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

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The increase in the frequency of wildfires on a global scale underscores the need for
2 advancements in fire monitoring techniques for disaster management, environmental
3 protection and to mitigate negative health outcomes. This research introduces
4 an innovative, data-driven framework that leverages the semi-supervised method,
5 pseudo-labeling, to generate smoke plume annotations in geostationary satellite
6 imagery. Unlike many pseudo-labeling applications that aim to increase the la-
7 beled dataset size, the primary objective is use pseudo-labels to refine an existing
8 National Oceanic and Atmospheric Administration smoke dataset that provides
9 temporal and geographical information on individual smoke plumes but at variable
10 and, primarily, low temporal resolution. We use deep learning and pseudo-labels to
11 pinpoint the singular, most representative, satellite image that optimally illustrates
12 the smoke annotation within the given time window. By identifying the most
13 representative imagery of smoke plumes for a given smoke annotation, the study
14 seeks to create an accurate and relevant machine learning dataset. The resulting
15 dataset is anticipated to be an instrumental tool in developing further machine
16 learning models, such as an automated system capable of real-time monitoring and
17 annotation of smoke plumes directly from streaming satellite imagery.

18

1 Introduction

19 In recent years, the escalation of wildfire incidents worldwide has become a prominent environmental
20 and public health concern. The combustion process in wildfires releases smoke containing fine
21 particulate matter (PM2.5) and harmful gases, posing severe hazards to human health and air quality.
22 These risks underscore the necessity for efficient and effective monitoring methods to mitigate the
23 adverse health impacts associated with wildfire smoke.

24 Traditionally, wildfire monitoring has relied on ground-based methods, such as forest service patrols,
25 manned lookout towers, and aviation surveillance [1]. While these methods provide valuable localized
26 insights, they are constrained by geographical and logistical limitations, often failing to deliver timely
27 and comprehensive data, especially over large and remote areas. In contrast, satellite imagery offers
28 a vantage point that overcomes these limitations, providing continuous, wide-area coverage and
29 real-time data crucial for assessing and responding to the health risks posed by wildfire smoke.

30 Satellite imagery, equipped with state-of-the-art sensors, such as the Advanced Baseline Imager
31 (ABI) on the Geostationary Operational Environmental Satellites (GOES) [8], have revolutionized
32 environmental monitoring. These tools enable the detailed observation of smoke plumes, their
33 particulate density, and the extent of smoke spread. These satellite-based systems offer the capabilities

34 to provide critical insights into the concentration and movement of smoke particulates, facilitating
 35 real-time assessments of air quality.
 36 The integration of satellite imagery in wildfire smoke monitoring is not only instrumental in providing
 37 real-time data but also plays a significant role in public health planning and response. By mapping
 38 the spread and density of smoke, health authorities can issue timely warnings, implement evacuation
 39 protocols, and deploy resources effectively to mitigate health risks. Furthermore, long-term data
 40 gathered from satellite observations can aid in understanding the broader impacts of wildfire smoke
 41 on public health, influencing policy decisions and preventive measures.
 42 Currently, multi-channel thresholding is a popular method to distinguish smoke pixels from pixels
 43 containing dust, clouds or other phenomenon with similar signatures [28]. Thresholds are determined
 44 by using historical, labeled data to extract optimal radiance values for each channel that corresponds
 45 with the labeled class. These methods are tuned to particular biogeographies and often have issues
 46 with generalization to new locations with varying fuel types [18].
 47 In contrast to the numerical thresholding approach, human visual inspection of satellite imagery is
 48 another commonly used method for smoke identification. Trained analyst inspect satellite imagery
 49 and label the smoke by hand. An example of hand-labeled annotations is the National Oceanic
 50 and Atmospheric Administration (NOAA) Hazard Mapping System (HMS) fire and smoke product
 51 [15, 23]. For the HMS smoke product, trained satellite analysts use movement characteristics to
 52 help identify smoke by scanning through a time series of satellite imagery. When visual inspection
 53 indicates smoke, the analyst will draw a polygon that corresponds to the geolocation and density
 54 of smoke. By design of the product, the HMS annotations have varying time resolution and are
 55 released on a rolling but undefined schedule ranging from one to multiple times a day as observation
 56 conditions permit. This method is potentially not as scalable as an automated approach and is limited
 57 by the availability of analysts and their time.
 58 To address the challenges associated with thresholding and manual labels, we can look towards
 59 innovative approaches and recent technological advancements in computer vision. Machine learning
 60 methods have shown potential in improving the accuracy and efficiency of satellite-based wildfire
 61 smoke detection and monitoring. For instance, SmokeNet, uses a convolutional neural network
 62 (CNN) based framework to determine if a scene of MODIS satellite imagery contains smoke [2].
 63 Another study, that looked at a singular wildfire event, also used a CNN to identify smoke on a
 64 pixel-wise basis using imagery from Himiware-8 [12]. Additionally, Wen et al. developed a CNN
 65 architecture that takes GOES-East imagery as input and the HMS-generated annotations for the target
 66 labels during training [26].
 67 The success of deep learning methods, such as CNNs, relies heavily on the availability of a large,
 68 representative dataset [24]. As laid out in table 1, prior studies use relatively small numbers of
 69 samples, from 47 [25] to 6825 [26], where one sample represents a satellite image with a singular
 70 time and geolocation. In contrast, benchmark datasets for image classification contain tens of
 71 thousands (CIFAR-10 and MNIST) to millions (CIFAR-100 and ImageNet) of data samples [11],
 72 [6], [5]. Keeping in mind the correlation between both the quality and quantity of data with model
 73 performance, we introduce the largest known smoke dataset, SmokeViz, containing over 130,000
 74 samples.

