LaTeX Author Guidelines for CVPR Proceedings

Anonymous CVPR submission

Paper ID *****

Abstract

The global increase in the frequency and intensity of wildfires 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 human-generated 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. We present an innovative solution for refining the temporal and spatial resolution in the given analyst annotations by leveraging the semi-supervised method, pseudo-labeling. Unlike typical pseudo-labeling applications that aim to increase the number of labeled samples, the objective is to use pseudo-labels to refine an existing but course-grained set of annotations. We train a deep learning model to generate pseudo-labels that pinpoint the singular, most representative, satellite image to match the smoke annotation within the given temporal range. By identifying the most representative imagery of smoke plumes for a given smoke annotation, the study seeks to create an accurate and relevant machine learning dataset. The resulting SmokeViz dataset is anticipated to be an instrumental tool in developing further machine learning models and is publically 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[1]. Despite those improvements, the contribution of wildfire smoke to $PM_{2.5}$ concentrations in the US has been calculated to have more than doubled between 2010 to 2020, accounting for up to half of the overall $PM_{2.5}$ exposure in Western regions [2]. 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 [3]. These risks underscore the necessity for efficient and effective monitoring methods to mitigate the adverse health impacts associated with wildfire smoke.

Traditionally, wildfire monitoring has relied on ground-based methods, such as forest service patrols, manned lookout towers, and aviation surveillance [4]. While these methods provide valuable localized insights, they are constrained by geographical and logistical limitations, often failing to deliver timely and comprehensive data, especially over large and remote areas. In contrast, satellite imagery offers a vantage point that overcomes these limitations, providing continuous, wide-area coverage and real-time data 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) [5], have revolutionized environmental monitoring. Compared to orbiting satellites such as the Suomi or Sentinel satellites, geostationary satellites maintain constant observation over a fixed area. GOES offers the advantage being able to reliably and consistently 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, models for real-time smoke dispersion currently have no smoke analysis product available for data assimilation [6, 7]. This can cause 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 [8]. 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 [9].

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 [10, 11]. 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. This method is potentially not as scalable as an automated approach and is limited by the availability of analysts and their time.

NOAA manages environmental satellite programs such as the HMS program, the HMS program is an operational system that uses an aggregation of satellite data to generate active fire and smoke data. To train our model, we implement 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 [12].

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 [13]. 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 [14]. 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 [15].

The success of deep learning methods, such as CNNs, relies heavily on the availability of a large, representative dataset [16]. As laid out in table 1, prior studies use relatively small numbers of samples, from 47 [17] to 6825 [15], where one sample represents a satellite image with a singular time and geolocation. In contrast, benchmark datasets for image classification contain tens of thousands (CIFAR-10 and MNIST) to millions (CIFAR-100 and ImageNet) of data samples [18–20]. Keeping in mind the correlation between both the quality and quantity of data with model performance, we introduce the largest known smoke dataset, SmokeViz, containing over 130,000 samples.

Semi-supervised learning is an approach that can be used to increase the number of labeled samples in a dataset. This is done by leveraging a labeled dataset to generate new labels for an often larger, but unlabeled, dataset. Pseudo-labeling, a form of semi-supervised learning, uses labeled data to train an initial model, then runs that model on unlabeled data to predict pseudo-labels, and finally trains a new model using the pseudo-labels [21]. Since we do not know of any studies that have used this technique in this way, we introduce a variation of pseudo-labeling, not to increase the size, but to increase the quality of our dataset by generating pseudo-labels to select the best satellite image out of a given time-window to represent each smoke plume annotation.

3. Methods

3.1. Datasets

In order 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 these annotations is to show the

Table 1. Comparison of different studies including method used, dataset size, satellite source, number of channels used and if classification is performed at a pixel or image level.

Reference	Method	# Samples	Satellite	# Channels	Level
[13]	CNN	6255	MODIS	5	image
[15]	CNN	6825	GOES-East	5	pixel
[14]	CNN	975	Himiwari-8	7	pixel
[17]	U-Net	47	Landsat-8	13	pixel
SmokeViz	U-Net	207,106	GOES-East/West	3	pixel

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 way to approach 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 require a large amount of additional memory and computational resources. Instead of using the original analysts' many satellite image inputs to one annotated output, we develop a one-to-one input-to-output by finding the optimal singular satellite image input to represent the annotation.

