
SmokeViz: A Large-Scale Satellite Dataset for Wildfire Smoke Detection and Segmentation

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The global rise in wildfire frequency and intensity over the past decade underscores
2 the need for improved fire monitoring techniques. To advance deep learning re-
3 search on wildfire detection and its associated human health impacts, we introduce
4 **SmokeViz**, a large-scale machine learning dataset of smoke plumes in satellite
5 imagery. The dataset is derived from expert annotations created by smoke analysts
6 at the National Oceanic and Atmospheric Administration, which provide coarse
7 temporal and spatial approximations of smoke presence. To enhance annotation
8 precision, we propose **pseudo-label dimension reduction (PLDR)**, a generalizable
9 method that applies pseudo-labeling to refine datasets with mismatching temporal
10 and/or spatial resolutions. Unlike typical pseudo-labeling applications that aim to
11 increase the number of labeled samples, PLDR maintains the original labels but
12 increases the dataset quality by solving for intermediary pseudo-labels (IPLs) that
13 align each annotation to the most representative input data. For SmokeViz, a parent
14 model produces IPLs to identify the single satellite image within each annotations
15 time window that best corresponds with the smoke plume. This refinement process
16 produces a succinct and relevant deep learning dataset consisting of over 180,000
17 manual annotations. The SmokeViz dataset is expected to be a valuable resource to
18 develop further wildfire-related machine learning models and is publicly available
19 at [\[aws download link\]](#).

20

1 Introduction

21 Due in part to environmental regulation, average fine particulate matter ($PM_{2.5}$) levels in the United
22 States have declined over recent decades [2]. However, from 2010 to 2020, the contribution of
23 wildfire smoke to $PM_{2.5}$ concentrations more than doubled, accounting for up to half of total exposure
24 in parts of the Western U.S. [6].

25 Due in part to public policy, average fine particulate matter ($PM_{2.5}$) levels in the United States have
26 declined over recent decades [2]. However, from 2010 to 2020, the contribution of wildfire smoke to
27 $PM_{2.5}$ concentrations more than doubled, accounting for up to half of total $PM_{2.5}$ exposure in Western
28 U.S. [6]. This is particularly concerning, as ambient $PM_{2.5}$ is a leading environmental risk factor for
29 adverse health outcomes and premature mortality [10]. These trends/risks highlight the urgent need
30 for scalable and timely smoke monitoring systems to mitigate public health risks.

31 Satellite imagery offers the spatial coverage and temporal frequency needed for large-scale smoke
32 monitoring. Geostationary satellites such as the GOES series [11] are especially well-suited to
33 this task, providing persistent observation over fixed regions—essential for capturing the dynamic
34 behavior of wildfire smoke.

35 Satellite imagery offers the spatial coverage and temporal frequency needed for large-scale smoke
36 monitoring. In comparison to polar-orbiting satellites like Suomi or Sentinel, Geostationary satellites
37 such as the GOES series [11] are especially well-suited to this task, providing persistent observation
38 over fixed regions—essential for capturing the dynamic behavior of wildfire smoke plumes. The
39 high temporal resolution and wide coverage of GOES imagery enable real-time tracking of smoke
40 concentration and movement, supporting air quality assessments and early warning systems.

41 Even with the advances in remote sensing, existing deep learning satellite datasets for wildfire smoke
42 detection face several limitations. They are often small in scale, restricted to specific regions or events,
43 and focus on scene-level classification rather than pixel-level segmentation. Most do not differentiate
44 between smoke density levels, are not publicly available, and lack standardized benchmarks for
45 semantic segmentation. While NOAA’s Hazard Mapping System (HMS) provides a large-scale,
46 expert-labeled dataset, its annotations span multi-hour time windows that vary in duration. This
47 creates a temporal mismatch between the labels and individual satellite frames, complicating their
48 direct use for supervised learning.

49 To address these challenges, we introduce **SmokeViz**, a large-scale satellite dataset for semantic
50 segmentation of wildfire smoke plumes. SmokeViz includes over 180,000 annotated samples derived
51 from GOES-East and GOES-West imagery, aligned with HMS analyst annotations. To resolve the
52 temporal ambiguity in the original labels, we propose a semi-supervised method called **pseudo-label**
53 **dimension reduction (PLDR)**, which uses intermediary pseudo-labels to select the satellite image
54 that best matches each smoke annotation. The resulting dataset provides one-to-one image-to-label
55 pairs with ordinal smoke density masks, suitable for supervised deep learning.

56 **SmokeViz** serves as a benchmark for wildfire smoke segmentation and as a resource for the broader
57 machine learning community working with geospatial, temporal, and remote sensing data. It supports
58 new directions in ordinal segmentation, semi-supervised learning with temporal uncertainty, and
59 pretraining for Earth observation tasks involving dynamic atmospheric phenomena.