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
[2]	CNN	6255	MODIS	5	image
[26]	CNN	6825	GOES-East	5	pixel
[12]	CNN	975	Himiware-8	7	pixel
[25]	U-Net	47	Landsat-8	13	pixel
SmokeViz	U-Net	133,871	GOES-East/West	3	pixel

75 Semi-supervised learning is an approach that can be used to increase the number of labeled samples
 76 in a dataset. This is done by leveraging a labeled dataset to generate new labels for an often larger,
 77 but unlabeled, dataset. Pseudo-labeling, a form of semi-supervised learning, uses labeled data to
 78 train an initial model, then runs that model on unlabeled data to predict pseudo-labels, and finally

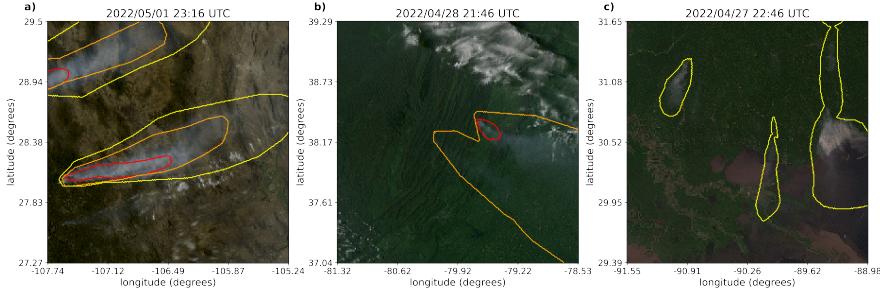


Figure 1: Satellite imagery captured by GOES-East within a few days of each other. The yellow, orange and red contours indicate the extent of Light, Medium and Heavy smoke. a) shows a canonical example of a smoke plume. b) and c) show observable variations in the density labels.

79 trains a new model using the pseudo-labels [13]. We introduce a variation of pseudo-labeling, not to
 80 increase the size, but to increase the quality of our dataset by generating pseudo-labels to select the
 81 best satellite image out of a given time-window to represent each smoke plume annotation.

82 2 Methods

83 Dataset

84 The initial source for smoke labels, discussed in further detail in the HMS Smoke Labels section, is
 85 uniquely characterized by each annotation, y , having corresponding imagery ranging between 1-60
 86 frames, where each frame, x , captures 5 minutes of exposure. Additionally, we have two satellites
 87 that overlap in coverage area, GOES-East and GOES-West, effectively doubling the number of frames
 88 for a single annotation. For the set of smoke annotations, \mathcal{Y} , $y \in \mathcal{Y}$ uses one or more $x \in \mathcal{X}$ where
 89 \mathcal{X} is the entire set of satellite imagery corresponding to the set of time windows defined by the
 90 labels. We apply pseudo-labeling to develop a subset of \mathcal{X} , denoted as \mathcal{X}_p , that has a one-to-one
 91 annotation-to-image ratio such that $|\mathcal{X}_p| = |\mathcal{Y}|$, where we choose the satellite image that has the
 92 maximum overlap between the geolocation of smoke in the imagery and the analyst annotation.

93 Dataset development came in three stages. First, we create an initial dataset, \mathcal{X}_M , that leverages light
 94 scattering physics to determine which singular satellite image would be in the optimal configuration
 95 for smoke detection. Second, we used \mathcal{X}_M to train an initial parent model, f_o , that identifies smoke in
 96 satellite imagery. Third, we use f_o to label each satellite image in a given annotation's time-window
 97 and the optimal satellite image is chosen based on which image's pseudo-labels has the greatest
 98 overlap with the analyst annotation for the given location and densities of smoke.