For the set of smoke annotations, $\mathcal{Y}$, $y \in \mathcal{Y}$ uses one or more $x \in \mathcal{X}$ where $\mathcal{X}$ is the entire set of satellite imagery corresponding to the set of time windows defined by the labels. In order to develop a one-to-one data-to-label dataset, we apply pseudo-labeling to develop a subset of $\mathcal{X}$, denoted as $\mathcal{X}_p$, that has a one-to-one ratio such that $|\mathcal{X}_p| = |\mathcal{Y}|$, where we choose the satellite image that has the maximum overlap between the geolocation of smoke in the imagery and the analyst annotation.

But in order to create pseudo-labels we need an initial parent model, $f_\circ$. To train $f_\circ$, we need a way of choosing $x \in \mathcal{X}$ that has a higher chance than random selection of being representative of y. Discussed in further detail in the Mie-Derived Dataset subsection, we do this by making a series of physics-driven choices on which satellite and timestamp would give the optimal angle between the sun, smoke and satellite to produce the strongest smoke signature for the geolocation and timestamp of the smoke plume. This dataset, $\mathcal{X}_M$ tells us that if there is smoke present during the entire time window, which timestamp would give the highest smoke signal-to-noise ratio.

But more importantly than knowing the timestamp for maximum signal-to-noise, we want to know which image actually has smoke present within the smoke label boundaries. We used $\mathcal{X}_M$ to train $f_\circ$, to identify smoke in satellite imagery, and then use that $f_\circ$ to create pseudo-labels of each satellite image in a given annotation's time-window. From those results, the optimal satellite image is chosen based on which image's pseudo-labels has the greatest overlap with the analyst annotation.

Table 2. To create a true color image, we use the following bands from the ABI Level 1b CONUS (ABI-L1b-RadC) product.

band	description	center $\lambda(\mu\mathrm{m})$	resolution (km)
C01	blue visible	0.47	1
C02	red visible	0.64	0.5
C03	veggie NIR	0.865	1

3.1.1. Satellite Imagery

The GOES satellites are operated by NOAA in order to support meteorology research and forecasting for the United States. We use the latest operational satellites, GOES-16 (East), 17 and 18 (West) that each carry the ABI, that measure 16 bands between the visible and infrared wavelengths. In improvement to the GOES predecessors, imagery is collected every 5 minutes for the contiguous United States and every 10 minutes for the full disk. Using PyTroll, a Python framework for processing satellite data [22], we input bands 1-3 (Table 2) to a GOES specific true color composite algorithm [23] to develop a, 1km resolution, true color image representation, similar to what is seen by HMS analysts. As discussed in further detail in the next section, the highest signal-to-noise ratio will come from the smallest wavelengths of light, larger wavelengths have lower smoke signal and higher noise (figure 5). For that reason, we only include the first 3 out of 16 available bands of data.

3.1.2. Mie-Derived Dataset

We used a physics-informed approach in selecting the initial GOES dataset, $\mathcal{X}_M$, which we call the Mie-derived dataset, for training an initial parent model, $f_\circ$, where if $\mathcal{X}$ represents all the GOES imagery corresponding to the HMS smoke annotation time window, $\mathcal{X}_M \subset \mathcal{X}$. Prior GOES ABI datasets for machine learning applications often include data from only one of the two GOES-series satellites, commonly opting for GOES-East [15], [24], [25]. Rather than using one satellite or the cumulative data from both GOES-West and GOES-East images, we select between one or the other based on the solar zenith angle. For smoke iden-

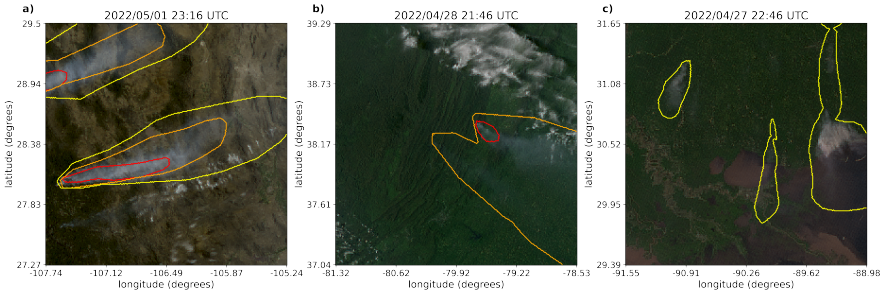

Figure 1. Satellite imagery captured by GOES-East within a few days of each other. The yellow, orange and red contours indicate the extent of light, medium and heavy density smoke. a) shows a canonical example of a smoke plume. b) and c) show observable variations in the density labels.