60 **Our contributions are:**

- 61 • We introduce **SmokeViz**, the largest satellite-based dataset for wildfire smoke segmentation,
62 with over 180,000 samples from GOES imagery.
- 63 • We propose **PLDR**, a physics-guided semi-supervised method for aligning coarse human
64 annotations with temporally optimal satellite imagery.
- 65 • We provide benchmark segmentation baselines and standardized training splits to support
66 reproducibility and downstream research.

67 **2 Related Work**

68 **2.1 Smoke Detection and Labeling Methods**

69 Multi-channel thresholding remains a widely used method for distinguishing smoke from similar
70 atmospheric signatures such as dust or clouds using channel-specific radiance values [29]. These
71 thresholds are typically derived from labeled historical data and are fine-tuned to specific regions
72 and fuel types, limiting their generalizability [20]. In contrast, the SmokeViz dataset spans a wide
73 range of biogeographies across North America and can serve as a source of refined analyst-labeled
74 examples for developing more generalizable thresholding techniques.

75 Large parameterized numerical models are used for forecasting smoke dispersion, but not for smoke
76 detection itself. Systems such as HRRR-Smoke and RRFS [14, 1] rely on computationally intensive
77 forecasts requiring nearly 200 dynamic meteorological inputs. A key limitation of these models is
78 the absence of a real-time smoke analysis product for data assimilation, resulting in delayed model
79 spin-up and compounded forecast errors. Model predictions from SmokeViz could help fill this gap,
80 offering a real-time, satellite-driven alternative to support data assimilation for operational smoke
81 dispersion forecasting.

82 Manual smoke labeling is performed by trained analysts through visual inspection of satellite imagery.
83 NOAA’s Hazard Mapping System (HMS) provides a analyst-labeled wildfire smoke dataset [18, 23].
84 HMS analysts examine GOES imagery sequences to track smoke plume movement and annotate the
85 approximate spatial extent and qualitative density of smoke (light, medium, heavy), as illustrated

in Figure 2.1. Annotations are issued on a rolling basis and span time windows ranging from instantaneous to over 20 hours [19]. While HMS provides high-quality expert annotations, its operational format introduces challenges for supervised learning: annotations are temporally coarse, vary in length, and lack one-to-one correspondence with satellite frames. SmokeViz refines HMS annotations into temporally resolved, frame-aligned labels, enabling real-time, continuous predictions of smoke extent and density.

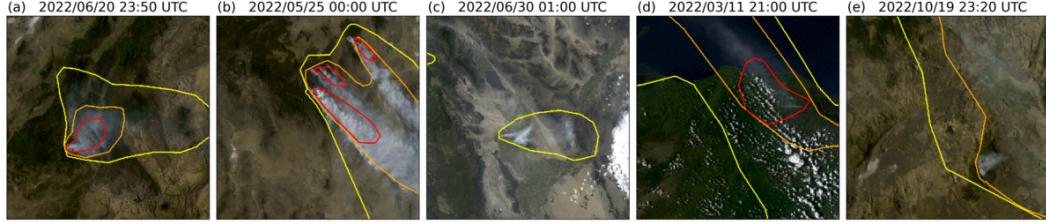


Figure 1: HMS smoke annotations overlaid on GOES imagery. Yellow, orange, and red contours indicate light, medium, and heavy smoke density, respectively. (a) and (b) show canonical smoke plumes; (c)–(e) illustrate density label variation across scenes.

2.2 Deep Learning Datasets and Models for Wildfire Smoke

Recent efforts have applied deep learning to wildfire smoke detection using a variety of satellite sources and label strategies. SmokeNet [3] employs a convolutional neural network (CNN) to classify MODIS image scenes as containing smoke or not, using student-provided labels. SatlasPretrain [5] includes a small set of Sentinel-2 images labeled for smoke as part of a larger multi-label pretraining dataset. While scene classification methods can provide wildfire detection information, they do not capture spatial characteristics of smoke plumes that segmentation would be more appropriate to capture.

Several datasets have been developed for smoke segmentation, but they are have been limited in scope. Wen et al. [27] trained a CNN on GOES-East imagery over California and Nevada using HMS annotations from the 2018 wildfire season. Larsen et al. [16] used Himawari-8 data to detect smoke at the pixel level for a single fire event, using a threshold-based algorithm as ground truth. Table 1 compares these datasets in terms of scale, source, and labeling. SmokeViz stands out by offering over 180,000 samples with analyst-generated, frame-aligned labels covering multiple fire seasons, regions, and biogeographies. It is, to our knowledge, the largest and most diverse dataset for smoke plume segmentation.

