

---

# 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  
18 to develop further wildfire-related machine learning models and is publicly avail-  
19 able at <https://noaa-gsl-experimental-pds.s3.amazonaws.com/index.html#SmokeViz/>.  
20

21 

## 1 Introduction

22 Due in part to public policy, average fine particulate matter ( $PM_{2.5}$ ) levels in the United States have  
23 declined over recent decades [1]. However, from 2010 to 2020, the contribution of wildfire smoke to  
24  $PM_{2.5}$  concentrations more than doubled, accounting for up to half of total  $PM_{2.5}$  exposure in Western  
25 U.S. [2]. This is particularly concerning, as ambient  $PM_{2.5}$  is a leading environmental risk factor for  
26 adverse health outcomes and premature mortality [3]. These trends/risks highlight the urgent need  
27 for scalable and timely smoke monitoring systems to mitigate public health risks.

28 Satellite imagery offers the spatial coverage and temporal frequency needed for large-scale smoke  
29 monitoring. In comparison to polar-orbiting satellites like Suomi or Sentinel, geostationary satellites  
30 such as the GOES series [4] are especially well-suited to this task, providing persistent observation  
31 over fixed regions, essential for capturing the dynamic behavior of wildfire smoke plumes. The  
32 high temporal resolution and wide coverage of GOES imagery enable real-time tracking of smoke  
33 concentration and movement, supporting air quality assessments and early warning systems.

34 Even with the advances in remote sensing, existing deep learning satellite datasets for wildfire smoke  
35 detection face several limitations. They are often small in scale, restricted to specific regions or events,  
36 and focus on scene-level classification rather than pixel-level segmentation. Most do not differentiate

37 between smoke density levels, are not publicly available, and lack standardized benchmarks for  
38 semantic segmentation. While NOAA’s Hazard Mapping System (HMS) provides a large-scale,  
39 expert-labeled dataset, its annotations span multi-hour time windows that vary in duration. This  
40 creates a temporal mismatch between the labels and individual satellite frames, complicating their  
41 direct use for supervised learning.

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

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

53 The contributions presented in this paper include **SmokeViz**, the largest satellite-based dataset for  
54 wildfire smoke segmentation, with over 180,000 samples from GOES imagery, our proposed **PLDR**,  
55 a physics-guided semi-supervised method for aligning coarse human annotations with temporally  
56 optimal satellite imagery and benchmark segmentation baselines with standardized training splits to  
57 support reproducibility and future studies.

## 58 2 Related Work

### 59 2.1 Smoke Detection and Labeling Methods

60 Multi-channel thresholding remains a widely used method for distinguishing smoke from similar  
61 atmospheric signatures such as dust or clouds using channel-specific radiance values [5]. These  
62 thresholds are typically derived from labeled historical data and are fine-tuned to specific regions and  
63 fuel types, limiting their generalizability [6]. In contrast, the SmokeViz dataset spans a wide range of  
64 biogeographies across North America and can serve as a source of refined analyst-labeled examples  
65 for developing more generalizable thresholding techniques.

66 Large parameterized numerical models are used for forecasting smoke dispersion, but not for smoke  
67 detection itself. Systems such as HRRR-Smoke and RRFS [7, 8] rely on computationally intensive  
68 forecasts requiring nearly 200 dynamic meteorological inputs. A key limitation of these models is  
69 the absence of a real-time smoke analysis product for data assimilation, resulting in delayed model  
70 spin-up and compounded forecast errors. Model predictions from SmokeViz could help fill this gap,  
71 offering a real-time, satellite-driven alternative to support data assimilation for operational smoke  
72 dispersion forecasting.

73 Manual smoke labeling is performed by trained analysts through visual inspection of satellite imagery.  
74 NOAA’s Hazard Mapping System (HMS) provides a analyst-labeled wildfire smoke dataset [9, 10].  
75 HMS analysts examine GOES imagery sequences to track smoke plume movement and annotate the  
76 approximate spatial extent and qualitative density of smoke (light, medium, heavy), as illustrated  
77 in Figure 2.1. Annotations are issued on a rolling basis and span time windows ranging from  
78 instantaneous to over 20 hours [11]. While HMS provides high-quality expert annotations, its  
79 operational format introduces challenges for supervised learning: annotations are temporally coarse,  
80 vary in length, and lack one-to-one correspondence with satellite frames. SmokeViz refines HMS  
81 annotations into temporally resolved, frame-aligned labels, enabling real-time, continuous predictions  
82 of smoke extent and density.