99 HMS Smoke Labels

100 NOAA manages environmental satellite programs such as the HMS program, the HMS program is an
 101 operational system that uses an aggregation of satellite data to generate active fire and smoke data.
 102 To train our model, we implement a supervised learning framework that uses the HMS analyst smoke
 103 product as truth labels during the model training process.

104 HMS smoke analysis data gives the coordinates of the smoke perimeter as a polygon and classifies
 105 the smoke by density within a given time window. The time windows can range from instantaneous
 106 (same start and end time) to lengths of 5 hours. While the true bounds of the smoke can change
 107 within the larger time spans, the analyst is making an approximation that should reflect the smoke
 108 coverage over the duration of the time window. The density information is qualitatively determined
 109 by each analyst based on the apparent smoke opacity in the satellite imagery and categorized as either
 110 light, medium or heavy as seen in figure 1a [16].

111 **Thermometer Encoding Smoke Densities**

112 One of the challenges introduced with using human generated qualitative smoke densities was that, as
113 seen in figure 1b and 1c, there are variations in what is labeled as heavy or light density smoke. More
114 generally, reproducing qualitative metrics with quantitative algorithms is a challenging problem, but
115 we apply mathematical approaches that mitigate some of the underlying complications of our specific
116 problem. Despite the fact that the smoke densities introduce qualitative complexities, we decided
117 that the density approximations were important to use in our dataset because of the differences in
118 signatures the densities produce. Within the satellite imagery, the appearance of a light density
119 smoke plume will look significantly different than a heavy density smoke plume as seen in figure 1.
120 Additionally, a light density smoke plume is expected to be more challenging to detect since it is easier
121 for it to be misclassified as not smoke. During the training process, the separate density categories
122 allows us to deferentially weight the penalization given to the model for incorrect classifications
123 based on category. For example, the model can be given a small penalization for misclassifying light
124 smoke as not smoke while given a higher penalization for misclassifying heavy smoke as not smoke.
125 In addition to the densities being ordered and categorical, the differences between the density
126 categories are not evenly distributed by a given metric, such as particulate matter per square meter.
127 The intervals between densities being unknown along with the hierarchical nature of the density labels
128 makes the labels ordinal instead of just categorical. This data property allows us to use thermometer
129 encoding [4], which leverages the idea that heavy density smoke includes both medium and light
130 density smoke, that heavy density smoke is closer to medium than it is to light, and automatically
131 weights the loss functions and incorporates the ranked ordering of the densities. As seen in Table 2,
132 one-hot encoding, commonly used for categorical data, doesn't take ordinal properties of the data
133 into consideration.

Table 2: 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]

134 **Time Windows For Smoke Annotations**

135 In order to take into account movement characteristics to help identify smoke, analysts use multi-
136 frame animations of the satellite imagery. The resulting annotations often have large time windows
137 over multiple hours to represent one smoke plume annotation. Since the goal of these annotations is
138 to show the general coverage over that time span, as shown in figure 2, the smoke boundaries don't
139 often match up with the satellite imagery over the entire time window. One way to approach this
140 problem would be to use all the satellite images the analysts used as input. Since the timespans are
141 non-uniform, this would vary the length in imagery inputs into the model, which would be difficult
142 with a CNN architecture. Moreover, this would require a large amount of additional memory and
143 computational resources. Instead of using the original analysts' many satellite image inputs to one
144 annotated output, we develop a one-to-one input-to-output by finding the optimal singular satellite
145 image input to represent the annotation. Discussed in further detail in the next section, we do this
146 by making physics-driven choices on which satellite and timestamp would give the optimal angle
147 between the sun and satellite that would produce the strongest smoke signature for the geolocation
148 and timestamp of the smoke plume.