Table 1: Comparison of satellite smoke plume datasets, detailing the number of smoke plume samples, satellite source, number of spectral bands, labeling method, classification type - scene classification (SC) or semantic segmentation (SS), and public availability.

reference	# samples	satellite	# bands	label	task	avail.
[3]	1016	MODIS	5	students	SC	no
[5]	125	Sentinel-2	3	crowd sourced	SC	yes
[27]	4095	GOES-East	5	HMS analysts	SS	no
[16]	975	Himawari-8	7	algorithm	SS	no
SmokeViz	183,672	GOES-East/West	3	HMS analysts	SS	yes

In addition to its relevance for wildfire applications, SmokeViz contributes a challenging benchmark for general-purpose remote sensing vision tasks. Unlike many existing datasets that avoid cloudy scenes [25, 13] or focus on sharply bounded features such as cropland [13], infrastructure [30], or oceanic clouds [15, 24], smoke has amorphous, fading boundaries in both space and time. Incorporating smoke segmentation into large-scale pretraining corpora, such as SatlasPretrain [5], could enhance generalizable models for Earth observation.

114 **2.3 Pseudo-labeling**

115 Semi-supervised learning techniques such as pseudo-labeling have been widely used to expand
116 training data by leveraging unlabeled samples [17]. Typically, a parent model is trained on labeled
117 data and then used to generate pseudo-labels for an unlabeled dataset, which are in turn used to train
118 subsequent models in an iterative process.

119 In contrast, we propose a non-iterative variation focused not on data expansion, but dataset data-to-
120 label precision. Our method, **pseudo-label dimension reduction (PLDR)**, generates intermediary
121 pseudo-labels (IPLs) for each satellite frame within the HMS annotation window. Rather than using
122 these labels for training, we use them to identify the satellite image with the greatest alignment to
123 the analyst annotation. This enables the construction of SmokeViz, a temporally disambiguated,
124 one-to-one image-to-label dataset. The resulting dataset methodically pairs the analyst-generated
125 smoke plume labels with selected GOES imagery, enabling high-resolution, temporally accurate
126 segmentation model training.

127 **3 Methods**

128 **3.1 Datasets**

129 We use imagery from the latest GOES satellites, GOES-16 (East), GOES-17 and GOES-18 (West),
130 each equipped with the ABI, which captures 16 spectral bands from visible to infrared wavelengths
131 (λ s) every 10 minutes. We process bands 1-3 using PyTroll [21] to generate 1km true-color
132 composites [4], matching the composite viewed by HMS analysts. These bands correspond to the
133 shortest λ s available on ABI and yield the highest signal-to-noise ratio (SNR).

134 To leverage movement characteristics of smoke, HMS analysts use multi-frame animations of the
135 satellite imagery. The time windows are defined by approximate number of frames the analyst
136 views smoke, often with buffers before the fire starts. Annotations span varying time windows,
137 often covering multiple hours, averaging three hours per annotation. Since the HMS annotations are
138 designed to reflect overall plume extent during a time window rather than at any specific moment,
139 smoke boundaries in individual frames may not align well with the annotation (Figure 2). A naive
140 modeling approach would use all frames within each time window as input, but this introduces
141 non-uniform sequence lengths and significantly increases memory and computational demands,
142 which complicates the use of deep learning architectures and model usage in real-time operations.
143 Instead, we establish a one-to-one mapping by identifying the single satellite frame that best matches
144 each analyst annotation.

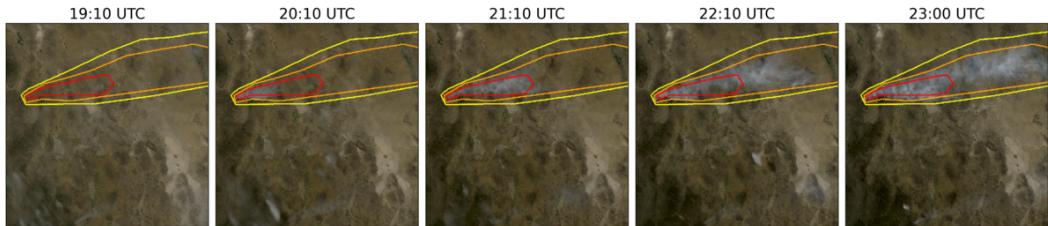


Figure 2: True color GOES-East imagery from May 5th, 2022, Southeast New Mexico (31.38°N, 107.87°W) during the start of the Foster Fire. The red, orange and yellow lines represent the heavy, medium and low density HMS smoke annotations that span 19:10–23:00 UTC.

145 We select either GOES-East or GOES-West based on the solar zenith angle (SZA) to optimize for
146 forward Mie scattering. For smoke identification, this approach can achieve a much higher SNR than
147 imaging the earth’s surface from an arbitrary angle. While the atmosphere is composed primarily of
148 molecules with size $< 1\text{nm}$, smoke particles are larger in comparison, varying from $100\text{ nm} - 10\text{ }\mu\text{m}$
149 in diameter, d . When the λ of light $\lambda < d$, the elastic scattering of light against matter is modeled
150 through Mie rather than Rayleigh scattering (Figure 3). Therefore, smoke observation favors forward
151 scattering and short λ so that we select (1) the satellite that would observe the most Mie forward
152 scattering as demonstrated in Figures 4(a) vs 4(b) that show GOES-East displaying a higher SNR

153 near sunset compared to GOES-West. (2) the three shortest λ ABI bands (C01-C03: 0.47, 0.64, and
 154 0.865 μm) that give the highest smoke SNR as seen in Figures 4(c)-4(e).