### 83 2.2 Deep Learning Datasets and Models for Wildfire Smoke

84 Recent efforts have applied deep learning to wildfire smoke detection using a variety of satellite  
85 sources and label strategies. SmokeNet [12] employs a convolutional neural network (CNN) to classify  
86 MODIS image scenes as containing smoke or not, using student-provided labels. SatlasPretrain [13]  
87 includes a small set of Sentinel-2 images labeled for smoke as part of a larger multi-label pretraining

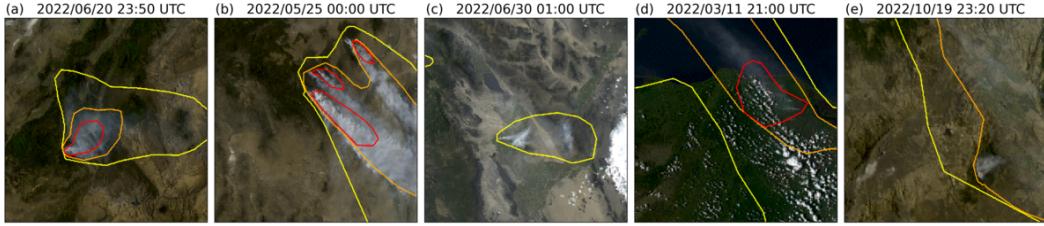


Figure 1: HMS smoke annotations overlaid on GOES imagery. Yellow, orange, and red contours indicate light, medium, and heavy smoke density, respectively. (a) and (b) show canonical smoke plumes; (c)–(e) illustrate density label variation across scenes.

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 limited in scope. Wen et al. [14] trained a CNN on GOES-East imagery over California and Nevada using HMS annotations from the 2018 wildfire season. Larsen et al. [15] 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. Not only do we use geostationary satellites with persistent observations, but we choose either GOES-East or GOES-West based on which satellite has optimal observational conditions of the event. 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 (polar orbiting (P) or geostationary (G)), 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. |
|-----------|-----------|--------------------|---------|---------------|------|--------|
| [12]      | 1016      | MODIS (P)          | 5       | students      | SC   | no     |
| [13]      | 125       | Sentinel-2 (P)     | 3       | crowd sourced | SC   | yes    |
| [14]      | 4095      | GOES-East (G)      | 5       | HMS analysts  | SS   | no     |
| [15]      | 975       | Himiwari-8 (G)     | 7       | algorithm     | SS   | no     |
| SmokeViz  | 183,672   | GOES-East+West (G) | 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 [16, 17] or focus on sharply bounded features such as cropland [17], infrastructure [18], or oceanic clouds [19, 20], smoke has amorphous, fading boundaries in both space and time. Incorporating smoke segmentation into large-scale pretraining corpora, such as SatlasPretrain [13], could enhance generalizable models for Earth observation.

### 2.3 Pseudo-labeling and Semi-Supervised Learning

Semi-supervised learning techniques such as pseudo-labeling have been widely used to expand training data by leveraging unlabeled samples [21]. Typically, a parent model is trained on labeled data and then used to generate pseudo-labels for an unlabeled dataset, which are in turn used to train subsequent models in an iterative process.

In contrast, we propose a non-iterative variation focused not on data expansion, but dataset data-to-label precision. Our method, **pseudo-label dimension reduction (PLDR)**, generates intermediary pseudo-labels (IPLs) for each satellite frame within the HMS annotation window. Rather than using these labels for training, we use them to identify the satellite image with the greatest alignment to the analyst annotation. This enables the construction of SmokeViz, a temporally disambiguated, one-to-one image-to-label dataset. The resulting dataset methodically pairs the analyst-generated

117 smoke plume labels with selected GOES imagery, enabling high-resolution, temporally accurate  
118 segmentation model training.

119 Beyond wildfire smoke segmentation, PLDR offers a general framework for aligning coarse or  
120 weakly matched datasets. This is particularly useful in domains such as remote sensing, medical  
121 imaging, and video analysis, where annotations often span temporal or spatial intervals rather than  
122 individual frames. In Earth observation specifically, atmospheric parameters are often combined  
123 from disparate sources with inconsistent spatial and temporal resolutions, making it difficult to  
124 integrate them into unified training datasets. By using intermediary pseudo-labels to identify the  
125 most representative input sample, PLDR transforms many-to-one or one-to-many supervision into  
126 clean one-to-one mappings. This enables more precise alignment between data and labels, facilitating  
127 integration across heterogeneous sources without requiring additional hand-labeling. As presented,  
128 PLDR serves as a practical preprocessing strategy for repurposing historical legacy datasets with  
129 temporal ambiguity into precise training resources for modern deep learning models.

### 130 3 Methods

#### 131 3.1 Datasets

132 We use imagery from the latest GOES satellites—GOES-16 (East), GOES-17, and GOES-18 (West),  
133 each equipped with the Advanced Baseline Imager (ABI), which captures 16 spectral bands from  
134 visible to infrared wavelengths every 10 minutes. We process bands 1-3 using PyTroll [22] to generate  
135 1km true-color composites [23], matching the imagery reviewed by HMS analysts. These bands  
136 correspond to the shortest wavelengths available on ABI and yield the highest signal-to-noise ratio  
137 (SNR).

