
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) [10], 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 [32]. 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 [22].
 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 [19, 25]. 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 [3].
 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 [15]. 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 [30].
 67 The success of deep learning methods, such as CNNs, relies heavily on the availability of a large,
 68 representative dataset [27]. As laid out in table 1, prior studies use relatively small numbers of
 69 samples, from 47 [29] to 6825 [30], 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 [14],
 72 [8], [7]. 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
[3]	CNN	6255	MODIS	5	image
[30]	CNN	6825	GOES-East	5	pixel
[15]	CNN	975	Himiware-8	7	pixel
[29]	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

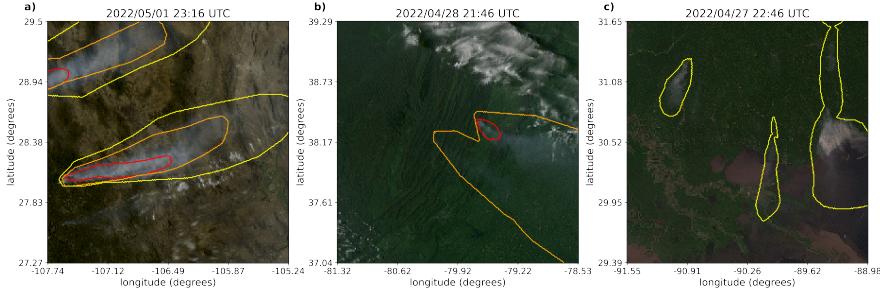


Figure 1: Satellite imagery captured by GOES-East within a few days of each other. The yellow, orange and red contours indicate the extent of Light, Medium and Heavy smoke. a) shows a canonical example of a smoke plume. b) and c) show observable variations in the density labels.

79 trains a new model using the pseudo-labels [16]. 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 [20].

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 [5], 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

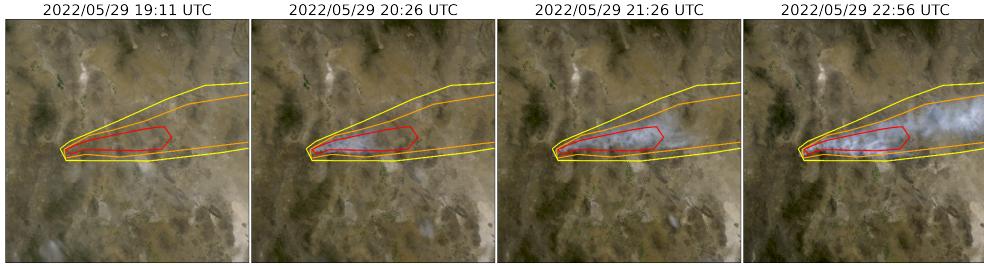


Figure 2: True color GOES-East imagery from May 2022, Southeast New Mexico (31°N , 100°W) during the start of the Foster Fire. The red, orange and yellow lines represent the heavy, medium and low density HMS smoke annotations that span 19:10–23:00 UTC.

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 [23], we input bands 1-3 (Table 3) to a GOES specific true color composite
 156 algorithm [4] to develop a, 1km resolution, true color image representation, similar to what is seen by
 157 HMS analysts. As discussed in further detail in the next section, the highest signal-to-noise ratio will
 158 come from the smallest wavelengths of light, higher wavelengths have lower smoke signal and higher
 159 noise (figure 5). For that reason, we only include the first 3 out of 16 available bands of data.

160 **Mie-Derived Dataset**

161 We used a physics-informed approach in selecting the initial GOES dataset, \mathcal{X}_M , which we call the
 162 Mie-derived dataset, for training an initial parent model, f_o , where if \mathcal{X} represents all the GOES
 163 imagery corresponding to the HMS smoke annotation time window, $\mathcal{X}_M \subset \mathcal{X}$. Prior GOES ABI
 164 datasets for machine learning applications often include data from only one of the two GOES-series
 165 satellites, commonly opting for GOES-East [30], [21], [17]. Rather than using one satellite or the
 166 cumulative data from both GOES-West and GOES-East images, we select between one or the other
 167 based on the solar zenith angle. For smoke identification, this approach can achieve a much higher
 168 signal-to-noise than imaging the earth’s surface from an arbitrary angle. The elastic scattering of
 169 light is the primary mechanism to account for - while the atmosphere is composed of molecules
 170 with size $< 1\text{nm}$, smoke particles can vary from $100\text{ nm} - 10\text{ }\mu\text{m}$ in diameter, d . The GOES ABI
 171 covers spectral bands from $0.47\text{ }\mu\text{m} - 13.3\text{ }\mu\text{m}$, so atmospheric and smoke particle sizes occupy two
 172 very different regimes with respect to the imaging wavelength λ . In the extreme limit of $\lambda \gg d$, the
 173 physics of scattering of light off a small sphere is captured by Rayleigh scattering. This process has
 174 two critical consequences: (1) the scattering cross section of light is strongly wavelength dependent
 175 (scaling with λ^{-4}), meaning that photons with wavelength closer to the ultraviolet are scattered more
 176 strongly than infrared photons. (2) the scattering cross section scales with an angular dependent
 177 cross section of $(1 + \cos^2 \theta)$. Scattered photons follow the emission distribution of a radiating dipole,
 178 scattering more strongly in the forward and backwards directions ($\theta = 0, \pi$) than orthogonal to the
 179 direction of propagation ($\theta = \pi/2, 3\pi/2$), see figure 3 for a Rayleigh scattering schematic.