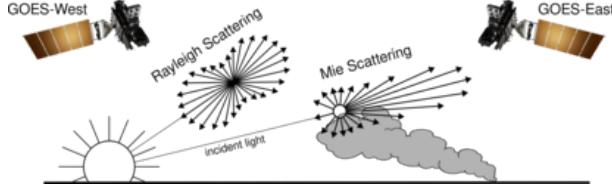


Figure 3: If the particle size is $< \frac{1}{10}$ the λ of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than the λ of light. This schematic demonstrates that when the sun is setting in the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

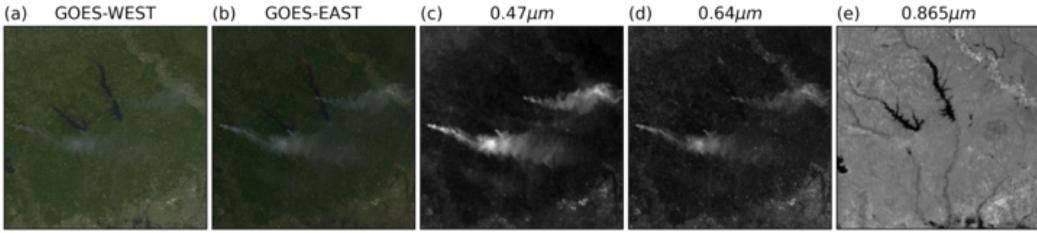


Figure 4: True color (a) GOES-WEST and (b) GOES-EAST imagery from March 23rd, 2022 centered at (31.1°, -93.8°) in Texas, USA taken at 23:20 UTC. The GOES-EAST raw band imagery for (c) blue, (d) red and (e) veggie bands show variations in the SNR for smoke detection in relation to the λ of light being measured.

155 For the existing dataset of HMS smoke plume annotations and all corresponding satellite imagery,
 156 $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, for each sample, a label $y_i \in \mathcal{Y}$ has a set of satellite imagery $[x_{(i,t_0)}, \dots, x_{(i,t_N)}] \in \mathcal{X}$
 157 over the analyst defined time window t . To develop a one-to-one data-to-label dataset, we generate
 158 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
 159 we aim to find the satellite image that has the maximum overlap between the geolocation of smoke in
 160 the imagery and the analyst annotation using the PLDR method shown in Figure 5.

161 To train an initial parent model, f_o , we developed a method of leveraging Mie scattering physics
 162 to select $x_{(i,t_M)} \in \mathcal{X}$ (Figure 5(b)) so that $x_{(i,t_M)}$ has a higher chance than random selection out
 163 of the set of imagery $[x_{(i,t_0)}, \dots, x_{(i,t_N)}]$ to be representative of y_i . This Mie-Derived Dataset, \mathcal{D}_M
 164 produces a training set so that if there is smoke present during the entire time window, the selected
 165 timestamp t_M would give the highest smoke SNR.

166 More importantly than finding the timestamp for maximum the SNR, we want to determine which
 167 image actually has smoke present within the smoke label boundaries. We use \mathcal{D}_M to train f_o (Figure
 168 5(d)), to identify smoke in satellite imagery, and then use that f_o to create IPLs of each satellite image
 169 in a given annotation's time-window (Figure 5(e)). From those results, the optimal satellite image
 170 is chosen based on which image's IPLs has the greatest overlap with the analyst annotation (Figure
 171 5(f)-(h)).

172 3.1.1 From Full Dataset \mathcal{D} to Mie-Derived Dataset \mathcal{D}_M

173 Let $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ be the original dataset, where each label $y_i \in \mathcal{Y}$ corresponds to multiple satellite
 174 images $[x_{(i,t_0)}, \dots, x_{(i,t_N)}] \in \mathcal{X}$ over a given time window. Using Mie scattering principles, we select
 175 the image $x_{(i,t_M)}$ with the highest expected smoke SNR to form a one-to-one dataset $\mathcal{D}_M = \{\mathcal{X}_M, \mathcal{Y}\}$
 176 such that $\mathcal{X}_M \subset \mathcal{X}$ and $|\mathcal{X}_M| = |\mathcal{Y}|$. The Mie-derived satellite image is chosen based on which
 177 frame within the annotation time window would exhibit the strongest forward scattering geometry
 178 and thus the highest potential smoke SNR if smoke were present.