138 To approximate the dynamic movement of smoke, HMS analysts annotate plumes using multi-frame  
139 satellite animations. These annotations span varying time windows, averaging three hours. Since the  
140 HMS annotations are designed to reflect overall plume extent during a time window rather than at  
141 any specific moment, smoke boundaries in individual frames may not align well with the annotation  
142 (Figure 2). A naive modeling approach would use all frames within each time window as input, but  
143 this introduces non-uniform sequence lengths and significantly increases memory and computational  
144 demands and complicates the use of CNN architectures. Instead, we establish a one-to-one mapping  
145 by identifying the single satellite frame that best matches each analyst annotation.

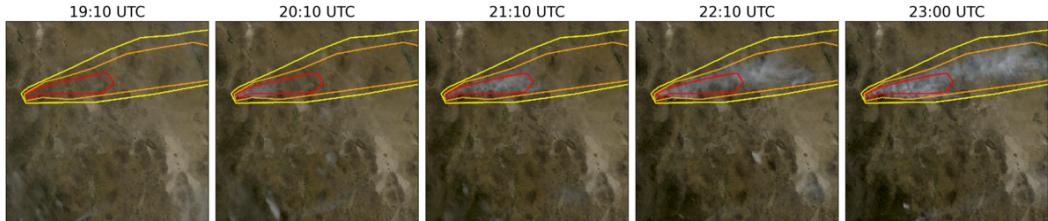


Figure 2: True color GOES-East imagery from May 5th, 2022, Southeast New Mexico ( $31.38^{\circ}\text{N}$ ,  $107.87^{\circ}\text{W}$ ) during the start of the Foster Fire. The red, orange and yellow lines represent the heavy, medium and low density HMS smoke annotations that span 19:10–23:00 UTC.

146 We select either GOES-East or GOES-West based on the solar zenith angle (SZA) to optimize  
147 for forward Mie scattering, which enhances smoke visibility in satellite imagery. Smoke particles  
148 ( $100\text{nm}-10\mu\text{m}$ ) scatter light predominantly via Mie scattering when  $\lambda < d$ , favoring short wavelengths  
149 and forward angles (Figure 3). To generate the Mie-derived dataset, we evaluate the available satellite  
150 platforms for each annotation time window and choose the satellite (East or West) that is expected  
151 to observe the strongest forward scattering geometry based on sun-satellite alignment. This ensures  
152 selection of the satellite view with the highest potential smoke SNR if smoke were present. Therefore,  
153 we select (1) the satellite expected to yield the strongest Mie forward scattering (Figures 4(a) vs  
154 4(b)) and (2) the three shortest wavelength ABI bands (C01-C03: 0.47, 0.64, and  $0.865\mu\text{m}$ ) (Figures  
155 4(c)-4(e)).

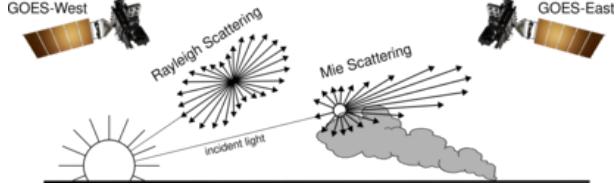


Figure 3: If the particle size is  $< \frac{1}{10}$  the  $\lambda$  of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than the  $\lambda$  of light. This schematic demonstrates that when the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

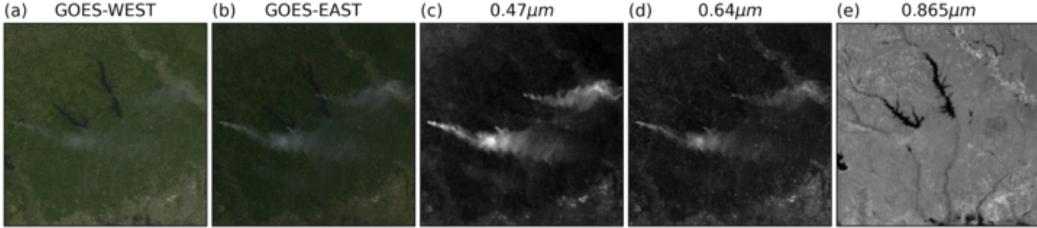


Figure 4: True color (a) GOES-WEST and (b) GOES-EAST imagery from March 23<sup>rd</sup>, 2022 centered at  $(31.1^\circ, -93.8^\circ)$  in Texas, USA taken at 23:20 UTC. The GOES-EAST raw band imagery for (c) blue, (d) red and (e) veggie bands show variations in the SNR for smoke detection in relation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