180 The significance of these scalings is that the observer, or detector, will receive blue photons in most
 181 directions orthogonal to the source. Equivalently, photons traveling colinearly with line of sight to
 182 the emission source will mostly have wavelengths in the infrared band. In the converse regime of
 183 $d > \lambda$, the elastic scattering of light against matter is modeled through Mie scattering. In comparison
 184 to Rayleigh scattering, Mie scattering is largely wavelength-independent and has a more complicated
 185 radiation pattern where the cross section has a maximal amplitude in the forward direction. An

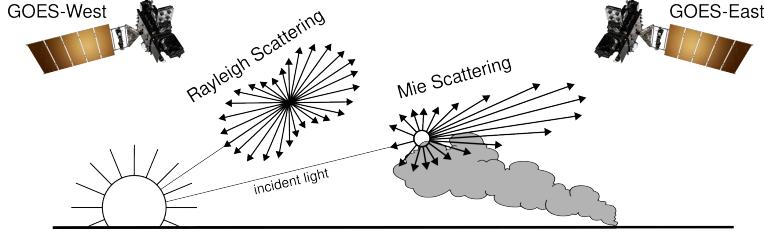


Figure 3: If the particle size is $< \frac{1}{10}$ the wavelength of the interacting light, then the primary scattering will be Rayleigh. Mie scattering is the predominant scattering mechanism when the particle size is larger than the wavelength of light. This schematic demonstrates that when the sun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

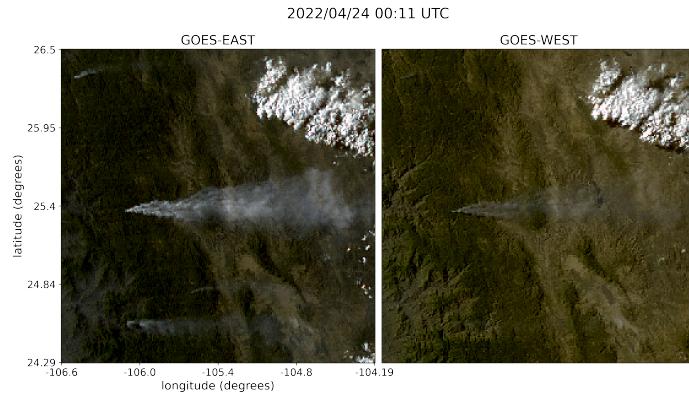


Figure 4: True color GOES-East (left) and GOES-West (right) imagery from April 24th, 2022 in Durango, Mexico. The images were taken ~ 0.5 hours before sunset (01:43 UTC) for this geolocation and time of year.

- 186 observer downstream of this scatterer will collect more photons than one positioned directly behind it.
 187 In the context of smoke identification, a sunrise or sunset will lead to a higher Mie scattered signal in
 188 GOES-West and GOES-East respectively, as shown with a smoke plume producing a stronger signal
 189 in GOES-East imagery near sunset in figure 4.
- 190 Smoke identification therefore amounts to extracting a signal of $d > \lambda$ photons from the $\lambda \gg d$
 191 background. Positioning a detector along line of sight to the scatterer will result in a higher signal
 192 from smoke particles (figure 3). Filtering the imaged wavelength can enhance this signal; photons
 193 collected in the blue spectrum will have a naturally lower background along the line of sight to the
 194 illumination source do their high level of Rayleigh scattering as. Therefore, as demonstrated in figure
 195 5, this configuration results in the highest signal to noise imaging for smoke particles.
- 196 Based solely on these criteria, the optimal strategy would be to pull data from GOES-West right after
 197 sunrise and from GOES-East right before sunset. Another factor to consider is that the time when the
 198 sun is in optimal alignment with the satellite for smoke detection coincides with when solar zenith
 199 angle is maximized. Larger angles between the satellite and sun result in an increase in noise due
 200 to increased atmospheric interactions [24]. This is shown in figure 6, while we optimize for smoke
 201 signal detection, due to the high solar zenith angle, we introduce atmospheric interaction noise that
 202 obfuscate the smoke signal. To reduce the noise from large solar zenith angles, if given multiple
 203 frames to choose from, we choose the image with the largest solar zenith angle that is $< 80^\circ$.
- 204 The resulting image selection process takes into account atmospheric properties and light scattering
 205 physics to generate an estimate of which singular satellite image within the analyst time-window could
 206 give the highest smoke signal-to-noise ratio. The resulting Mie-derived dataset, $\mathcal{X}_M = \{X_M, Y\}$,
 207 was then used to train a model, f_o , that would generate N pseudo-labels, y^* , for every sample,
 208 where N is determined by how many images, taken at a 10 minute interval, fit within the analyst

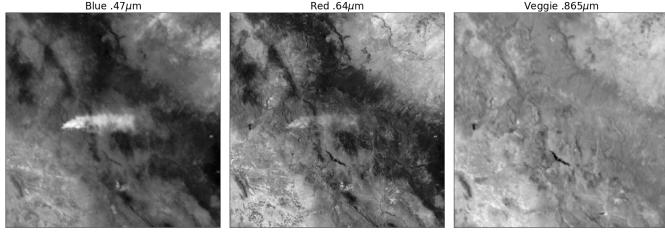


Figure 5: Three bands of GOES-East data are the raw input to generate a true color image. These plots show variations in the signal-to-noise ratio for smoke detection in relation to the wavelength, λ , of light being measured.

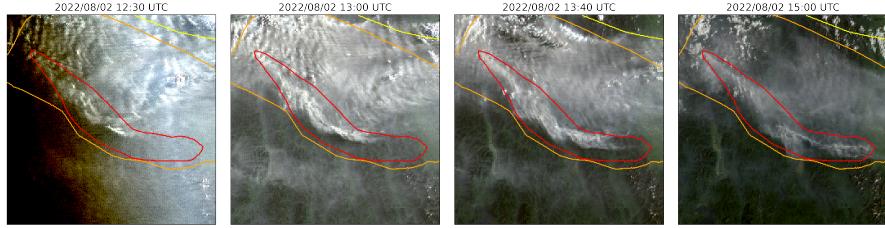


Figure 6: A smoke annotation projected onto GOES-West imagery from August 2022 that spans from 11:00 UTC to 15:00 UTC, sunrise on August 2nd, 2022 at coordinates (49°24'N, 115°29'W) was 12:15 UTC.

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

211 Machine Learning Model

212 We implement a deep learning architecture that uses the encoder from EfficientNetV2 [28] and a
 213 semantic segmentation classifier from the DeepLabV3 model [6]. Transfer learning has shown to
 214 reduce the time and resources needed to train a model by leveraging information from pre-trained
 215 models [31], [26]. We initialize the values of our model weights using the pre-trained values originally
 216 trained on the ImageNet dataset [7], containing 1.2 million images and 1000 categories. Our model
 217 was developed using the Segmentation Models PyTorch package [12] that was written as a high level
 218 API for implementing models for semantic segmentation problems. We input 256x256x3 snapshots of
 219 1km resolution true color GOES imagery that contains smoke and output a 256x256x3 classification
 220 map that predicts if a pixel contains smoke and if so, what the density of that smoke is. As mentioned
 221 earlier, we apply the thermometer encoding shown in table 2 to encode the smoke densities and apply
 222 binary cross entropy as the loss function per density of smoke.