179 Based on forward scattering criteria, the trivial strategy would be to pull imagery from GOES-West
 180 right after sunrise and from GOES-East right before sunset when the SZA is closest to 90°. However,

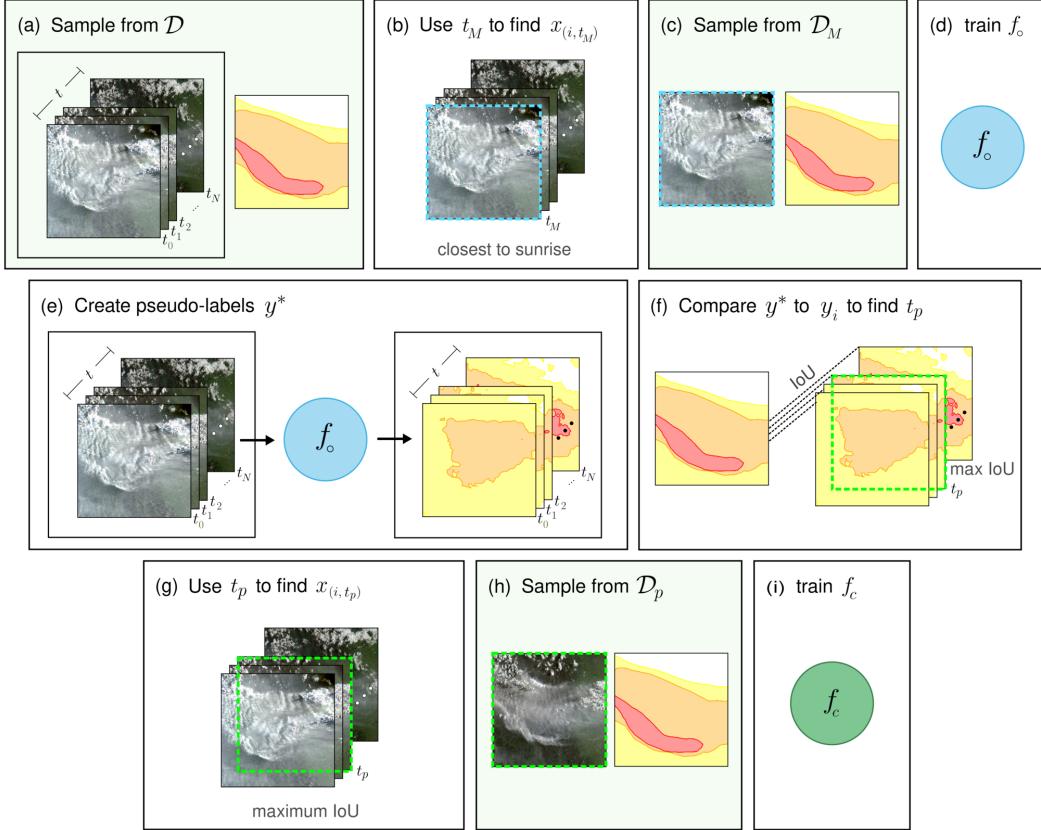


Figure 5: PLDR applied to create the SmokeViz dataset. Green boxes indicate dataset stages. (a) For original dataset \mathcal{D} - analyst annotation y_i corresponds to N satellite images across time window t so that $([x_{(i,t_0)}, \dots, x_{(i,t_N)}], y_i) \in \mathcal{D}$; (b) use Mie scattering to find the time, t_M , that corresponds with satellite image $x_{(i,t_M)}$ that would produce the highest possible SNR if smoke was present; (c) resulting \mathcal{D}_M is one-to-one $(x_{(i,t_M)}, y_i) \in \mathcal{D}_M$; (d) parent model f_o is trained; (d) parent model f_o is trained on \mathcal{D}_M such that $f_o(x_{(i,t_M)}) = y_i$; (e) apply a greedy algorithm $f_o([x_{(i,t_0)}, \dots, x_{(i,t_N)}]) = [y_{(i,t_0)}^*, \dots, y_{(i,t_N)}^*]$ to create IPLs y^* for each candidate image; (f) compute the intersection over union (IoU) between y^* and y_i to identify the time t_p where the IPL and analyst annotation have the maximum IoU; (g) best match image $x_{(i,t_p)}$ selected; (g) match t_p to its corresponding image $x_{(i,t_p)}$ that should best match the analyst annotation; (h) SmokeViz dataset \mathcal{D}_p created; (i) child model f_c is trained on \mathcal{D}_p such that $f_c(x_{(i,t_p)}) = y_i$ is used to detect and classify the density of wildfire smoke plumes in GOES imagery.

181 the time when the SZA approaches 90° coincides with when there are maximized atmospheric
 182 interactions that cause an increase in noise and artifacts [22]. To avoid image artifacts caused by
 183 extreme SZA, we exclude scenes with $SZA > 88^\circ$ if there are lower SZA images available. The
 184 resulting dataset \mathcal{D}_M contains over 200,000 samples and is used to train the parent model f_o .

185 3.1.2 PLDR Dataset \mathcal{D}_p

186 We train f_o on \mathcal{D}_M (table 3) to find smoke in GOES imagery. The model is then applied across each
 187 satellite image within the time window $[x_{(i,t_0)}, \dots, x_{(i,t_N)}]$ to produce intermediary pseudo-labels
 188 (IPLs), $[y_{(i,t_0)}^*, \dots, y_{(i,t_N)}^*]$. We compute the intersection over union (IoU) between each y^* and the
 189 corresponding analyst label y_i , selecting the timestamp t_p that yields the maximum IoU to define
 190 $\mathcal{D}_p = \{x_{(i,t_p)}, y_i\}$.