### 156 3.1.1 From Full Dataset $\mathcal{D}$ to Mie-Derived Dataset $\mathcal{D}_M$

157 Let  $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$  be the original dataset, where each label  $y_i \in \mathcal{Y}$  corresponds to multiple satellite  
 158 images  $[x_{(i,t_0)}, \dots, x_{(i,t_N)}] \in \mathcal{X}$  over a given time window. Using Mie scattering principles, we select  
 159 the image  $x_{(i,t_M)}$  with the highest expected smoke SNR to form a one-to-one dataset  $\mathcal{D}_M = \{\mathcal{X}_M, \mathcal{Y}\}$   
 160 such that  $\mathcal{X}_M \subset \mathcal{X}$  and  $|\mathcal{X}_M| = |\mathcal{Y}|$ . Based on forward scattering criteria, the trivial strategy would  
 161 be to pull imagery from GOES-West right after sunrise and from GOES-East right before sunset  
 162 when the SZA is closest to  $90^\circ$ . To avoid image artifacts caused by extreme SZA, we exclude scenes  
 163 with  $SZA > 88^\circ$  [24]. The resulting dataset  $\mathcal{D}_M$  (Table 3) contains over 200,000 samples where the  
 164 satellite image is chosen based on which frame within the annotation time window would exhibit  
 165 the strongest forward scattering geometry and thus the highest potential smoke SNR if smoke were  
 166 present.

### 167 3.1.2 PLDR Dataset $\mathcal{D}_p$

168 The  $\mathcal{D}_M$  data selection process introduces a potential bias for resulting models to limit smoke  
 169 identification to higher SZAs. Additionally,  $\mathcal{D}_M$  is limited to providing the timestamp for maximum  
 170 possible smoke SNR, it does not give information to point to which image aligns best with the  
 171 smoke label. To address these limitations, we propose using  $\mathcal{D}_M$  as a intermediary dataset in the  
 172 PLDR workflow (Figure 5) that will predict the satellite image that best matches the analyst's smoke  
 173 annotation to produce  $\mathcal{D}_p$ .

174 To build  $f_o$ , we implement Segmentation Models PyTorch [25] with EfficientNetV2 [26] as the  
 175 encoder and PSPNet [27] as the decoder. Input images are  $256 \times 256 \times 3$  true-color snapshots; the  
 176 output is a  $256 \times 256 \times 3$  classification map predicting categorical smoke density. We use thermometer  
 177 encoding (Table 2) and apply binary cross-entropy loss across density levels. Thermometer encoding  
 178 is chosen over one-hot encoding because it captures the ordinal structure of smoke density categories  
 179 (none < light < medium < heavy). In thermometer encoding, each higher class includes all lower  
 180 class activations (e.g., heavy = [1 1 1]), allowing the model to learn not just class distinctions, but the  
 181 relative severity of smoke. We use a confidence threshold of IoU > 0.01 [28] to exclude samples with  
 182 negligible overlap.

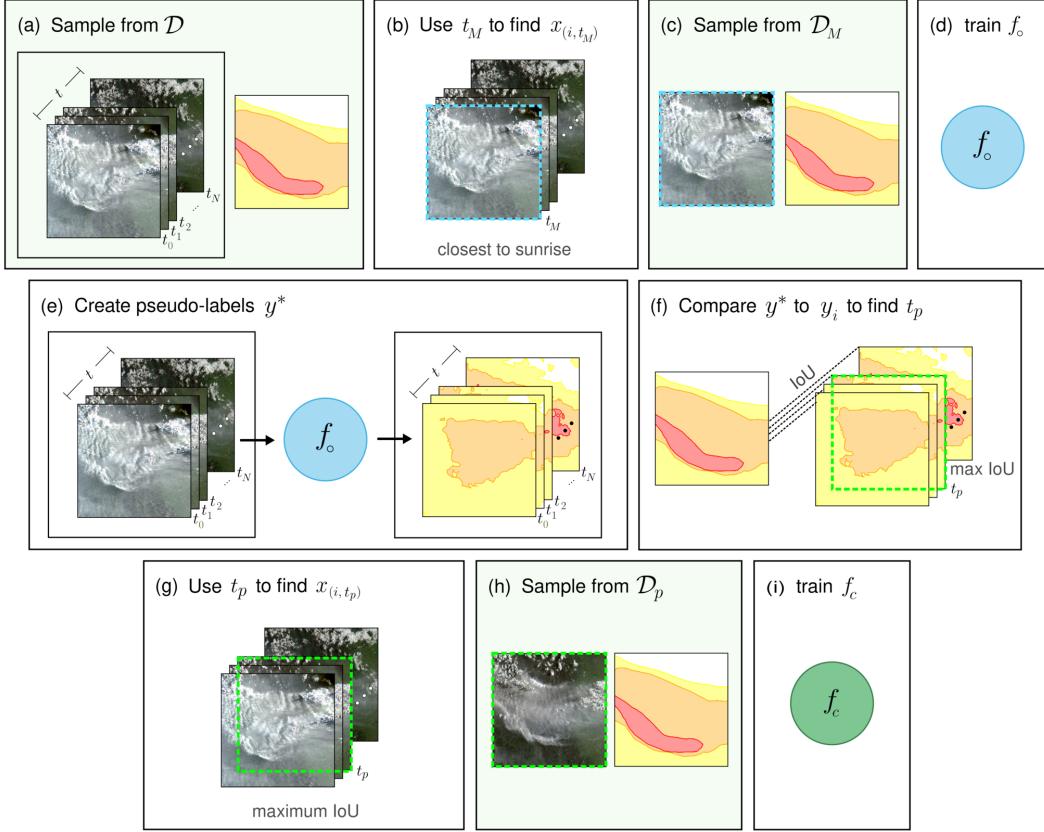


Figure 5: PLDR applied to create the SmokeViz dataset. Green boxes indicate dataset stages. (a) For original dataset  $\mathcal{D}$  - analyst annotation  $y_i$  corresponds to  $N$  satellite images across time window  $t$  so that  $([x_{(i,t_0)}, \dots, x_{(i,t_N)}], y_i) \in \mathcal{D}$ ; (b) use 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 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) match  $t_p$  to its corresponding image  $x_{(i,t_p)}$  that is predicted to 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.