223 The dataset, \mathcal{X}_M , contained over 130,000 samples. To train f_o , we split \mathcal{X}_M into training (118,691
 224 samples), validation (8,335 samples) and testing (7,474 samples) datasets. Training data contains
 225 data from the years 2018, 2019, 2020, 2021 and 2023 while the data from 2022 is split into validation
 226 and testing sets by taking data from alternating 10 days of the year. In order to make sure we include
 227 the monthly variations in wildfire trends over a full year, we split 2022 data up by every 10 days.
 228 This allowed us to: (1) allocate an additional full year of data for the training set, (2) show yearlong
 229 trends in both the validation and testing sets and (3) keep the validation and testing datasets relatively
 230 independent from one another since only two out of every ten days of data will have adjacent days in
 231 validation and testing.

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

	f_o		f_c	
	\mathcal{X}_M	\mathcal{X}_p	\mathcal{X}_M	\mathcal{X}_p
Light	0.394	0.551	0.437	0.583
Medium	0.283	0.392	0.345	0.431
Heavy	0.233	0.290	0.275	0.332
Overall	0.365	0.510	0.412	0.539

232 To determine which image out of the relevant imagery for the given time window best represents
 233 the analyst annotation, we implement a greedy algorithm by running f_o on each x to generate a
 234 pseudo-label, y^* . The output of f_o , y^* , give predictions on if smoke is in the image, and if there is
 235 smoke, where the smoke is in that image and the density of that smoke. y^* serve as pseudo-labels
 236 for each density of smoke and are compared to the analyst annotations, y . To compare y^* and y , we
 237 calculate the IoU using the total set of pixels for y^* at that density of smoke and the entire set of
 238 pixels for y for a particular smoke density in each image as shown in equation 1. The image with the
 239 highest IoU score is chosen as the image, x_p , that best represents the analyst smoke annotation, y .
 240 Often used for pseudo-labeling, a confidence threshold value is defined to determine if a pseudo-label
 241 should to be included in a dataset [9]. We chose a confidence threshold that would include the sample,
 242 x_p , in \mathcal{X}_p if the maximum overall IoU (equation 1) between y^* and y over all densities was over 0.01.

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

243 Finally, we use \mathcal{X}_p to train an additional child model, f_c . We use the same dataset split method and
 244 model setup but change \mathcal{X}_M to \mathcal{X}_p to train the child model. For training both f_c and f_p we train each
 245 model over 10 epochs using the Adam optimizer on a single Nvidia A100 GPU allocating 10GB of
 246 memory over 80 hours of allotted training time.

247 Results

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

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

261 3 Limitations

262 One of the concerns that comes with using pseudo-labeling methods is that you can perpetuate biases
 263 from the parent model into subsequent child models. Due to the increase in detectable forward
 264 scattered light off smoke particular matter, we expect the model to have a bias towards producing a
 265 higher success rate for smoke detection at larger solar zenith angles. Another concern is the possibility
 266 of a data leak between the adjacent days every 10 days for validation and testing set. Finally, the

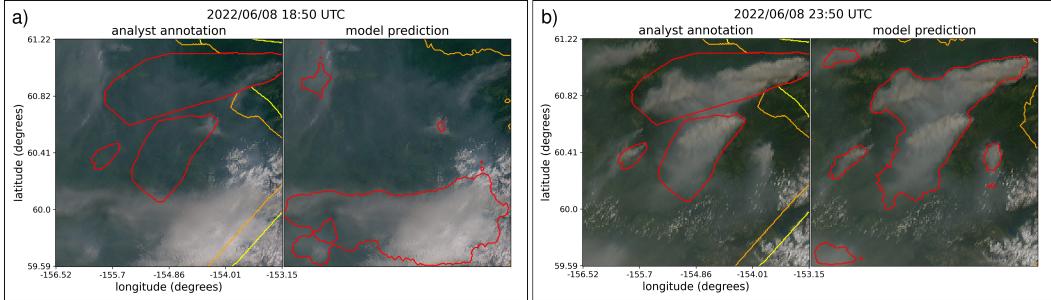


Figure 7: GOES-West imagery showing smoke on June 8th, 2022 in Alaska where, at this geolocation, daylight was between 12:43-7:53 UTC. The HMS smoke annotations displayed span from 18:50 to 23:50 UTC. a) shows the imagery that was selected using the Mie-derived data selection process b) shows the image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

267 original HMS dataset is not split by type of fire and includes a large portion of small, controlled burns.
 268 This can be a limitation to consider if the dataset is being used to detect large wildfires. All these
 269 limitations are discussed and analyzed further in the Appendix.

270 4 Conclusion

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

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

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

303 **5 Acknowledgments and Disclosure of Funding**

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

312 **References**

- 313 [1] R. S. Allison, J. M. Johnston, G. Craig, and S. Jennings. Airborne optical and thermal remote
314 sensing for wildfire detection and monitoring. *Sensors*, 16(8):1310, 2016.
- 315 [2] N. Andela, J. Kaiser, G. Van der Werf, and M. Wooster. New fire diurnal cycle characterizations
316 to improve fire radiative energy assessments made from modis observations. *Atmospheric
317 Chemistry and Physics*, 15(15):8831–8846, 2015.
- 318 [3] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo. Smokenet: Satellite smoke scene detection using
319 convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 11(14):
320 1702, 2019.
- 321 [4] M. Bah, M. Gunshor, and T. Schmit. Generation of goes-16 true color imagery without a green
322 band. *Earth and Space Science*, 5(9):549–558, 2018.
- 323 [5] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to
324 resist adversarial examples. In *International conference on learning representations*, 2018.
- 325 [6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic
326 image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.
327 *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- 328 [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical
329 image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages
330 248–255. Ieee, 2009.
- 331 [8] L. Deng. The mnist database of handwritten digit images for machine learning research [best of
332 the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.
- 333 [9] R. E. Ferreira, Y. J. Lee, and J. R. Dórea. Using pseudo-labeling to improve performance of
334 deep neural networks for animal identification. *Scientific Reports*, 13(1):13875, 2023.
- 335 [10] S. J. Goodman, T. J. Schmit, J. Daniels, and R. J. Redmon. *The GOES-R series: a new
336 generation of geostationary environmental satellites*. Elsevier, 2019.
- 337 [11] E. J. Hyer, J. S. Reid, E. M. Prins, J. P. Hoffman, C. C. Schmidt, J. I. Miettinen, and L. Giglio.
338 Patterns of fire activity over indonesia and malaysia from polar and geostationary satellite
339 observations. *Atmospheric research*, 122:504–519, 2013.
- 340 [12] P. Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.
- 341 [13] J. E. Keeley and A. D. Syphard. Large california wildfires: 2020 fires in historical context. *Fire
342 Ecology*, 17:1–11, 2021.
- 343 [14] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- 344 [15] A. Larsen, I. Hanigan, B. J. Reich, Y. Qin, M. Cope, G. Morgan, and A. G. Rappold. A deep
345 learning approach to identify smoke plumes in satellite imagery in near-real time for health risk
346 communication. *Journal of exposure science & environmental epidemiology*, 31(1):170–176,
347 2021.