149 **Satellite Imagery**

150 The GOES satellites are operated by NOAA in order to support meteorology research and forecasting
151 for the United States. We use the latest operational satellites, GOES-16 (East), 17 and 18 (West)
152 that each carry the ABI, that measure 16 bands between the visible and infrared wavelengths. In
153 improvement to the GOES predecessors, imagery is collected every 5 minutes for the contiguous
154 United States and every 10 minutes for the full disk. Using PyTroll, a Python framework for

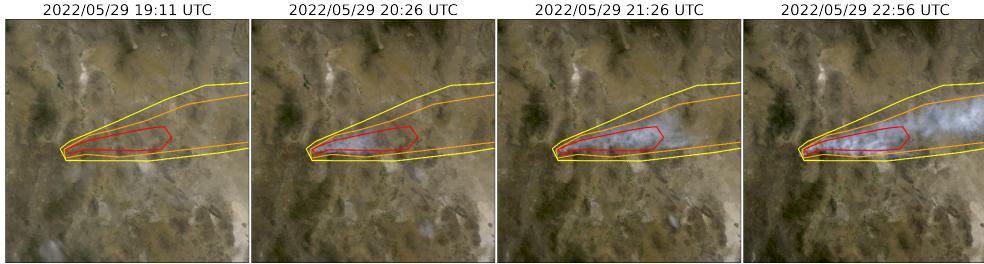


Figure 2: True color GOES-East imagery from May 2022, Southeast New Mexico (31°N , 100°W) during the start of the Foster Fire. The red, orange and yellow lines represent the heavy, medium and low density HMS smoke annotations that span 19:10–23:00 UTC.

Table 3: To create a true color image, we use the following bands from the ABI Level 1b CONUS (ABI-L1b-RadC) product.

band	description	center wavelength (μm)	spatial resolution (km)
C01	blue visible	0.47	1
C02	red visible	0.64	0.5
C03	veggie near infrared	0.865	1

155 processing satellite data [19], we input bands 1-3 (Table 3) to a GOES true color composite algorithm
 156 to develop a true color image representation [3], similar to what is seen by HMS analysts.

157 Mie-Derived Dataset

158 We used a physics-informed approach in selecting the initial GOES dataset, \mathcal{X}_M , which we call the
 159 Mie-derived dataset, for training an initial parent model, f_o , where if \mathcal{X} represents all the GOES
 160 imagery corresponding to the HMS smoke annotation time window, $\mathcal{X}_M \subset \mathcal{X}$. Prior GOES ABI
 161 datasets for machine learning applications often include data from only one of the two GOES-series
 162 satellites, commonly opting for GOES-East [26], [17], [14]. Rather than using one satellite or the
 163 cumulative data from both GOES-West and GOES-East images, we select between one or the other
 164 based on the solar zenith angle. For smoke identification, this approach can achieve a much higher
 165 signal-to-noise than imaging the earth’s surface from an arbitrary angle. The elastic scattering of
 166 light is the primary mechanism to account for - while the atmosphere is composed of molecules
 167 with size $< 1\text{nm}$, smoke particles can vary from $100\text{ nm} – 10\text{ }\mu\text{m}$ in diameter, d . The GOES ABI
 168 covers spectral bands from $0.47\text{ }\mu\text{m} – 13.3\text{ }\mu\text{m}$, so atmospheric and smoke particle sizes occupy two
 169 very different regimes with respect to the imaging wavelength λ . In the extreme limit of $\lambda \gg d$, the
 170 physics of scattering of light off a small sphere is captured by Rayleigh scattering. This process has
 171 two critical consequences: (1) the scattering cross section of light is strongly wavelength dependent
 172 (scaling with λ^{-4}), meaning that photons with wavelength closer to the ultraviolet are scattered more
 173 strongly than infrared photons. (2) the scattering cross section scales with an angular dependent
 174 cross section of $(1 + \cos^2 \theta)$. Scattered photons follow the emission distribution of a radiating dipole,
 175 scattering more strongly in the forward and backwards directions ($\theta = 0, \pi$) than orthogonal to the
 176 direction of propagation ($\theta = \pi/2, 3\pi/2$), see figure 3 for a Rayleigh scattering schematic.

177 The significance of these scalings is that the observer, or detector, will receive blue photons in most
 178 directions orthogonal to the source. Equivalently, photons traveling colinearly with line of sight to
 179 the emission source will mostly have wavelengths in the infrared band. In the converse regime of
 180 $d > \lambda$, the elastic scattering of light against matter is modeled through Mie scattering. In comparison
 181 to Rayleigh scattering, Mie scattering is largely wavelength-independent and has a more complicated
 182 radiation pattern where the cross section has a maximal amplitude in the forward direction. An
 183 observer downstream of this scatterer will collect more photons than one positioned directly behind it.
 184 In the context of smoke identification, a sunrise or sunset will lead to a higher Mie scattered signal in

