wildfire detection and monitoring in along with a highly needed smoke product for data assimilation into smoke dispersion models. Additionally, this dataset can be used as a benchmark for how well remote sensing segmentation models can perform on dispersed edges such as smoke. On a broader scale, we show how pseudo-labeling can be used to optimize a dataset when the resolution for the data and corresponding labels do not match. This could be useful in similar applications involving time-series/video data with a singular label where the data can be compressed while still remaining representative of the label. All data is made publically available at [aws download link] and all code can be found at `https://github.com/anonymous-smokeviz/SmokeViz`.

References

[1] R. Ahmadov, H. Li, J. Romero-Alvarez, J. Schnell, S. Bhimireddy, E. James, K. Y. Wong, M. Hu, J. Carley, P. Bhattacharjee, et al. Forecasting smoke and dust in noaa's next-generation high-resolution coupled numerical weather prediction model. Technical report, Copernicus Meetings, 2024.

[2] J. E. Aldy, M. Auffhammer, M. Cropper, A. Fraas, and R. Morgenstern. Looking back at 50 years of the clean air act. *Journal of Economic Literature*, 60(1):179–232, 2022.

[3] R. S. Allison, J. M. Johnston, G. Craig, and S. Jennings. Airborne optical and thermal remote sensing for wildfire detection and monitoring. *Sensors*, 16(8):1310, 2016.

[4] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo. Smokenet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 11(14): 1702, 2019.

[5] M. Bah, M. Gunshor, and T. Schmit. Generation of goes-16 true color imagery without a green band. *Earth and Space Science*, 5(9):549–558, 2018.

[6] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023.

[7] M. Burke, A. Driscoll, S. Heft-Neal, J. Xue, J. Burney, and M. Wara. The changing risk and burden of wildfire in the united states. *Proceedings of the National Academy of Sciences*, 118 (2):e2011048118, 2021.

[8] A. Chaurasia and E. Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] L. Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[11] T. Fan, G. Wang, Y. Li, and H. Wang. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8:179656–179665, 2020.

[12] R. E. Ferreira, Y. J. Lee, and J. R. Dórea. Using pseudo-labeling to improve performance of deep neural networks for animal identification. *Scientific Reports*, 13(1):13875, 2023.

[13] E. Gakidou, A. Afshin, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, A. M. Abdulle, S. F. Abera, V. Aboyans, et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1345–1422, 2017.

[14] S. J. Goodman, T. J. Schmit, J. Daniels, and R. J. Redmon. *The GOES-R series: a new generation of geostationary environmental satellites*. Elsevier, 2019.

[15] P. Iakubovskii. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019.

[16] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, et al. Foundation models for generalist geospatial artificial intelligence. arxiv 2023. *arXiv preprint arXiv:2310.18660*.

[17] E. James, R. Ahmadov, and G. A. Grell. Realtime wildfire smoke prediction in the united states: The hrrr-smoke model. In *EGU General Assembly Conference Abstracts*, page 19526, 2018.

[18] A. Kitamoto, J. Hwang, B. Vuillod, L. Gautier, Y. Tian, and T. Clanuwat. Digital typhoon: Long-term satellite image dataset for the spatio-temporal modeling of tropical cyclones. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] A. Larsen, I. Hanigan, B. J. Reich, Y. Qin, M. Cope, G. Morgan, and A. G. Rappold. A deep learning approach to identify smoke plumes in satellite imagery in near-real time for health risk communication. *Journal of exposure science & environmental epidemiology*, 31(1):170–176, 2021.

[21] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[22] Y. Lee, C. D. Kummerow, and I. Ebert-Uphoff. Applying machine learning methods to detect convection using geostationary operational environmental satellite-16 (goes-16) advanced baseline imager (abi) data. *Atmospheric Measurement Techniques*, 14(4):2699–2716, 2021.

[23] D. McNamara, G. Stephens, M. Ruminski, and T. Kasheta. The hazard mapping system (hms) - noaa's multi-sensor fire and smoke detection program using environmental satellites. *Conference on Satellite Meteorology and Oceanography*, 01 2004.

[24] J. Y.-H. Ng, K. McCloskey, J. Cui, V. R. Meijer, E. Brand, A. Sarna, N. Goyal, C. Van Arsdale, and S. Geraedts. Contrail detection on goes-16 abi with the opencontrails dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2023.

[25] NOAA. Hazard mapping system fire and smoke product. URL `https://www.ospo.noaa.gov/Products/land/hms.html#about`.

[26] T. C. Phan and T. T. Nguyen. Remote sensing meets deep learning: exploiting spatio-temporal-spectral satellite images for early wildfire detection. 2019.

[27] T. Randriambelo, S. Baldy, M. Bessafi, M. Petit, and M. Despinoy. An improved detection and characterization of active fires and smoke plumes in south-eastern africa and madagascar. *International Journal of Remote Sensing*, 19(14):2623–2638, 1998.

[28] M. Raspaud, D. Hoese, A. Dybbroe, P. Lahtinen, A. Devasthale, M. Itkin, U. Hamann, L. Ø. Rasmussen, E. S. Nielsen, T. Leppelt, et al. Pytroll: An open-source, community-driven python framework to process earth observation satellite data. *Bulletin of the American Meteorological Society*, 99(7):1329–1336, 2018.

[29] A. Royer, P. Vincent, and F. Bonn. Evaluation and correction of viewing angle effects on satellite measurements of bidirectional reflectance. *Photogrammetric engineering and remote sensing*, 51(12):1899–1914, 1985.

[30] W. Schroeder, M. Ruminski, I. Csiszar, L. Giglio, E. Prins, C. Schmidt, and J. Morisette. Validation analyses of an operational fire monitoring product: The hazard mapping system. *International Journal of Remote Sensing*, 29(20):6059–6066, 2008.

[31] B. Stevens, S. Bony, H. Brogniez, L. Hentgen, C. Hohenegger, C. Kiemle, T. S. L'Ecuyer, A. K. Naumann, H. Schulz, P. A. Siebesma, et al. Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds. *Quarterly Journal of the Royal Meteorological Society*, 146(726): 141–152, 2020.

[32] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pages 5901–5904. IEEE, 2019.

[33] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[34] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[35] J. Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery. *ArXiv*, abs/2109.01637, 2021. URL `https://api.semanticscholar.org/CorpusID:237416777`.

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[37] T. X.-P. Zhao, S. Ackerman, and W. Guo. Dust and smoke detection for multi-channel imagers. *Remote Sensing*, 2(10):2347–2368, 2010. ISSN 2072-4292. doi: 10.3390/rs2102347. URL `https://www.mdpi.com/2072-4292/2/10/2347`.

[38] S. Zorzi, S. Bazrafkan, S. Habenschuss, and F. Fraundorfer. Polyworld: Polygonal building extraction with graph neural networks in satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1848–1857, 2022.