- 349 [16] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep
 350 neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07
 351 2013.
- 352 [17] Y. Lee, C. D. Kummerow, and I. Ebert-Uphoff. Applying machine learning methods to detect
 353 convection using geostationary operational environmental satellite-16 (goes-16) advanced
 354 baseline imager (abi) data. *Atmospheric Measurement Techniques*, 14(4):2699–2716, 2021.
- 355 [18] J. L. McCarty, C. O. Justice, and S. Korontzi. Agricultural burning in the southeastern united
 356 states detected by modis. *Remote Sensing of Environment*, 108(2):151–162, 2007.
- 357 [19] D. McNamara, G. Stephens, M. Ruminski, and T. Kasheta. The hazard mapping system (hms) -
 358 noaa's multi-sensor fire and smoke detection program using environmental satellites. *Conference
 359 on Satellite Meteorology and Oceanography*, 01 2004.
- 360 [20] NOAA. Hazard mapping system fire and smoke product. URL <https://www.ospo.noaa.gov/Products/land/hms.html#about>.
- 362 [21] T. C. Phan and T. T. Nguyen. Remote sensing meets deep learning: exploiting spatio-temporal-
 363 spectral satellite images for early wildfire detection. 2019.
- 364 [22] T. Randriambelo, S. Baldy, M. Bessafi, M. Petit, and M. Despinoy. An improved detection
 365 and characterization of active fires and smoke plumes in south-eastern africa and madagascar.
 366 *International Journal of Remote Sensing*, 19(14):2623–2638, 1998.
- 367 [23] M. Raspaud, D. Hoese, A. Dybbroe, P. Lahtinen, A. Devasthale, M. Itkin, U. Hamann, L. Ø.
 368 Rasmussen, E. S. Nielsen, T. Leppelt, et al. Pytroll: An open-source, community-driven python
 369 framework to process earth observation satellite data. *Bulletin of the American Meteorological
 370 Society*, 99(7):1329–1336, 2018.
- 371 [24] A. Royer, P. Vincent, and F. Bonn. Evaluation and correction of viewing angle effects on
 372 satellite measurements of bidirectional reflectance. *Photogrammetric engineering and remote
 373 sensing*, 51(12):1899–1914, 1985.
- 374 [25] W. Schroeder, M. Ruminski, I. Csizar, L. Giglio, E. Prins, C. Schmidt, and J. Morisette.
 375 Validation analyses of an operational fire monitoring product: The hazard mapping system.
 376 *International Journal of Remote Sensing*, 29(20):6059–6066, 2008.
- 377 [26] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an
 378 astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision
 379 and pattern recognition workshops*, pages 806–813, 2014.
- 380 [27] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in
 381 deep learning era. In *Proceedings of the IEEE international conference on computer vision*,
 382 pages 843–852, 2017.
- 383 [28] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *International
 384 conference on machine learning*, pages 10096–10106. PMLR, 2021.
- 385 [29] Z. Wang, P. Yang, H. Liang, C. Zheng, J. Yin, Y. Tian, and W. Cui. Semantic segmentation and
 386 analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery.
 387 *Remote Sensing*, 14(1):45, 2022.
- 388 [30] J. Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery.
 389 *ArXiv*, abs/2109.01637, 2021. URL [https://api.semanticscholar.org/CorpusID:
 390 237416777](https://api.semanticscholar.org/CorpusID:237416777).
- 391 [31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural
 392 networks? *Advances in neural information processing systems*, 27, 2014.
- 393 [32] T. X.-P. Zhao, S. Ackerman, and W. Guo. Dust and smoke detection for multi-channel imagers.
 394 *Remote Sensing*, 2(10):2347–2368, 2010. ISSN 2072-4292. doi: 10.3390/rs2102347. URL
 395 <https://www.mdpi.com/2072-4292/2/10/2347>.

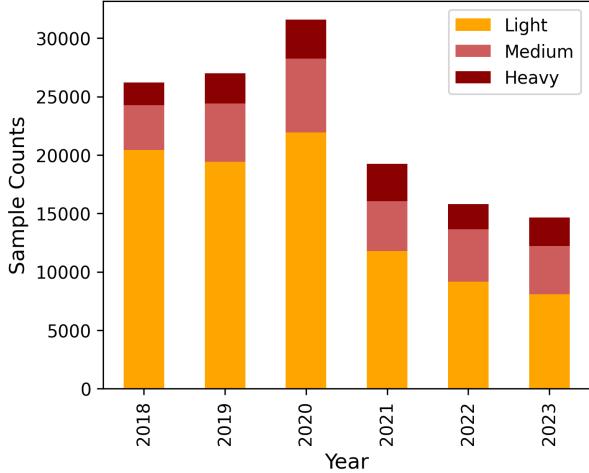


Figure 8: Sample count per year

396 A Appendix

397 A.1 Original Data and Software Licenses

398 The HMS Smoke product does not have a license attached to it. For GOES imagery, NOAA states
 399 "There are no restrictions on the use of this data" and does not provide a license. Pytroll is distributed
 400 under the GNU General Public License v3.0 license while Segmentation Models Pytorch is distributed
 401 under the MIT License.