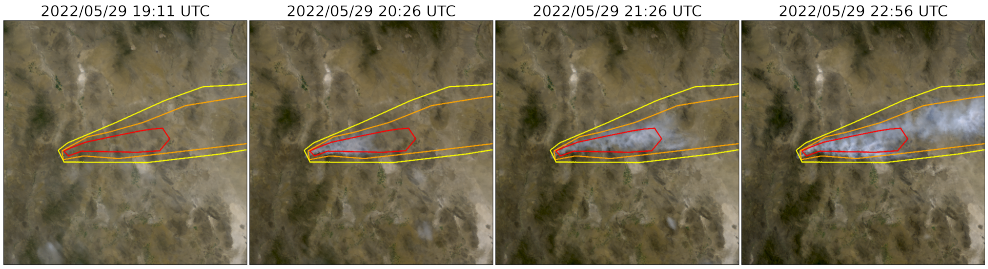

Figure 2. True color GOES-East imagery from May 2022, Southeast New Mexico (31°N, 100°W) during the start of the Foster Fire. The red, orange and yellow lines represent the heavy, medium and low density HMS smoke annotations that span 19:10–23:00 UTC.

tification, this approach can achieve a much higher signal-to-noise than imaging the earth's surface from an arbitrary angle. The elastic scattering of light is the primary mechanism to account for - while the atmosphere is composed of molecules with size <1nm, smoke particles can vary from 100 nm – 10 μm in diameter, d. The GOES ABI covers spectral bands from 0.47 μm – 13.3 μm, so atmospheric and smoke particle sizes occupy two very different regimes with respect to the imaging wavelength λ. In the extreme limit of $\lambda \gg d$, the physics of scattering of light off a small sphere is captured by Rayleigh scattering. This process has two critical consequences: (1) the scattering cross section of light is strongly wavelength dependent (scaling with λ^{-4}), meaning that photons with wavelength closer to the ultraviolet are scattered more strongly than infrared photons. (2) the scattering cross section scales with an angular dependent cross section of $(1 + \cos^2\theta)$. Scattered photons follow the emission distribution of a radiating dipole, scattering more strongly in the forward and backwards directions $(\theta = 0, \pi)$than orthogonal to the direction of propagation $(\theta = \pi/2, 3\pi/2)$, see figure 3 for a Rayleigh scattering schematic.

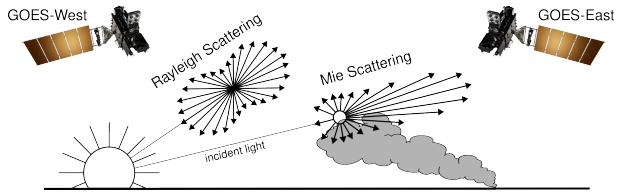

Figure 3. If the particle size is $< \frac{1}{10}$ the wavelength of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than the wavelength of light. This schematic demonstrates that when the sun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

The significance of these scalings is that the observer, or detector, will receive blue photons in most directions orthogonal to the source. Equivalently, photons traveling colinearly with line of sight to the emission source will mostly have wavelengths in the infrared band. In the converse regime of $d > \lambda$, the elastic scattering of light against matter is modeled through Mie scattering. In comparison to Rayleigh scattering, Mie scattering is largely wavelength-independent and has a more complicated radiation pattern

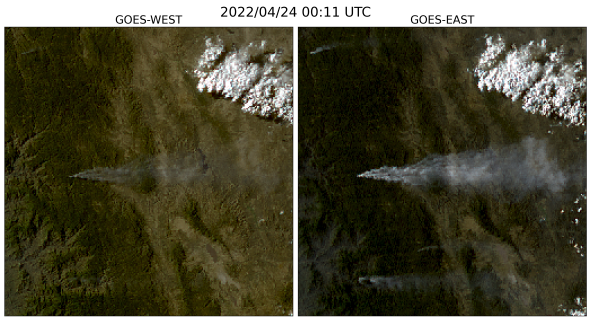

Figure 4. True color GOES-West (left) and GOES-East (right) imagery from April 24^{th}, 2022 in Durango, Mexico. The images were taken ~ 1.5 hours before sunset (01:43 UTC) for this geolocation and time of year.