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

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

191 To build f_o , we implement Segmentation Models PyTorch [12] with EfficientNetV2 [26] as the
 192 encoder and PSPNet [28] as the decoder. Input images are $256 \times 256 \times 3$ true-color snapshots;
 193 the output is a $256 \times 256 \times 3$ classification map predicting categorical smoke density. We use
 194 thermometer encoding to capture the ordinal nature of smoke density classes (Table 2) and apply
 195 binary cross-entropy loss across the classes. We use a confidence threshold of $\text{IoU} > 0.01$ [9] to
 196 exclude samples with negligible overlap.

197 The same dataset split choices (table 3) and model setup that were used for \mathcal{D}_M and f_o were
 198 implemented for to train a child model f_c on \mathcal{D}_p . To assess if training with \mathcal{D}_p can produce a more
 199 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
 200 sets.

201 3.2 Benchmark Models

202 We benchmark the SmokeViz dataset \mathcal{D}_p by varying the semantic segmentation architectures. We
 203 train Linknet [7], PSPNet [28] and MANet [8] using the same encoder for f_c and f_o , EfficientNetV2.
 204 Each model is trained over 100 epochs using a batch size of 32 and the Adam optimizer on 8 16GB
 205 memory Nvidia P100 GPUs over 12 hours of allotted training time. We choose these architectures
 206 because of their abilities to capture multi-scale objects such the varying spatial extents of smoke
 207 plumes.

208 4 Results

209 To interpret the performance of f_o , we report the IoU metrics in table 4 that were computed by
 210 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
 211 pixels that f_o predicts as that density of smoke and the entire set of pixels labeled by the analyst
 212 as a particular smoke density over all imagery contained in the testing dataset. Additionally, we
 213 compute the overall IoU for all densities by first computing the number of pixels that intersect their
 214 corresponding density and divide that by the total number of pixels that make up the union of model
 215 predicted and analyst labeled smoke in the testing dataset.

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

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
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

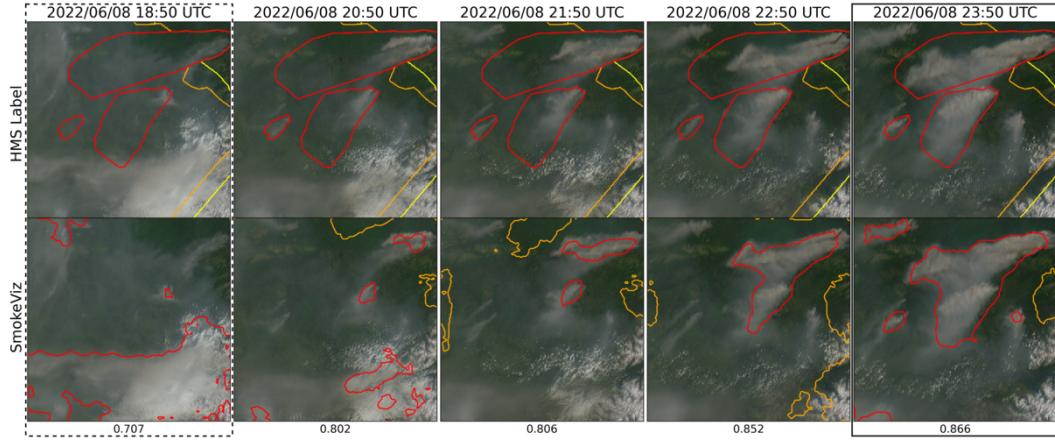


Figure 6: GOES-West imagery showing smoke on June 8th, 2022 in Alaska where, at this geolocation (61.06°N , 156.12°W), daylight was between 12:43-7:53 UTC. The HMS smoke annotations (top row) span from 18:50 to 23:50 UTC and are compared to the f_o generated pseudo-labels (bottom row). The first column (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.

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

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

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

229 5 Limitations

230 One of the concerns that comes with using pseudo-labeling methods is that you can perpetuate biases
 231 from the parent model into subsequent child models. Due to the increase in detectable forward
 232 scattered light off smoke particular matter, we expect the model to have a bias towards producing a
 233 higher success rate for smoke detection at larger solar zenith angles. The original HMS annotations
 234 do not distinguish by type of fire and include a large representation of controlled agricultural burns.