402 A.2 Statistical Visualizations for SmokeViz Dataset

403 Figures 8, 9, 10, 11 provide some statistical analysis on \mathcal{X}_p . As seen in figure 8, we see the highest
 404 number of samples for the year 2020 that showed a high volume of available annotations that year
 405 likely due to the large number of wildfires [13] during 2020. The peak for number of samples shown
 406 in figure 9 is March and April, coming right before the typical wildfire season that usually goes from
 407 late Spring through Fall. This may be due to the increase in prescribed agricultural burns before
 408 plants emerge from winter dormancy [18]. The HMS analysts do not have a way of distinguishing
 409 between planned or uncontrolled fire, so many of the annotations represent small agricultural burns
 410 along with wildfires.

411 As shown in figure 10, the states with the highest number of samples are California, Georgia and
 412 Florida. The high frequency in fires in the Southeast may be due to the aforementioned prescribed
 413 agricultural burns. Analysts are looking not only at the United States, but also Canada and Mexico,
 414 figure 11 shows a breakdown of the number of samples that originate from each country.

415 A.3 Model Performance Analysis

416 In order to get a better understanding of the dataset, we use the deep learning models to analyze
 417 certain data characteristics. Figure 12 shows variations in overall IoU values running f_o on the \mathcal{X}_p
 418 test set data. The highest IoU are during the typical wildfire season and outside the typical window
 419 for prescribed agricultural burns.

420 We report on how many samples come from each satellite in table 5, along with the \mathcal{X}_p test set
 421 IoU in comparison to the HMS analyst annotations. While GOES-EAST provides over triple the
 422 number of training samples, f_c performs better on GOES-WEST samples out of the test set. The
 423 signal observed by a single satellite vary diurnally and annually in the amount of atmospheric noise
 424 and solar radiation. In turn, if provided with enough samples, this could create a more robust and
 425 generalizable model to the extent of being able to perform well on two different sensors with varying
 426 calibrations and line of sights.

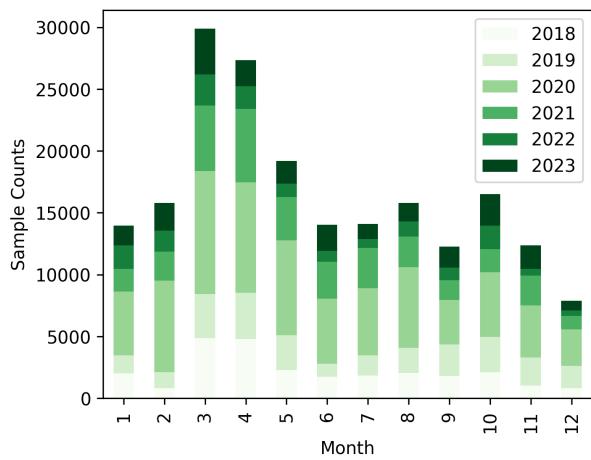


Figure 9: Sample count per month.

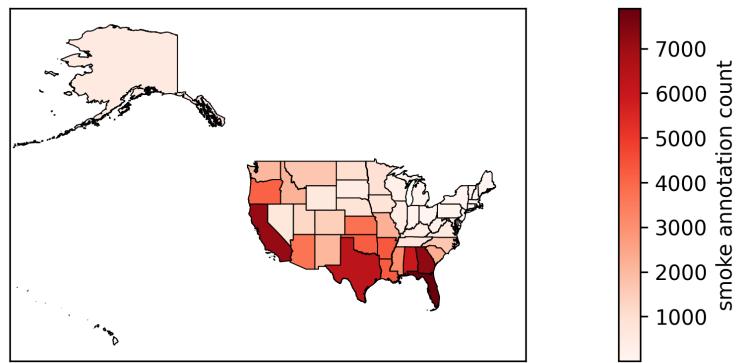


Figure 10: Sample count per US state.

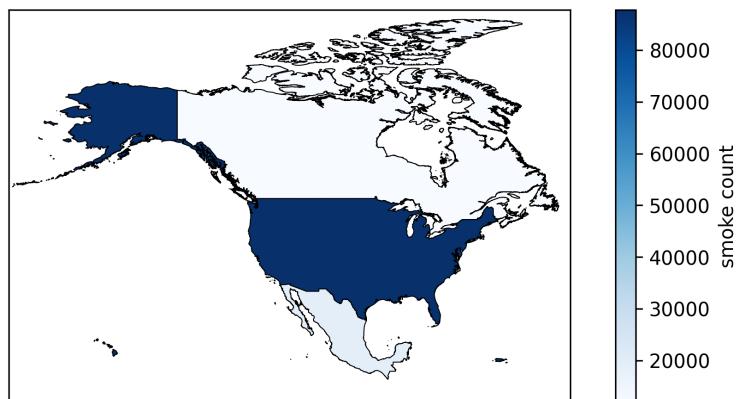


Figure 11: Sample count per North American country.

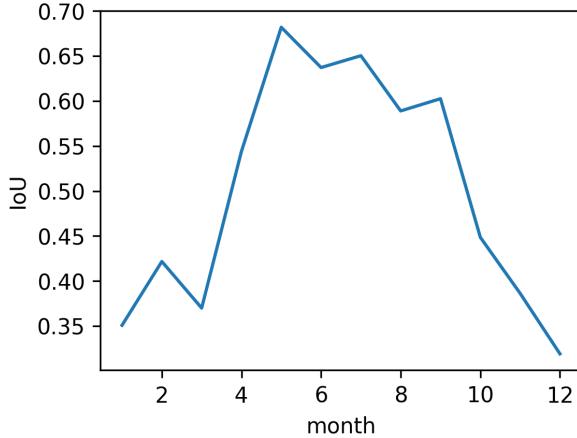


Figure 12: IoU between f_c predictions and analyst annotations per month for \mathcal{X}_p test set.

Table 5: Sample count along with variations in f_c performance depending on which GOES satellite data is used.