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. 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         | 2018-21, 24 |
| validation | 20,056          | 19,223          | 2023        |
| testing    | 21,541          | 20,224          | 2022        |

### 183 3.2 Benchmark Models

184 We benchmark the SmokeViz dataset  $\mathcal{D}_p$  using DeepLabV3+ [29] and PSPNet [27] with Efficient-  
185 NetV2 [26], DPT [30] with ViT [31], Segformer [32] and UperNet [33] with EfficientVit [34]. Each

186 model is trained for 100 epochs using a batch size of 16 and the Adam optimizer on 8 16GB Nvidia  
 187 P100 GPUs. These architectures are selected for their relatively low memory requirements and  
 188 effectiveness in segmenting multi-scale objects such as smoke plumes.

## 189 4 Results

190 We evaluate the performance of  $f_o$  and  $f_c$  using Intersection over Union (IoU) metrics on the test  
 191 sets of both  $\mathcal{D}_M$  and  $\mathcal{D}_p$ , as shown in Table 4. For each smoke density class, IoU is calculated as the  
 192 pixel-level intersection between model predictions and HMS analyst labels, divided by their union,  
 193 aggregated over all test samples. Overall IoU is computed by summing intersections across all density  
 194 classes and dividing by the total union of predicted and labeled smoke pixels.

Table 4: IoU results per smoke density and overall, comparing  $f_o$  and  $f_c$  run on  $\mathcal{D}_M$  and  $\mathcal{D}_p$  test sets.

|             | $f_o$           |                 | $f_c$           |                 |
|-------------|-----------------|-----------------|-----------------|-----------------|
|             | $\mathcal{D}_M$ | $\mathcal{D}_p$ | $\mathcal{D}_M$ | $\mathcal{D}_p$ |
| heavy IoU   | 0.1510          | 0.2179          | 0.1649          | 0.2604          |
| medium IoU  | 0.2572          | 0.3417          | 0.2786          | 0.3965          |
| light IoU   | 0.3933          | 0.5054          | 0.4395          | 0.5873          |
| overall IoU | 0.3483          | 0.4499          | 0.3898          | 0.5250          |
| precision   | 0.6990          | 0.7875          | 0.6942          | 0.7907          |
| recall      | 0.4098          | 0.5121          | 0.4706          | 0.6098          |

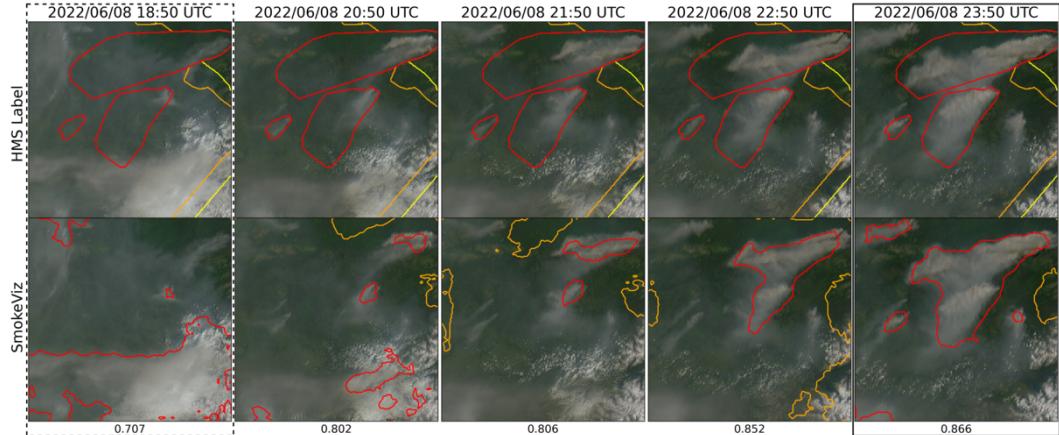


Figure 6: GOES-West imagery from June 8, 2022, over Alaska ( $61.06^{\circ}\text{N}$ ,  $156.12^{\circ}\text{W}$ ). Daylight spanned 12:43-7:53 UTC. The HMS annotations (top row) span 18:50-23:50 UTC and are compared with  $f_o$ -generated smoke predictions (bottom row). The leftmost frame (dotted) represents the Mie-derived image; the rightmost frame (solid) was selected via PLDR and achieves higher IoU.