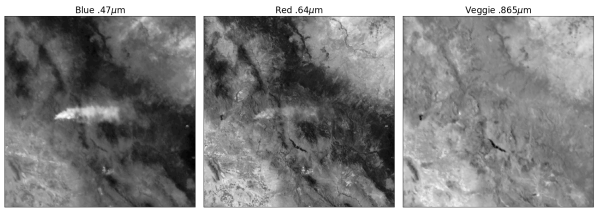

Figure 5. Three bands of GOES-East data are the raw input to generate a true color image. These plots show variations in the signal-to-noise ratio for smoke detection in relation to the λ of light being measured.

where the cross section has a maximal amplitude in the forward direction. An observer downstream of this scatterer will collect more photons than one positioned directly behind it. In the context of smoke identification, a sunrise or sunset will lead to a higher Mie scattered signal in GOES-West and GOES-East respectively, as shown with a smoke plume producing a stronger signal in GOES-East imagery near sunset in figure 4.

Smoke identification therefore amounts to extracting a signal of $d > \lambda$ photons from the $\lambda \gg d$ background. Positioning a detector along line of sight to the scatterer will result in a higher signal from smoke particles (figure 3). Filtering the imaged wavelength can enhance this signal; photons collected in the blue spectrum will have a naturally lower background along the line of sight to the illumination source do their high level of Rayleigh scattering as. Therefore, as demonstrated in figure 5, this configuration results in the highest signal to noise imaging for smoke particles.

Based solely on these criteria, the optimal strategy would be to pull data from GOES-West right after sunrise and from GOES-East right before sunset. Another factor to consider is that the time when the sun is in optimal alignment with the satellite for smoke detection coincides with when solar zenith angle is close to $90°$. Larger angles between the satellite and sun result in an increase in noise due to increased atmospheric interactions [26]. To reduce the noise from large solar zenith angles, if given multiple frames to choose from, we choose the image with the largest solar zenith angle that is $< 88°$.

The resulting image selection process takes into account atmospheric properties and light scattering physics to generate an estimate of which singular satellite image within the analyst time-window could give the highest smoke signal-to-noise ratio. The resulting Mie-derived dataset, $\mathcal{X}_M = \{X_M, Y\}$, was then used to train a model, $f_\circ$, that would generate N pseudo-labels, y^*, for every sample, where N is determined by how many images, taken at a 10 minute interval, fit within the analyst time-window for that sample. Chosen from the N images, x_p is the image with the highest alignment between the $f_\circ$ prediction of smoke, y^*, in the image and the HMS analysts' annotation y.

3.1.3. Thermometer Encoding Smoke Densities

One of the challenges introduced with using human generated qualitative smoke densities was that, as seen in figure 1b and 1c, there are variations in what is labeled as heavy or light density smoke. More generally, reproducing qualitative metrics with quantitative algorithms is a challenging problem, but we apply mathematical approaches that mitigate some of the underlying complications of our specific problem. Despite the smoke densities introduce qualitative complexities, we decided that the density approximations were important to use in our dataset because of the differences in signatures the densities produce. Within the satellite imagery, the appearance of a light density smoke plume will look significantly different than a heavy density smoke plume as seen in figure 1. Additionally, a light density smoke plume is expected to be more challenging to detect since it is easier for it to be misclassified as not smoke. During the training process, the separate density categories allows us to deferentially weight the penalization given to the model for incorrect classifications based on category. For example, the model can be given a small penalization for misclassifying light smoke as not smoke while given a higher penalization for misclassifying heavy smoke as not smoke.

In addition to the densities being ordered and categorical, the differences between the density categories are not evenly distributed by a given metric, such as $PM_{2.5}$ density. The intervals between densities being unknown along with the hierarchical nature of the density labels makes the labels ordinal instead of just categorical. This data property allows us to use thermometer encoding [27], which leverages the idea that heavy density smoke includes both medium and light density smoke, that heavy density smoke is closer to medium than it is to light, and automatically weights the

loss functions and incorporates the ranked ordering of the densities. As seen in Table 3, one-hot encoding, commonly used for categorical data, doesn't take ordinal properties of the data into consideration.