Satellite	Test IoU	\mathcal{X}_p Test Samples	\mathcal{X}_p Samples
GOES-WEST	0.645	1827	30640
GOES-EAST	0.483	5647	119040

427 As mentioned in the limitations, there may have been a bias introduced towards correctly classifying
 428 imagery close to sunrise or sunset. This bias may not only be introduced by our Mie-derived dataset
 429 that was used to train f_o , but also in the original HMS annotations. The configuration of the sun,
 430 smoke and satellite give the highest signal-to-noise ratio at the times near the sunrise and sunset,
 431 making smoke more easily observable. In contrast, the diurnal variations of wildfires cause the
 432 fire radiative power to be highest around solar noon [2]. Table 6 shows how the IoU between f_c
 433 predictions and analyst annotations for the test data from either \mathcal{X}_M or \mathcal{X}_p are not significantly
 434 affected by being within 2 hours to sunrise/sunset. The main difference we see from table 6 is the
 435 split of closer to daylight boundaries is shifted towards midday between \mathcal{X}_M to \mathcal{X}_p . This is because,
 436 for \mathcal{X}_p , we are choosing the imagery with the best overlap to the analyst product rather than the image
 437 from \mathcal{X}_M that optimized for highest possible signal-to-noise ratio if given constant signal.

438 In order to observe geographical regional variations we create quadrants, Northwest (NW), Southwest
 439 (SW), Northeast (NE) and Southeast (SE) in relation to the midpoint (40, -100) and show the
 440 sample distribution and model performance for each region in table 7. The table shows the worst f_c
 441 performance in the SE quadrant despite representing this largest fraction of the training data. This
 442 is likely due to the large number of aforementioned prescribed burns in that area. If the goal of
 443 the dataset is to be used to train a model to detect and monitor large wildfires, a weakness in the
 444 dataset would be that it likely consists of a lot more small, controlled agricultural burns that aren't
 445 representative of the intended task.

446 A weakness in the dataset split for 2022 validation and testing sets is that there are adjacent days
 447 between the rotating 10 day splits. This is a weakness because wildfires often last more than one day,
 448 smoke from the same fires are likely to leak between the datasets. The choice to split the dataset
 449 every 10 days was a trade off between being able to keep another day for training and keeping the
 450 validation and test set completely independent. Another consideration for the choice was that we
 451 expect the diurnal variations in smoke characteristics to vary largely enough at either ends of the
 452 nocturnal stagnations in fire activity [11]. The scope of this paper was to use the deep learning models
 453 as a way of optimizing the dataset and comparing the datasets against each other. While the data leak
 454 is not likely to have high consequences for this particular application (as suggested in table 8), we
 455 encourage users of SmokeViz to split validation and test sets so that they are completely independent,
 456 especially as new years of data are added.

Table 6: Variations in f_c performance depending on temporal proximity to sunrise or sunset.

Time difference	\mathcal{X}_M Test Set IoU	\mathcal{X}_p Test Set IoU	\mathcal{X}_M Test Samples	\mathcal{X}_p Test Samples
<2 hours	0.412	0.546	3923 (63%)	3436 (46%)
>2 hours	0.411	0.538	2280 (37%)	4038 (54%)

Table 7: Along with sample count we show variations in f_c performance depending on quadrant.

Quadrant	\mathcal{X}_p Test IoU	\mathcal{X}_p Test Samples	\mathcal{X}_p Samples
NW	0.5932	1425	23335
SW	0.6094	1131	26577
NE	0.4726	252	8392
SE	0.4706	4666	76130

457 A.4 Machine Learning Reproducibility

458 The models presented in this paper are not optimized for performance, but are intended to create
 459 sufficient pseudo-labels to develop the SmokeViz dataset and then compare the performance of
 460 SmokeViz against the original dataset. We did not perform any experimentation for deciding on
 461 architecture or hyperparameters shown in table 9, but did make educated decisions. We chose
 462 DeepLabV3+ because smoke varies in scale and the DeepLabV3+ backbone uses a atrous spatial
 463 pyramid pooling module that allows for varying scales of the same type of object. We use the Adam
 464 optimizer that will adapt the learning rate during training and is suited for problems with large
 465 amounts of data. Batch size was chosen due to the necessity to run the model on limited resources.

466 A.5 Datasheet for SmokeViz

467 Questions from the <https://arxiv.org/abs/1803.09010> paper, v7.

468 A.5.1 Motivation

469 The questions in this section are primarily intended to encourage dataset creators to clearly articulate
 470 their reasons for creating the dataset and to promote transparency about funding interests.

471 For what purpose was the dataset created?

472 SmokeViz was created to serve as a large labeled dataset to be used in creating wildfire smoke plume
 473 related machine learning models. Applications include wildfire smoke detection or smoke dispersion
 474 modeling.

475 Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., 476 company, institution, organization)?

477 SmokeViz was created a group of researchers that at the time of the dataset creation were affiliated
 478 with The National Oceanic and Atmospheric Administration and The University of Colorado, Boulder,
 479 and The Cooperative Institute for Research in Environmental Sciences that connects CU, Boulder to
 480 NOAA.

481 Who funded the creation of the dataset?

482 This work was funded by the National Oceanic and Atmospheric Administration and The Cooperative
 483 Institute for Research in Environmental Sciences.

Table 8: Comparison of the IoU and loss between the full \mathcal{X}_p test set and the \mathcal{X}_p test set with adjacent
 days between the validation and test set removed.

\mathcal{X}_p Test Set	Overall IoU	Testing Loss
full test set	0.539	0.870
adjacent days removed	0.547	0.895

Table 9: Hyperparameters used to create f_o and f_c .

parameter	value
epochs	10
learning rate	1e-2
batch size	32
optimizer	Adam

484 **Any other comments?**

485 None.

486 **A.5.2 Composition**

487 Most of these questions are intended to provide dataset consumers with the information they need to
 488 make informed decisions about using the dataset for specific tasks. The answers to some of these
 489 questions reveal information about compliance with the EU’s General Data Protection Regulation
 490 (GDPR) or comparable regulations in other jurisdictions.

491 **What do the instances that comprise the dataset represent (e.g., documents, photos, people,
 492 countries)?**

493 Each instance is a 256x256x3 RGB image from GOES imagery with an accompanying 256x256x3
 494 binary masks corresponding to density of smoke. There are 3 densities of smoke - Light, Medium
 495 and Heavy.

496 **How many instances are there in total (of each type, if appropriate)?**

497 There are 134500 samples, 90810 for light, 28023 for medium and 15667 for Heavy density smoke.

498 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of
 499 instances from a larger set?**

500 It is intended to contain all smoke data from 2018 through 2023 but we cut out imagery if it is too
 501 bright or too dim based on photon count.

502 **What data does each instance consist of?**

503 The data is processed to correct for Rayleigh scattering, solar zenith angle and projected so each pixel
 504 is representative of the same area of land. The algorithm is referenced in the SmokeViz paper.