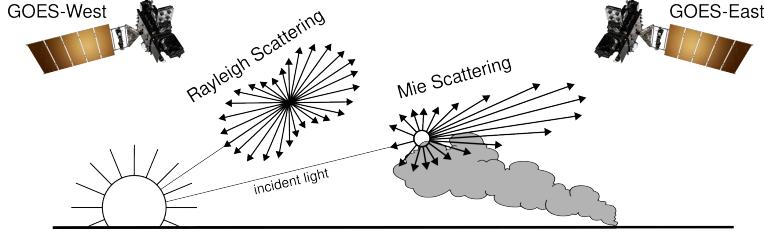


Figure 3: If the particle size is $< \frac{1}{10}$ the wavelength of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than the wavelength of light. This schematic demonstrates that when the sun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

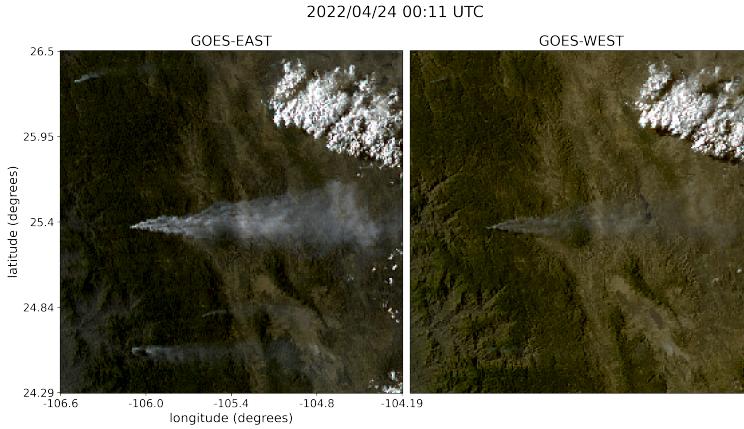


Figure 4: True color GOES-East (left) and GOES-West (right) imagery from April 24th, 2022 in Durango, Mexico. The images were taken ~ 0.5 hours before sunset (01:43 UTC) for this geolocation and time of year.

185 GOES-West and GOES-East respectively, as shown with a smoke plume producing a stronger signal
186 in GOES-East imagery near sunset in figure 4.

187 Smoke identification therefore amounts to extracting a signal of $d > \lambda$ photons from the $\lambda \gg d$
188 background. Positioning a detector along line of sight to the scatterer will result in a higher signal
189 from smoke particles (figure 3). Filtering the imaged wavelength can enhance this signal; photons
190 collected in the blue spectrum will have a naturally lower background along the line of sight to the
191 illumination source do their high level of Rayleigh scattering as. Therefore, as demonstrated in figure
192 5, this configuration results in the highest signal to noise imaging for smoke particles.

193 Based solely on these criteria, the optimal strategy would be to pull data from GOES-West right after
194 sunrise and from GOES-East right before sunset. Another factor to consider is that the time when the
195 sun is in optimal alignment with the satellite for smoke detection coincides with when solar zenith
196 angle is maximized. Larger angles between the satellite and sun result in an increase in noise due
197 to increased atmospheric interactions [22]. This is shown in figure 6, while we optimize for smoke
198 signal detection, due to the high solar zenith angle, we introduce atmospheric interaction noise that
199 obfuscate the smoke signal. To reduce the noise from large solar zenith angles, if given multiple
200 frames to choose from, we choose the image with the largest solar zenith angle that is $< 80^\circ$.

201 The resulting image selection process takes into account atmospheric properties and light scattering
202 physics to generate an estimate of which singular satellite image within the analyst time-window could
203 give the highest smoke signal-to-noise ratio. The resulting Mie-derived dataset, $\mathcal{D}_M = \{X_M, Y\}$,
204 was then used to train a model, f_o , that would generate N pseudo-labels, y^* , for every sample,
205 where N is determined by how many images, taken at a 10 minute interval, fit within the analyst

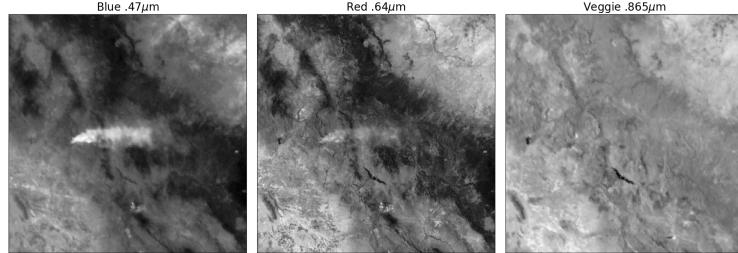


Figure 5: Three bands of GOES-East data are the raw input to generate a true color image. These plots show variations in the signal-to-noise ratio for smoke detection in relation to the wavelength, λ , of light being measured.