195 Figure 6 illustrates a case in which the PLDR-selected frame better represents the HMS annotation  
 196 than the Mie-derived selection. Here, the heavy smoke IoU improves from 0.01 to 0.59. While the  
 197 Mie-derived image is selected based on its proximity to sunrise, PLDR chooses the frame with the  
 198 highest overlap between the model-generated intermediary pseudo-label and the analyst annotation.  
 199 This example highlights PLDR’s advantage in resolving temporal ambiguity.

200 To further examine the performance of  $f_c$ , we can qualitatively compare its predictions against HMS  
 201 annotations for samples from  $\mathcal{D}_p$  in Figure 7. The model outputs capture more spatially detailed and  
 202 coherent smoke boundaries compared to the coarser, polygon-based analyst labels.

203 To benchmark performance across segmentation architectures, we evaluate several encoder-decoder  
 204 models trained on  $\mathcal{D}_p$ . Table 5 reports IoU scores by smoke density and overall. While DeepLabV3+  
 205 achieves the highest IoU for heavy smoke, PSPNet yields the best overall performance. Results across

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

| encoder<br>decoder | EfficientNetV2 [26]<br>DeepLabV3+ [29] | [26]<br>PSPNet [27] | EfficientViT [34]<br>Segformer [32] | [34]<br>UperNet [33] | ViT [31]<br>DPT [30] |
|--------------------|--|---------------------|-------------------------------------|----------------------|----------------------|
| heavy              | 0.2766                                 | 0.2604              | 0.2351                              | 0.2374               | 0.2450               |
| medium             | 0.3854                                 | 0.3965              | 0.3707                              | 0.3444               | 0.3745               |
| light              | 0.5912                                 | 0.5873              | 0.5259                              | 0.5541               | 0.5803               |
| overall            | 0.5246                                 | 0.5250              | 0.4732                              | 0.4886               | 0.5136               |

206 models are relatively consistent, highlighting the robustness of the SmokeViz dataset for training  
207 diverse architectures.

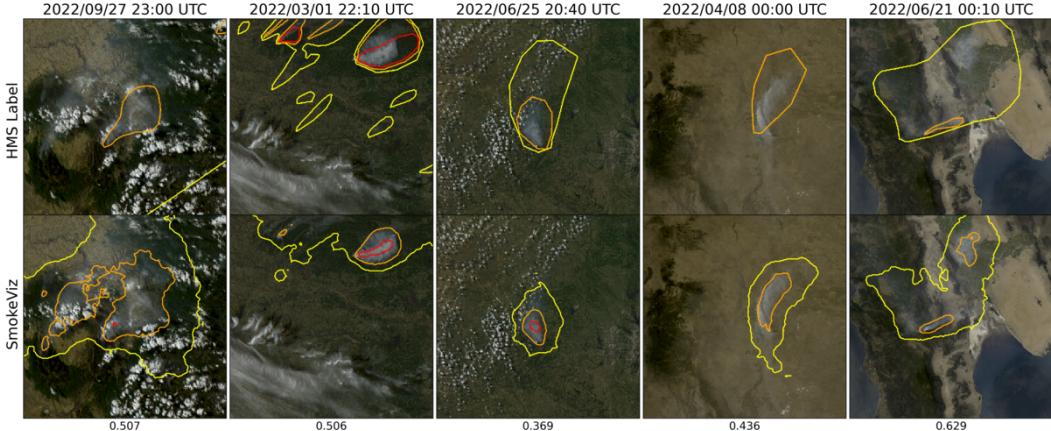


Figure 7: Examples of HMS annotations (top row) vs  $f_c$  output (bottom row) on  $\mathcal{D}_p$  samples. The overall IoU score is reported at the bottom of each column.

## 208 5 Limitations

209 More discussion and analysis on the two primary limitations can be found in the Supplementary Materials.  
210 First, pseudo-labeling methods may propagate biases from the parent model into downstream  
211 models. In our case, the increased detectability of forward-scattered light from smoke particulates  
212 may bias the model toward higher performance at larger solar zenith angles. Second, the HMS annotations  
213 do not distinguish between fire types and include a large number of controlled agricultural  
214 burns, which may limit the dataset’s applicability for targeting large-scale wildfires.

215 Several additional limitations remain important directions for future work. One is evaluating the  
216 model’s ability to distinguish smoke from dust. Another is to investigate uncertainty in the analyst  
217 annotations. Finally, while we limit the dataset to true-color GOES imagery (bands C01–C03) due  
218 to signal-to-noise and dataset size considerations, future studies should investigate the benefits of  
219 incorporating additional spectral bands, especially C07.