505 **Is there a label or target associated with each instance?**

506 Yes, there are no instances that do not contain smoke.

507 **Is any information missing from individual instances?**

508 We have seen imagery where smoke is labeled but there’s adjacent smoke plumes that were unlabeled.
 509 With human labels comes human errors.

510 **Are relationships between individual instances made explicit (e.g., users’ movie ratings, social
 511 network links)?**

512 Some instances can overlap in geographic location, there can be multiple smoke plumes in one
 513 instance, but the index of the HMS smoke annotation is listed and can be mapped back to the original
 514 dataset for geolocation information.

515 **Are there recommended data splits (e.g., training, development/validation, testing)?**

516 We recommend using full years of data for training, validation and testing, but split testing and
 517 validation every 10 days for 2022 in order to keep more data in the training set.

518 **Are there any errors, sources of noise, or redundancies in the dataset?**

519 The HMS smoke annotations that are used as truth are a source of noise as explained in the SmokeViz
 520 paper. These include approximations of smoke polygons mismatching actual location and time
 521 windows being too large that smoke moves during the time window. There is also noise caused by

522 atmospheric interactions with light. Redundancies occur when there more than one smoke plume and
523 annotation in one image.

524 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g.,
525 websites, tweets, other datasets)?**

526 The dataset is self-contained.

527 **Does the dataset contain data that might be considered confidential (e.g., data that is pro-
528 tected by legal privilege or by doctor-patient confidentiality, data that includes the content of
529 individuals' non-public communications)?**

530 No.

531 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening,
532 or might otherwise cause anxiety?**

533 No.

534 **Does the dataset relate to people?**

535 No, not directly, wildfires do affect people, but these images are at 1km resolution.

536 **Does the dataset identify any subpopulations (e.g., by age, gender)?**

537 No.

538 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or
539 indirectly (i.e., in combination with other data) from the dataset?**

540 No.

541 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that
542 reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or
543 union memberships, or locations; financial or health data; biometric or genetic data; forms of
544 government identification, such as social security numbers; criminal history)?**

545 No.

546 **Any other comments?**

547 No.

548 **A.5.3 Collection process**

549 The answers to questions here may provide information that allow others to reconstruct the dataset
550 without access to it.

551 **How was the data associated with each instance acquired?**

552 The labeled from HMS smoke product is not validated or verified but is used as verification for
553 numerical smoke dispersion modeling. The GOES imagery is collected by the ABI sensor and is
554 corrected for any anomalies and also converted from photon count to radiance values.

555 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or
556 sensor, manual human curation, software program, software API)?**

557 Original low temporal resolution annotations were manual human analyst curated. To create the high
558 temporal resolution annotations, we use pseudo-labeling discussed in the SmokeViz paper.

559 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic,
560 probabilistic with specific sampling probabilities)?**

561 The HMS smoke analysts are only looking for smoke during the daytime.

562 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and
563 how were they compensated (e.g., how much were crowdworkers paid)?**

564 The NOAA employed analysts are compensated as salaried federal employees.

565 **Over what timeframe was the data collected?**

566 2018-2023

567 **Were any ethical review processes conducted (e.g., by an institutional review board)?**

568 No.

569 **A.5.4 Preprocessing/cleaning/labeling**

570 The questions in this section are intended to provide dataset consumers with the information they
571 need to determine whether the “raw” data has been processed in ways that are compatible with their
572 chosen tasks. For example, text that has been converted into a “bag-of-words” is not suitable for tasks
573 involving word order.

574 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing,
575 tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing
576 of missing values)?**

577 The data was processed according to the GOES True Color paper referenced in the SmokeViz methods
578 section.

579 **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support
580 unanticipated future uses)?**

581 The raw data is available from the NOAA AWS webpage. <https://registry.opendata.aws/noaa-goes/>
582 The HMS smoke annotations are available here: <https://www.ospo.noaa.gov/products/land/hms.html>

583 **Is the software used to preprocess/clean/label the instances available?**

584 Yes, Pytroll implements the algorithm discussed in the GOES True Color paper referenced in the
585 SmokeViz paper.

586 **Any other comments?** None.

587 **A.5.5 Uses**

588 These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset
589 should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset
590 consumers to make informed decisions, thereby avoiding potential risks or harms.

591 **Has the dataset been used for any tasks already?**

592 Not yet.

593 **Is there a repository that links to any or all papers or systems that use the dataset?**

594 No.

595 **What (other) tasks could the dataset be used for?** Smoke dispersion modeling, automated wildfire
596 smoke detection.

597 **Is there anything about the composition of the dataset or the way it was collected and prepro-
598 cessed/cleaned/labeled that might impact future uses?**

599 No.

600 **Are there tasks for which the dataset should not be used?**

601 No. **Any other comments?** None

602 **A.5.6 Distribution**

603 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution,
604 organization) on behalf of which the dataset was created?**

605 No.

606 **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**

607 Amazon Web Services hosted by NOAA.

608 **When will the dataset be distributed?**

- 609 It is currently available.
- 610 **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
- 612 No. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
- 614 No.
- 615 **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
- 617 No.
- 618 **Any other comments?**
- 619 None.
- 620 Maintenance
- 621 These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.
- 623 **Who is supporting/hosting/maintaining the dataset?**
- 624 NOAA.
- 625 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
- 626 rey.koki@noaa.gov
- 627 **Is there an erratum?**
- 628 No.
- 629 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
- 630 yes
- 631 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
- 634 Not applicable.
- 635 **Will older versions of the dataset continue to be supported/hosted/maintained?**
- 636 No, if it needs to be updated, it is too large to keep multiple versions.
- 637 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
- 639 We encourage anyone that would like to contribute to SmokeViz to reach out to Rey Koki at
640 rey.koki@noaa.gov
- 641 **Any other comments?**
- 642 None

643 **NeurIPS Paper Checklist**

644 **1. Claims**

645 Question: Do the main claims made in the abstract and introduction accurately reflect the
646 paper's contributions and scope?

647 Answer: [Yes]

648 Justification: The claims of using pseudolabels to create a more robust dataset is reflected in
649 the paper's contributions.