Table 3. A comparison of one-hot encoding used for categorical data to thermometer encoding for ordinal data.

category	one-hot	thermometer
No Smoke	[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]

3.1.4. Pseudo-label Dataset

We implement a deep learning architecture that uses the encoder from EfficientNetV2 [28] and a semantic segmentation classifier from the DeepLabV3 model [29]. Transfer learning has shown to reduce the time and resources needed to train a model by leveraging information from pre-trained models [30], [31]. We initialize the values of our model weights using the pre-trained values originally trained on the ImageNet dataset [20], containing 1.2 million images and 1000 categories. Our model was developed using the Segmentation Models PyTorch package [32] that was written as a high level API for implementing models for semantic segmentation problems. We input 256x256x3 snapshots of 1km resolution true color GOES imagery that contains smoke and output a 256x256x3 classification map that predicts if a pixel contains smoke and if so, what the density of that smoke is. As mentioned earlier, we apply the thermometer encoding shown in table 3 to encode the smoke densities and apply binary cross entropy as the loss function per density of smoke.

The dataset, $\mathcal{X}_M$, contains 207,106 samples as shown in the dataset split in table 4.

Table 4. Dataset split for $\mathcal{X}_M$ and $\mathcal{X}_p$, samples for 2024 go up to November 1st. We use an entire year of data for both validation and testing sets to capture year-long wildfire trends.

dataset	$\mathcal{X}_M$	$\mathcal{X}_p$	years
training	165,609	144,225	2018-2021, 2024
validation	20,056	19,223	2023
testing	21,541	16,855	2022

To determine which image out of the relevant imagery for the given time window best represents the analyst annotation, we implement a greedy algorithm by running $f_\circ$ on each x to generate a pseudo-label, y^*. The output of $f_\circ$, y^* give predictions on if smoke is in the image, and if there is smoke, where the smoke is in that image and the density of that smoke. y^* serve as pseudo-labels for each density of smoke and are compared to the analyst annotations, y. To compare y^* and y, we calculate the IoU using the total set of pixels for y^* at that density of smoke and the entire set of pixels for y for a particular smoke density in each image as shown in equation 1. The image with the highest IoU score is chosen as the image, x_p, that best represents the analyst smoke annotation, y. Often used for pseudo-labeling, a confidence threshold value is defined to determine if a pseudo-label should to be included in a dataset [33]. We chose a confidence threshold that would include the sample, x_p, in $\mathcal{X}_p$ if the maximum overall IoU (equation 1) between y^* and y over all densities was over 0.01.

$$IoU_{\text{overall}} = \frac{\sum_{i=\text{light}}^{\text{heavy}} |y_i \cap y_i^*|}{\sum_{i=\text{light}}^{\text{heavy}} |y_i| \cup |y_i^*|} \quad (1)$$

We use $\mathcal{X}_p$ to train an additional child model, f_c in order to assess if training with $\mathcal{X}_p$ can produce a more robust semantic segmentation model compared to training on $\mathcal{X}_M$. We use the same dataset split method and model setup but change $\mathcal{X}_M$ to $\mathcal{X}_p$ to train f_c.

3.2. Benchmark Models

While this dataset is anticipated to be primarily useful for solving various wildfire smoke applications, this dataset could be a uniquely insightful test case for remote sensing semantic segmentation. Many deep learning satellite image datasets are focused on objects with sharp contrasts such as crops [34], human infrastructure [35], or even clouds over oceans [36, 37], but smoke has indistinct boundaries that often fade both spatially and temporally.

We benchmark the SmokeViz dataset, $\mathcal{X}_p$ by varying the semantic segmentation classification heads. We train Linknet [38], PSPNet [39] and MANet [40] using the same encoder used 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.

4. Results

To interpret the performance of $f_\circ$, we report the IoU metrics in table 5 that were computed by running $f_\circ$ and f_c on $\mathcal{X}_M$ and $\mathcal{X}_p$. 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

Table 5. IoU results per density of smoke and over all densities using $f_\circ$ and f_c with $\mathcal{X}_M$ and $\mathcal{X}_p$.

	$f_\circ$		f_c	
	$\mathcal{X}_M$	$\mathcal{X}_p$	$\mathcal{X}_M$	$\mathcal{X}_p$
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

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 $\mathcal{X}_p$ in figure 6. 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.

Table 6. Comparison of semantic segmentation model IoU performance on $\mathcal{X}_p$.

	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

The results for the benchmarking models (table 6) 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.