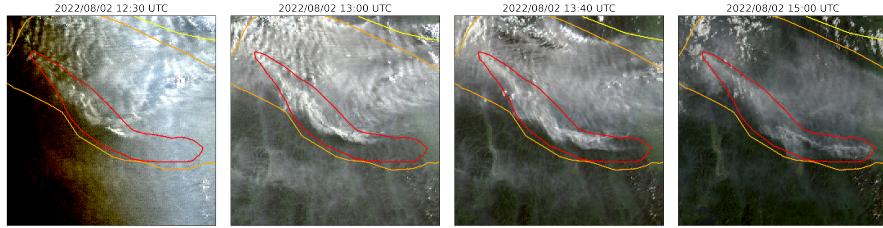


Figure 6: A smoke annotation projected onto GOES-West imagery from August 2022 that spans from 11:00 UTC to 15:00 UTC, sunrise on August 2nd, 2022 at coordinates (49°24'N, 115°29'W) was 12:15 UTC.

206 time-window for that sample. Chosen from the N images, x_p is the image with the highest alignment
 207 between the f_o prediction of smoke, y^* , in the image and the HMS analysts' annotation y .

208 Machine Learning Model

209 We implement a deep learning architecture that uses the encoder from the ResNet model [9] and a
 210 semantic segmentation classifier from the U-Net model [21]. Transfer learning has shown to reduce
 211 the time and resources needed to train a model by leveraging information from pre-trained models
 212 [27], [20]. We initialize the values of our model weights using the pre-trained values originally
 213 trained on the ImageNet dataset [5], containing 1.2 million images and 1000 categories. Our model
 214 was developed using the Segmentation Models PyTorch package [10] that was written as a high level
 215 API for implementing models for semantic segmentation problems. We input 256x256x3 snapshots
 216 of true color GOES imagery that contains smoke and output a 256x256x3 classification map that
 217 predicts if a pixel contains smoke and if so, what the density of that smoke is. As mentioned earlier,
 218 we apply the thermometer encoding shown in table 2 to encode the smoke densities and apply binary
 219 cross entropy as the loss function per density of smoke.

220 The dataset, \mathcal{D}_M , contained over 130,000 samples. To train f_o , we split \mathcal{D}_M into training (118,691
 221 samples), validation (8,100 samples) and testing (7,080) datasets. Training data contains data from
 222 the years 2018, 2019, 2020, 2021 and 2023 while the data from 2022 is split into validation and
 223 testing sets by taking data from alternating 10 days of the year. In order to make sure we include
 224 the monthly variations in wildfire trends over a full year, we split 2022 data up by every 10 days.
 225 This allowed us to: (1) allocate an additional full year of data for the training set, (2) show yearlong
 226 trends in both the validation and testing sets and (3) keep the validation and testing datasets relatively

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

	f_o		f_c	
	\mathcal{D}_M	\mathcal{D}_p	\mathcal{D}_M	\mathcal{D}_p
Light	0.394	0.551	0.437	0.560
Medium	0.283	0.392	0.345	0.417
Heavy	0.233	0.290	0.275	0.295
Overall	0.365	0.510	0.412	0.518

227 independent from one another since only two out of every ten days of data will have adjacent days in
228 validation and testing.

229 We trained the parent model, f_o , for 10 epochs. To determine which image out of the relevant imagery
230 for the given time window best represents the analyst annotation, we run f_o on each x to generate a
231 pseudo-label, y^* . The output of f_o , y^* , give predictions on if smoke is in the image, and if there is
232 smoke, where the smoke is in that image and the density of that smoke. y^* serve as pseudo-labels
233 for each density of smoke and are compared to the analyst annotations, y . To compare y^* and y , we
234 calculate the IoU using the total set of pixels for y^* at that density of smoke and the entire set of
235 pixels for y for a particular smoke density in each image as shown in equation 1. The image with the
236 highest IoU score is chosen as the image, x_p , that best represents the analyst smoke annotation, y .
237 Often used for pseudo-labeling, a confidence threshold value is defined to determine if a pseudo-label
238 should to be included in a dataset [7]. We chose a confidence threshold that would include the sample,
239 x_p , in \mathcal{X}_p if the maximum overall IoU (equation 1) between y^* and y over all densities was over 0.1.

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

240 Finally, we use \mathcal{D}_p to train an additional child model, f_c . We use the same dataset split method and
241 model setup but change \mathcal{D}_M to \mathcal{D}_p to train the model over 10 epochs.

242 Results

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

250 An illustration of a pseudo-label picked image better representing the analyst annotation when
251 compared to the Mie-derived image selection is evident in Figure 7, where the heavy density smoke
252 IoU increases from 0.01 to 0.59. The analyst annotation for these densities cover 5 hours of imagery,
253 the Mie-derived selection optimizes for the image closest to sunrise while the pseudo-label image
254 selection chooses the image with the highest overlap between the pseudo-label and the analyst
255 annotation.