## 220 6 Conclusion

221 In this study, we present **SmokeViz**, a refined satellite imagery dataset for semantic segmentation of  
222 wildfire smoke plumes. Starting from the original NOAA HMS annotations of coarse, many-to-one  
223 approximations of smoke boundaries, we transform the dataset into a one-to-one mapping between  
224 satellite frames and smoke annotations. While the Mie-derived dataset selection process maximized  
225 the potential for detecting smoke if present, it did not account for whether smoke was actually visible  
226 in the selected image, leading to a high incidence of label-image mismatch and associated training  
227 noise. To address this, we introduce **pseudo-label dimension reduction (PLDR)**, a physics-guided,  
228 semi-supervised method that uses a parent model trained on the Mie-derived dataset ( $\mathcal{D}_M$ ) to generate

229 pseudo-labels across each annotation's time window. We then select the image with the highest spatial  
230 overlap between the intermediary pseudo-label and the HMS annotation to construct a refined dataset  
231 ( $\mathcal{D}_p$ ). A child model trained on  $\mathcal{D}_p$  achieves higher segmentation performance than the original parent  
232 model, as measured by IoU on both test sets, demonstrating the value of pseudo-label-based temporal  
233 alignment.

234 SmokeViz serves as a robust and representative dataset for training models to detect wildfire smoke in  
235 GOES imagery at the frame level. In addition to supporting real-time smoke segmentation, this dataset  
236 has potential applications in early wildfire detection, air quality monitoring, and as a smoke analysis  
237 product for data assimilation into dispersion models. It also provides a challenging benchmark for  
238 remote sensing models tasked with segmenting diffuse, low-contrast features like smoke. More  
239 generally, this work illustrates how PLDR can be used to resolve resolution mismatches between data  
240 and labels, especially in settings with time-series or video data paired with coarse annotations. The  
241 dataset is publicly available at <https://noaa-gsl-experimental-pds.s3.amazonaws.com/index.html#SmokeViz/> with code available at <https://github.com/anonymous-smokeviz/SmokeViz>. HELLOW

## 244 References

- 245 [1] Joseph E Aldy, Maximilian Auffhammer, Maureen Cropper, Arthur Fraas, and Richard Mor-  
246 genstern. Looking back at 50 years of the clean air act. *Journal of Economic Literature*, 60(1):  
247 179–232, 2022.
- 248 [2] Marshall Burke, Anne Driscoll, Sam Heft-Neal, Jiani Xue, Jennifer Burney, and Michael Wara.  
249 The changing risk and burden of wildfire in the united states. *Proceedings of the National  
250 Academy of Sciences*, 118(2):e2011048118, 2021.
- 251 [3] Emmanuela Gakidou, Ashkan Afshin, Amanuel Alemu Abajobir, Kalkidan Hassen Abate,  
252 Cristiana Abbafati, Kaja M Abbas, Foad Abd-Allah, Abdishakur M Abdulle, Semaw Ferede  
253 Abera, Victor Aboyans, et al. Global, regional, and national comparative risk assessment  
254 of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks,  
255 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390  
256 (10100):1345–1422, 2017.
- 257 [4] Steven J Goodman, Timothy J Schmit, Jaime Daniels, and Robert J Redmon. *The GOES-R  
258 series: a new generation of geostationary environmental satellites*. Elsevier, 2019.
- 259 [5] Tom X.-P. Zhao, Steve Ackerman, and Wei Guo. Dust and smoke detection for multi-channel  
260 imagers. *Remote Sensing*, 2(10):2347–2368, 2010. ISSN 2072-4292. doi: 10.3390/rs2102347.  
261 URL <https://www.mdpi.com/2072-4292/2/10/2347>.
- 262 [6] T Randriambelo, S Baldy, M Bessafi, Michel Petit, and Marc Despinoy. An improved detection  
263 and characterization of active fires and smoke plumes in south-eastern africa and madagascar.  
264 *International Journal of Remote Sensing*, 19(14):2623–2638, 1998.
- 265 [7] Eric James, Ravan Ahmadov, and Georg A Grell. Realtime wildfire smoke prediction in the  
266 united states: The hrrr-smoke model. In *EGU General Assembly Conference Abstracts*, page  
267 19526, 2018.
- 268 [8] Ravan Ahmadov, Haiqin Li, Johana Romero-Alvarez, Jordan Schnell, Sudheer Bhimireddy,  
269 Eric James, Ka Yee Wong, Ming Hu, Jacob Carley, Partha Bhattacharjee, et al. Forecasting  
270 smoke and dust in noaa's next-generation high-resolution coupled numerical weather prediction  
271 model. Technical report, Copernicus Meetings, 2024.
- 272 [9] Donna McNamara, George Stephens, Mark Ruminski, and Tim Kasheta. The hazard mapping  
273 system (hms) - noaa's multi-sensor fire and smoke detection program using environmental  
274 satellites. *Conference on Satellite Meteorology and Oceanography*, 01 2004.
- 275 [10] W Schroeder, M Ruminski, I Csiszar, L Giglio, E Prins, C Schmidt, and J Morisette. Validation  
276 analyses of an operational fire monitoring product: The hazard mapping system. *International  
277 Journal of Remote Sensing*, 29(20):6059–6066, 2008.