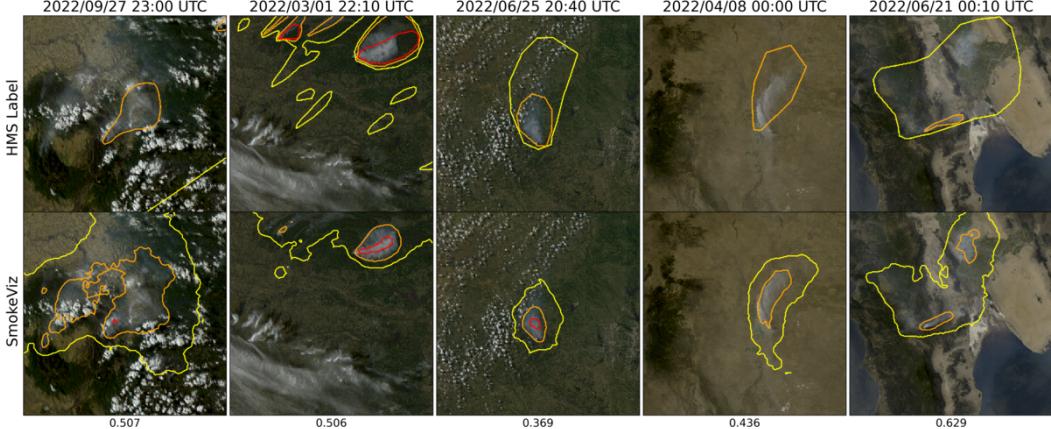


Figure 7: Examples of HMS annotations (top row) vs f_c output (bottom row) on \mathcal{D}_p samples. The overall IoU score is reported at the bottom of each column.

235 This can be a limitation to consider if the dataset is being trained to target detection of large wildfires.
 236 All these limitations are discussed and analyzed further in the Supplementary Materials. Additional
 237 work should be done to analyze the performance of SmokeViz derived models on dust vs smoke.

238 6 Conclusion

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

255 To improve the dataset’s capacity to accurately represent wildfire smoke plumes, we train a parent
 256 machine learning model, f_o , using the Mie-derived dataset, \mathcal{D}_M , and run it on the relevant satellite
 257 images within the time-frame. The image with the maximum IoU score between the model’s smoke
 258 predictions, or pseudo-label, and the analyst smoke annotations are used to create the pseudo-label
 259 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
 260 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
 261 performing model, f_c , that gave higher IoU metrics on both dataset’s testing sets in comparison to
 262 the original parent model, f_o .

263 The result of this study is a representative dataset, SmokeViz, that can be used to train machine
 264 learning models for various wildfire smoke applications. A future goal is to produce a robust
 265 and reliable machine learning based approach for detecting wildfires using satellite imagery. That
 266 information can be used for wildfire detection and monitoring in along with a highly needed smoke
 267 product for data assimilation into smoke dispersion models. Additionally, this dataset can be used as
 268 a benchmark for how well remote sensing segmentation models can perform on dispersed edges such
 269 as smoke. On a broader scale, we show how pseudo-labeling can be used to optimize a dataset when
 270 the resolution for the data and corresponding labels do not match. This could be useful in similar

271 applications involving time-series/video data with a singular label where the data can be compressed
272 while still remaining representative of the label. All data is made publicly available at [aws download
273 link] and all code can be found at <https://github.com/anonymous-smokeviz/SmokeViz>.

274 References

- 275 [1] R. Ahmadov, H. Li, J. Romero-Alvarez, J. Schnell, S. Bhimireddy, E. James, K. Y. Wong,
276 M. Hu, J. Carley, P. Bhattacharjee, et al. Forecasting smoke and dust in noaa's next-generation
277 high-resolution coupled numerical weather prediction model. Technical report, Copernicus
278 Meetings, 2024.
- 279 [2] J. E. Aldy, M. Auffhammer, M. Cropper, A. Fraas, and R. Morgenstern. Looking back at 50
280 years of the clean air act. *Journal of Economic Literature*, 60(1):179–232, 2022.
- 281 [3] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo. Smokenet: Satellite smoke scene detection using
282 convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 11(14):
283 1702, 2019.
- 284 [4] M. Bah, M. Gunshor, and T. Schmit. Generation of goes-16 true color imagery without a green
285 band. *Earth and Space Science*, 5(9):549–558, 2018.
- 286 [5] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi. Satlaspretrain: A large-scale
287 dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International
288 Conference on Computer Vision*, pages 16772–16782, 2023.
- 289 [6] M. Burke, A. Driscoll, S. Heft-Neal, J. Xue, J. Burney, and M. Wara. The changing risk and
290 burden of wildfire in the united states. *Proceedings of the National Academy of Sciences*, 118
291 (2):e2011048118, 2021.
- 292 [7] A. Chaurasia and E. Culurciello. Linknet: Exploiting encoder representations for efficient
293 semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*,
294 pages 1–4. IEEE, 2017.
- 295 [8] T. Fan, G. Wang, Y. Li, and H. Wang. Ma-net: A multi-scale attention network for liver and
296 tumor segmentation. *IEEE Access*, 8:179656–179665, 2020.
- 297 [9] R. E. Ferreira, Y. J. Lee, and J. R. Dórea. Using pseudo-labeling to improve performance of
298 deep neural networks for animal identification. *Scientific Reports*, 13(1):13875, 2023.
- 299 [10] E. Gakidou, A. Afshin, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah,
300 A. M. Abdulle, S. F. Abera, V. Aboyans, et al. Global, regional, and national comparative risk
301 assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters
302 of risks, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The
303 Lancet*, 390(10100):1345–1422, 2017.
- 304 [11] S. J. Goodman, T. J. Schmit, J. Daniels, and R. J. Redmon. *The GOES-R series: a new
305 generation of geostationary environmental satellites*. Elsevier, 2019.
- 306 [12] P. Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.
- 307 [13] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes,
308 G. Nyirjesy, B. Edwards, et al. Foundation models for generalist geospatial artificial intelligence.
309 arxiv 2023. *arXiv preprint arXiv:2310.18660*.
- 310 [14] E. James, R. Ahmadov, and G. A. Grell. Realtime wildfire smoke prediction in the united states:
311 The hrrr-smoke model. In *EGU General Assembly Conference Abstracts*, page 19526, 2018.
- 312 [15] A. Kitamoto, J. Hwang, B. Vuillod, L. Gautier, Y. Tian, and T. Klanuwat. Digital typhoon: Long-
313 term satellite image dataset for the spatio-temporal modeling of tropical cyclones. *Advances in
314 Neural Information Processing Systems*, 36, 2024.