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 Appendix. 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, $\mathcal{X}_M$, 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, $\mathcal{X}_p$. We then train a child model, f_c, using $\mathcal{X}_p$ and test $f_\circ$ and f_c on both the 2022 testing sets from $\mathcal{X}_M$ and $\mathcal{X}_p$. The results reported in table 5 suggest that $\mathcal{X}_p$ 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, $f_\circ$.

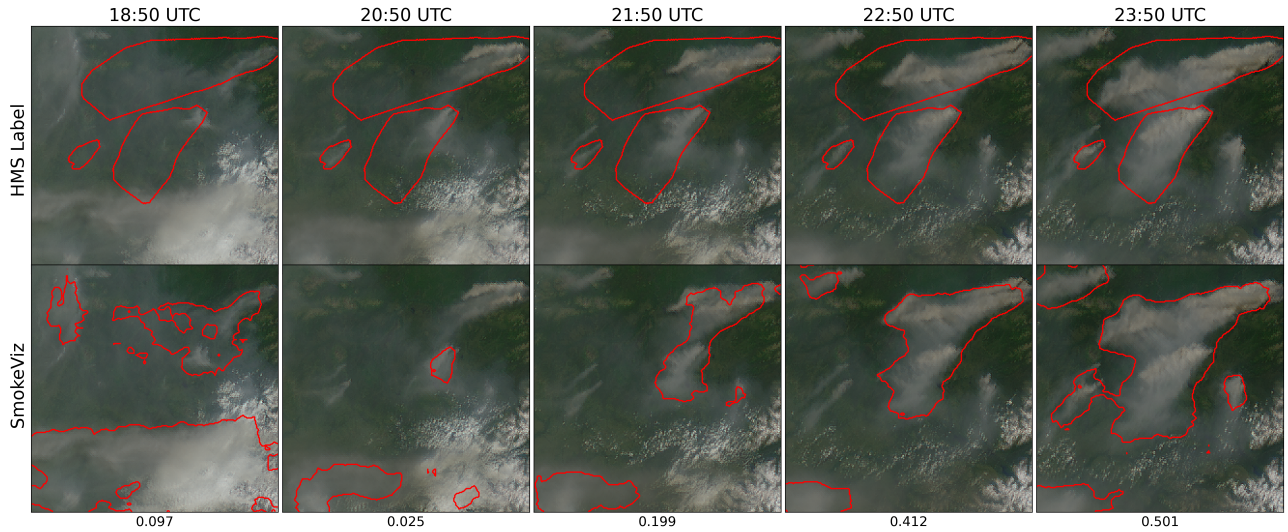

Figure 6. GOES-West imagery showing smoke on June 8th, 2022 in Alaska where, at this geolocation, daylight was between 12:43-7:53 UTC. The HMS smoke annotations (top row) span from 18:50 to 23:50 UTC and are compared to the $f_\circ$ generated pseudo-labels (bottom row). The first column would be the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