- 278 [11] NOAA. Hazard mapping system fire and smoke product, 2024. URL <https://www.ospo.noaa.gov/Products/land/hms.html#about>.
- 280 [12] Rui Ba, Chen Chen, Jing Yuan, Weiguo Song, and Siuming Lo. Smokenet: Satellite smoke  
281 scene detection using convolutional neural network with spatial and channel-wise attention.  
282 *Remote Sensing*, 11(14):1702, 2019.
- 283 [13] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi.  
284 Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of  
285 the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023.
- 286 [14] Jeff Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satel-  
287 lite imagery. *ArXiv*, abs/2109.01637, 2021. URL <https://api.semanticscholar.org/CorpusID:237416777>.
- 289 [15] Alexandra Larsen, Ivan Hanigan, Brian J Reich, Yi Qin, Martin Cope, Geoffrey Morgan, and  
290 Ana G Rappold. A deep learning approach to identify smoke plumes in satellite imagery in  
291 near-real time for health risk communication. *Journal of exposure science & environmental  
292 epidemiology*, 31(1):170–176, 2021.
- 293 [16] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-  
294 scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE  
295 international geoscience and remote sensing symposium*, pages 5901–5904. IEEE, 2019.
- 296 [17] J Jakubik, S Roy, C Phillips, P Fraccaro, D Godwin, B Zadrozny, D Szwarcman, C Gomes,  
297 G Nyirjesy, B Edwards, et al. Foundation models for generalist geospatial artificial intelligence.  
298 arxiv 2023. *arXiv preprint arXiv:2310.18660*, 2023.
- 299 [18] Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. Polyworld:  
300 Polygonal building extraction with graph neural networks in satellite images. In *Proceedings  
301 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1848–1857,  
302 2022.
- 303 [19] Asanobu Kitamoto, Jared Hwang, Bastien Vuillod, Lucas Gautier, Yingtao Tian, and Tarin  
304 Klanuwat. Digital typhoon: Long-term satellite image dataset for the spatio-temporal modeling  
305 of tropical cyclones. *Advances in Neural Information Processing Systems*, 36, 2024.
- 306 [20] Bjorn Stevens, Sandrine Bony, Hélène Brogniez, Laureline Hentgen, Cathy Hohenegger,  
307 Christoph Kiemle, Tristan S L’Ecuyer, Ann Kristin Naumann, Hauke Schulz, Pier A Siebesma,  
308 et al. Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds. *Quarterly  
309 Journal of the Royal Meteorological Society*, 146(726):141–152, 2020.
- 310 [21] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method  
311 for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning  
(WREPL)*, 07 2013.
- 313 [22] Martin Raspaud, David Hoesel, Adam Dybbroe, Panu Lahtinen, Abhay Devasthale, Mikhail  
314 Itkin, Ulrich Hamann, Lars Ørum Rasmussen, Esben Stigård Nielsen, Thomas Leppelt, et al.  
315 Pytroll: An open-source, community-driven python framework to process earth observation  
316 satellite data. *Bulletin of the American Meteorological Society*, 99(7):1329–1336, 2018.
- 317 [23] MK Bah, MM Gunshor, and TJ Schmit. Generation of goes-16 true color imagery without a  
318 green band. *Earth and Space Science*, 5(9):549–558, 2018.
- 319 [24] Alain Royer, Pierre Vincent, and Ferdinand Bonn. Evaluation and correction of viewing angle  
320 effects on satellite measurements of bidirectional reflectance. *Photogrammetric engineering  
321 and remote sensing*, 51(12):1899–1914, 1985.
- 322 [25] Pavel Iakubovskii. Segmentation models pytorch. [https://github.com/qubvel/segmentation\\_models.pytorch](https://github.com/qubvel/segmentation_models.pytorch), 2019.
- 324 [26] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International  
325 conference on machine learning*, pages 10096–10106. PMLR, 2021.

- 326 [27] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene  
327 parsing network. In *Proceedings of the IEEE conference on computer vision and pattern*  
328 *recognition*, pages 2881–2890, 2017.
- 329 [28] Rafael EP Ferreira, Yong Jae Lee, and João RR Dórea. Using pseudo-labeling to improve  
330 performance of deep neural networks for animal identification. *Scientific Reports*, 13(1):13875,  
331 2023.
- 332 [29] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille.  
333 Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and  
334 fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):  
335 834–848, 2017.
- 336 [30] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction.  
337 In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–  
338 12188, 2021.
- 339 [31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,  
340 Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.  
341 An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*  
342 *arXiv:2010.11929*, 2020.
- 343 [32] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo.  
344 Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances*  
345 *in neural information processing systems*, 34:12077–12090, 2021.
- 346 [33] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing  
347 for scene understanding. In *Proceedings of the European conference on computer vision*  
348 (*ECCV*), pages 418–434, 2018.
- 349 [34] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit:  
350 Memory efficient vision transformer with cascaded group attention. In *Proceedings of the*  
351 *IEEE/CVF conference on computer vision and pattern recognition*, pages 14420–14430, 2023.