- 316 [16] A. Larsen, I. Hanigan, B. J. Reich, Y. Qin, M. Cope, G. Morgan, and A. G. Rappold. A deep
317 learning approach to identify smoke plumes in satellite imagery in near-real time for health risk
318 communication. *Journal of exposure science & environmental epidemiology*, 31(1):170–176,
319 2021.
- 320 [17] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep
321 neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07
322 2013.
- 323 [18] D. McNamara, G. Stephens, M. Rumsby, and T. Kasheta. The hazard mapping system (hms) -
324 noaa's multi-sensor fire and smoke detection program using environmental satellites. *Conference
325 on Satellite Meteorology and Oceanography*, 01 2004.
- 326 [19] NOAA. Hazard mapping system fire and smoke product. URL <https://www.ospo.noaa.gov/Products/land/hms.html#about>.
- 327 [20] T. Randriambelo, S. Baldy, M. Bessafi, M. Petit, and M. Despinoy. An improved detection
328 and characterization of active fires and smoke plumes in south-eastern africa and madagascar.
329 *International Journal of Remote Sensing*, 19(14):2623–2638, 1998.
- 330 [21] M. Raspaud, D. Hoesel, A. Dybbroe, P. Lahtinen, A. Devasthale, M. Itkin, U. Hamann, L. Ø.
331 Rasmussen, E. S. Nielsen, T. Leppelt, et al. Pytroll: An open-source, community-driven python
332 framework to process earth observation satellite data. *Bulletin of the American Meteorological
333 Society*, 99(7):1329–1336, 2018.
- 334 [22] A. Royer, P. Vincent, and F. Bonn. Evaluation and correction of viewing angle effects on
335 satellite measurements of bidirectional reflectance. *Photogrammetric engineering and remote
336 sensing*, 51(12):1899–1914, 1985.
- 337 [23] W. Schroeder, M. Rumsby, I. Csizsar, L. Giglio, E. Prins, C. Schmidt, and J. Morisette.
338 Validation analyses of an operational fire monitoring product: The hazard mapping system.
339 *International Journal of Remote Sensing*, 29(20):6059–6066, 2008.
- 340 [24] B. Stevens, S. Bony, H. Brogniez, L. Hentgen, C. Hohenegger, C. Kiemle, T. S. L'Ecuyer, A. K.
341 Naumann, H. Schulz, P. A. Siebesma, et al. Sugar, gravel, fish and flowers: Mesoscale cloud
342 patterns in the trade winds. *Quarterly Journal of the Royal Meteorological Society*, 146(726):
343 141–152, 2020.
- 344 [25] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. Bigearthnet: A large-scale benchmark
345 archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international
346 geoscience and remote sensing symposium*, pages 5901–5904. IEEE, 2019.
- 347 [26] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *International
348 conference on machine learning*, pages 10096–10106. PMLR, 2021.
- 349 [27] J. Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery.
350 ArXiv, abs/2109.01637, 2021. URL [https://api.semanticscholar.org/CorpusID:
351 237416777](https://api.semanticscholar.org/CorpusID:237416777).
- 352 [28] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of
353 the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- 354 [29] T. X.-P. Zhao, S. Ackerman, and W. Guo. Dust and smoke detection for multi-channel imagers.
355 *Remote Sensing*, 2(10):2347–2368, 2010. ISSN 2072-4292. doi: 10.3390/rs2102347. URL
356 <https://www.mdpi.com/2072-4292/2/10/2347>.
- 357 [30] S. Zorzi, S. Bazrafkan, S. Habenschuss, and F. Fraundorfer. Polyworld: Polygonal building
358 extraction with graph neural networks in satellite images. In *Proceedings of the IEEE/CVF
359 Conference on Computer Vision and Pattern Recognition*, pages 1848–1857, 2022.
- 360