256 3 Limitations

257 One of the concerns that comes with using pseudo-labeling methods is that you can perpetuate biases
258 from the parent model into subsequent child models. Due to the increase in detectable forward
259 scattered light off smoke particular matter, we expect the model to have a bias towards producing a
260 higher success rate for smoke detection at larger solar zenith angles. This could potentially cause poor

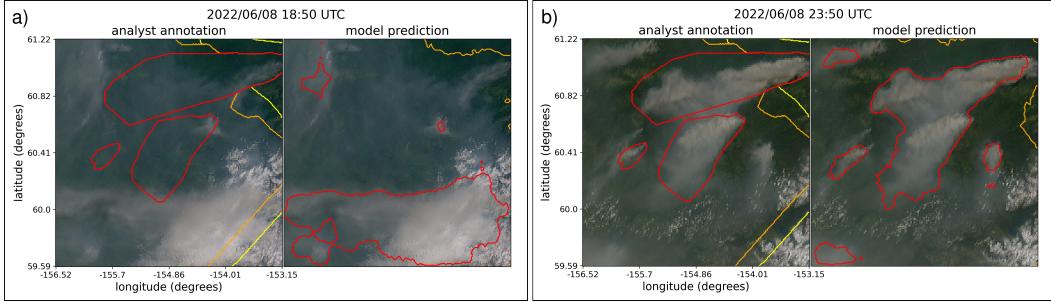


Figure 7: GOES-West imagery showing smoke on June 8th, 2022 in Alaska where, at this geolocation, daylight was between 12:43-7:53 UTC. The HMS smoke annotations displayed span from 18:50 to 23:50 UTC. a) shows the imagery that was selected using the Mie-derived data selection process b) shows the image that had the highest IoU score between the f_o generated pseudo-label and the analyst annotation.

261 model performance when monitoring smoke during the middle of the day and should be analyzed in
262 future work.

263 4 Conclusion

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

280 To improve the dataset’s capacity to accurately represent wildfire smoke plumes, we train a parent
281 machine learning model, f_o , using the Mie-derived dataset, \mathcal{D}_M , and run it on the relevant satellite
282 images within the time-frame. The image with the maximum IoU score between the model’s smoke
283 predictions, or pseudo-label, and the analyst smoke annotations are used to create the pseudo-label
284 generated dataset, \mathcal{D}_p . We then train a child model, f_c , using \mathcal{D}_p and test f_o and f_c on both the 2022
285 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
286 performing model, f_c , that gave higher IoU metrics on both dataset’s testing sets in comparison to
287 the original parent model, f_o .

288 The result of this study is a representative dataset that can be used to train machine learning models
289 for various wildfire smoke applications. A future goal is to produce a robust and reliable machine
290 learning based approach for detecting wildfires using satellite imagery. That information can be used
291 for wildfire monitoring and as data provided to public health officials for air quality assessments. On
292 a broader scale, we show how pseudo-labeling can be used to optimize a dataset when the resolution
293 for the data and corresponding labels do not match. This could be useful in similar applications
294 involving time-series/video data with a singular label where the data can be compressed while still
295 remaining representative of the label.

296 **5 Acknowledgments and Disclosure of Funding**

297 This research was supported in part by NOAA cooperative agreement NA22OAR4320151, for the
298 Cooperative Institute for Earth System Research and Data Science (CIESRDS). We thank Wilfrid
299 Schroeder and the Hazard Mapping Systems team for giving guidance on how they created their
300 smoke plume dataset. This work utilized the Alpine high performance computing resource at the
301 University of Colorado Boulder. Alpine is jointly funded by the University of Colorado Boulder, the
302 University of Colorado Anschutz, Colorado State University, and the National Science Foundation
303 (award 2201538). The statements, findings, conclusions, and recommendations are those of the
304 author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