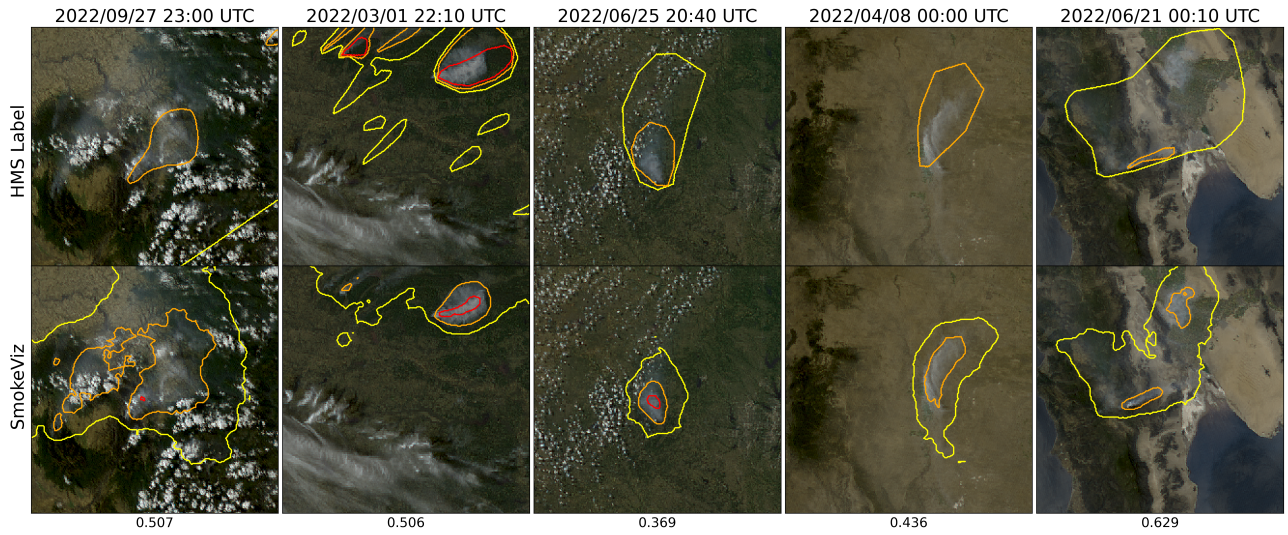

Figure 7. Examples of HMS annotations (top row) vs f_c output (bottom row) on $\mathcal{X}_p$ samples.

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.

References

[1] 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*, vol. 60, no. 1, pp. 179–

232, 2022. 1

[2] 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*, vol. 118, no. 2, p. e2011048118, 2021. 1

[3] 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*, vol. 390, no. 10100, pp. 1345–1422, 2017. 1

[4] R. S. Allison, J. M. Johnston, G. Craig, and S. Jennings, "Airborne optical and thermal remote sensing for wildfire detection and monitoring," *Sensors*, vol. 16, no. 8, p. 1310, 2016. 1

[5] 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. 1

[6] 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*, p. 19526, 2018. 1

[7] 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," tech. rep., Copernicus Meetings, 2024. 1

[8] T. X.-P. Zhao, S. Ackerman, and W. Guo, "Dust and smoke detection for multi-channel imagers," *Remote Sensing*, vol. 2, no. 10, pp. 2347–2368, 2010. 2

[9] 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*, vol. 19, no. 14, pp. 2623–2638, 1998. 2

[10] 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. 2

[11] 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*, vol. 29, no. 20, pp. 6059–6066, 2008. 2

[12] NOAA, "Hazard mapping system fire and smoke product." 2

[13] 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*, vol. 11, no. 14, p. 1702, 2019. 2, 3

[14] 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*, vol. 31, no. 1, pp. 170–176, 2021. 2, 3

[15] J. Wen and M. Burke, "Wildfire smoke plume segmentation using geostationary satellite imagery," *ArXiv*, vol. abs/2109.01637, 2021. 2, 3

[16] 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*, pp. 843–852, 2017. 2

[17] Z. Wang, P. Yang, H. Liang, C. Zheng, J. Yin, Y. Tian, and W. Cui, "Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery," *Remote Sensing*, vol. 14, no. 1, p. 45, 2022. 2, 3

[18] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009. 2

[19] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[20] 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*, pp. 248–255, Ieee, 2009. 2, 6

[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. 2

[22] 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*, vol. 99, no. 7, pp. 1329–1336, 2018. 3

[23] M. Bah, M. Gunshor, and T. Schmit, "Generation of goes-16 true color imagery without a green band," *Earth and Space Science*, vol. 5, no. 9, pp. 549–558, 2018. 3

[24] T. C. Phan and T. T. Nguyen, "Remote sensing meets deep learning: exploiting spatio-temporal-spectral satellite images for early wildfire detection," 2019. 3

[25] 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*, vol. 14, no. 4, pp. 2699–2716, 2021. 3

[26] 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*, vol. 51, no. 12, pp. 1899–1914, 1985. 5

[27] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in *International conference on learning representations*, 2018. 5

[28] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*, pp. 10096–10106, PMLR, 2021. 6

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with

deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 6

[30] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Advances in neural information processing systems*, vol. 27, 2014. 6

[31] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014. 6

[32] P. Iakubovskii, "Segmentation models pytorch." https://github.com/qubvel/segmentation_models.pytorch, 2019. 6

[33] 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*, vol. 13, no. 1, p. 13875, 2023. 6

[34] 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*. 6

[35] 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*, pp. 1848–1857, 2022. 6

[36] 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*, vol. 36, 2024. 6

[37] 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*, vol. 146, no. 726, pp. 141–152, 2020. 6

[38] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE visual communications and image processing (VCIP)*, pp. 1–4, IEEE, 2017. 6

[39] 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*, pp. 2881–2890, 2017. 6

[40] T. Fan, G. Wang, Y. Li, and H. Wang, "Ma-net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020. 6