305 **References**

- 306 [1] R. S. Allison, J. M. Johnston, G. Craig, and S. Jennings. Airborne optical and thermal remote
307 sensing for wildfire detection and monitoring. *Sensors*, 16(8):1310, 2016.
- 308 [2] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo. Smokenet: Satellite smoke scene detection using
309 convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 11(14):
310 1702, 2019.
- 311 [3] M. Bah, M. Gunshor, and T. Schmit. Generation of goes-16 true color imagery without a green
312 band. *Earth and Space Science*, 5(9):549–558, 2018.
- 313 [4] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to
314 resist adversarial examples. In *International conference on learning representations*, 2018.
- 315 [5] J. Deng, K. Li, M. Do, H. Su, and L. Fei-Fei. Construction and Analysis of a Large Scale Image
316 Ontology. Vision Sciences Society, 2009.
- 317 [6] L. Deng. The mnist database of handwritten digit images for machine learning research [best of
318 the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.
319 2211477.
- 320 [7] R. E. Ferreira, Y. J. Lee, and J. R. Dórea. Using pseudo-labeling to improve performance of
321 deep neural networks for animal identification. *Scientific Reports*, 13(1):13875, 2023.
- 322 [8] S. J. Goodman, T. J. Schmit, J. Daniels, and R. J. Redmon. *The GOES-R series: a new
323 generation of geostationary environmental satellites*. Elsevier, 2019.
- 324 [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.
- 325 [10] P. Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.
- 327 [11] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- 328 [12] A. Larsen, I. Hanigan, B. J. Reich, Y. Qin, M. Cope, G. Morgan, and A. G. Rappold. A deep
329 learning approach to identify smoke plumes in satellite imagery in near-real time for health risk
330 communication. *Journal of exposure science & environmental epidemiology*, 31(1):170–176,
331 2021.
- 332 [13] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep
333 neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07
334 2013.
- 335 [14] Y. Lee, C. D. Kummerow, and I. Ebert-Uphoff. Applying machine learning methods to detect
336 convection using geostationary operational environmental satellite-16 (goes-16) advanced
337 baseline imager (abi) data. *Atmospheric Measurement Techniques*, 14(4):2699–2716, 2021.
- 338 [15] D. McNamara, G. Stephens, M. Ruminski, and T. Kasheta. The hazard mapping system (hms) -
339 noaa’s multi-sensor fire and smoke detection program using environmental satellites. *Conference
340 on Satellite Meteorology and Oceanography*, 01 2004.

- 341 [16] NOAA. Hazard mapping system fire and smoke product. URL <https://www.ospo.noaa.gov/Products/land/hms.html#about>.
- 342 [17] T. C. Phan and T. T. Nguyen. Remote sensing meets deep learning: exploiting spatio-temporal-spectral satellite images for early wildfire detection. 2019.
- 343 [18] T. Randriambelo, S. Baldy, M. Bessafi, M. Petit, and M. Despinoy. An improved detection
344 and characterization of active fires and smoke plumes in south-eastern africa and madagascar.
345 *International Journal of Remote Sensing*, 19(14):2623–2638, 1998.
- 346 [19] M. Raspaud, D. Hoese, A. Dybbroe, P. Lahtinen, A. Devasthale, M. Itkin, U. Hamann, L. Ø.
347 Rasmussen, E. S. Nielsen, T. Leppelt, et al. Pytroll: An open-source, community-driven python
348 framework to process earth observation satellite data. *Bulletin of the American Meteorological Society*,
349 99(7):1329–1336, 2018.
- 350 [20] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an
351 astounding baseline for recognition, 2014.
- 352 [21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image
353 segmentation, 2015.
- 354 [22] A. Royer, P. Vincent, and F. Bonn. Evaluation and correction of viewing angle effects on
355 satellite measurements of bidirectional reflectance. *Photogrammetric engineering and remote
356 sensing*, 51(12):1899–1914, 1985.
- 357 [23] W. Schroeder, M. Ruminski, I. Csizar, L. Giglio, E. Prins, C. Schmidt, and J. Morisette.
358 Validation analyses of an operational fire monitoring product: The hazard mapping system.
359 *International Journal of Remote Sensing*, 29(20):6059–6066, 2008.
- 360 [24] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in
361 deep learning era, 2017.
- 362 [25] Z. Wang, P. Yang, H. Liang, C. Zheng, J. Yin, Y. Tian, and W. Cui. Semantic segmentation and
363 analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery.
364 *Remote Sensing*, 14(1):45, 2022.
- 365 [26] J. Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery.
366 *ArXiv*, abs/2109.01637, 2021. URL [https://api.semanticscholar.org/CorpusID:
367 237416777](https://api.semanticscholar.org/CorpusID:237416777).
- 368 [27] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural
369 networks?, 2014.
- 370 [28] T. X.-P. Zhao, S. Ackerman, and W. Guo. Dust and smoke detection for multi-channel imagers.
371 *Remote Sensing*, 2(10):2347–2368, 2010. ISSN 2072-4292. doi: 10.3390/rs2102347. URL
372 <https://www.mdpi.com/2072-4292/2/10/2347>.