650 Guidelines:

- 651 • The answer NA means that the abstract and introduction do not include the claims
652 made in the paper.
- 653 • The abstract and/or introduction should clearly state the claims made, including the
654 contributions made in the paper and important assumptions and limitations. A No or
655 NA answer to this question will not be perceived well by the reviewers.
- 656 • The claims made should match theoretical and experimental results, and reflect how
657 much the results can be expected to generalize to other settings.
- 658 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
659 are not attained by the paper.

660 **2. Limitations**

661 Question: Does the paper discuss the limitations of the work performed by the authors?

662 Answer: [Yes]

663 Justification: We address limitations of the dataset.

664 Guidelines:

- 665 • The answer NA means that the paper has no limitation while the answer No means that
666 the paper has limitations, but those are not discussed in the paper.
- 667 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 668 • The paper should point out any strong assumptions and how robust the results are to
669 violations of these assumptions (e.g., independence assumptions, noiseless settings,
670 model well-specification, asymptotic approximations only holding locally). The authors
671 should reflect on how these assumptions might be violated in practice and what the
672 implications would be.
- 673 • The authors should reflect on the scope of the claims made, e.g., if the approach was
674 only tested on a few datasets or with a few runs. In general, empirical results often
675 depend on implicit assumptions, which should be articulated.
- 676 • The authors should reflect on the factors that influence the performance of the approach.
677 For example, a facial recognition algorithm may perform poorly when image resolution
678 is low or images are taken in low lighting. Or a speech-to-text system might not be
679 used reliably to provide closed captions for online lectures because it fails to handle
680 technical jargon.
- 681 • The authors should discuss the computational efficiency of the proposed algorithms
682 and how they scale with dataset size.
- 683 • If applicable, the authors should discuss possible limitations of their approach to
684 address problems of privacy and fairness.
- 685 • While the authors might fear that complete honesty about limitations might be used by
686 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
687 limitations that aren't acknowledged in the paper. The authors should use their best
688 judgment and recognize that individual actions in favor of transparency play an impor-
689 tant role in developing norms that preserve the integrity of the community. Reviewers
690 will be specifically instructed to not penalize honesty concerning limitations.

691 **3. Theory Assumptions and Proofs**

692 Question: For each theoretical result, does the paper provide the full set of assumptions and
693 a complete (and correct) proof?

694 Answer: [NA]

695 Justification: No theoretical results are presented.

696 Guidelines:

- 697 • The answer NA means that the paper does not include theoretical results.
- 698 • All the theorems, formulas, and proofs in the paper should be numbered and cross-
- 699 referenced.
- 700 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 701 • The proofs can either appear in the main paper or the supplemental material, but if
- 702 they appear in the supplemental material, the authors are encouraged to provide a short
- 703 proof sketch to provide intuition.
- 704 • Inversely, any informal proof provided in the core of the paper should be complemented
- 705 by formal proofs provided in appendix or supplemental material.
- 706 • Theorems and Lemmas that the proof relies upon should be properly referenced.

707 4. Experimental Result Reproducibility

708 Question: Does the paper fully disclose all the information needed to reproduce the main ex-
709 perimental results of the paper to the extent that it affects the main claims and/or conclusions
710 of the paper (regardless of whether the code and data are provided or not)?

711 Answer: [Yes]

712 Justification: We provide the code to create the datasets along with the final dataset hosted
713 on AWS by NOAA.

714 Guidelines:

- 715 • The answer NA means that the paper does not include experiments.
- 716 • If the paper includes experiments, a No answer to this question will not be perceived
- 717 well by the reviewers: Making the paper reproducible is important, regardless of
- 718 whether the code and data are provided or not.
- 719 • If the contribution is a dataset and/or model, the authors should describe the steps taken
- 720 to make their results reproducible or verifiable.
- 721 • Depending on the contribution, reproducibility can be accomplished in various ways.
- 722 For example, if the contribution is a novel architecture, describing the architecture fully
- 723 might suffice, or if the contribution is a specific model and empirical evaluation, it may
- 724 be necessary to either make it possible for others to replicate the model with the same
- 725 dataset, or provide access to the model. In general, releasing code and data is often
- 726 one good way to accomplish this, but reproducibility can also be provided via detailed
- 727 instructions for how to replicate the results, access to a hosted model (e.g., in the case
- 728 of a large language model), releasing of a model checkpoint, or other means that are
- 729 appropriate to the research performed.
- 730 • While NeurIPS does not require releasing code, the conference does require all submis-
- 731 sions to provide some reasonable avenue for reproducibility, which may depend on the
- 732 nature of the contribution. For example
 - 733 (a) If the contribution is primarily a new algorithm, the paper should make it clear how
 - 734 to reproduce that algorithm.
 - 735 (b) If the contribution is primarily a new model architecture, the paper should describe
 - 736 the architecture clearly and fully.
 - 737 (c) If the contribution is a new model (e.g., a large language model), then there should
 - 738 either be a way to access this model for reproducing the results or a way to reproduce
 - 739 the model (e.g., with an open-source dataset or instructions for how to construct
 - 740 the dataset).
 - 741 (d) We recognize that reproducibility may be tricky in some cases, in which case
 - 742 authors are welcome to describe the particular way they provide for reproducibility.
 - 743 In the case of closed-source models, it may be that access to the model is limited in
 - 744 some way (e.g., to registered users), but it should be possible for other researchers
 - 745 to have some path to reproducing or verifying the results.

746 5. Open access to data and code

747 Question: Does the paper provide open access to the data and code, with sufficient instruc-

748 tions to faithfully reproduce the main experimental results, as described in supplemental

749 material?

750 Answer: [Yes]

751 Justification: Pseudo-labeled derived dataset is released along with code to recreate it.

752 Guidelines:

- 753 • The answer NA means that paper does not include experiments requiring code.
- 754 • Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- 755 • While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- 756 • The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- 757 • The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 758 • The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- 759 • At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- 760 • Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

762 6. Experimental Setting/Details

763 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
764 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
765 results?

766 Answer: [Yes]

767 Justification: Dataset splits, hyperparameters, optimizer are specified.

768 Guidelines:

- 769 • The answer NA means that the paper does not include experiments.
- 770 • The experimental setting should be presented in the core of the paper to a level of detail
771 that is necessary to appreciate the results and make sense of them.
- 772 • The full details can be provided either with the code, in appendix, or as supplemental
773 material.

774 7. Experiment Statistical Significance

775 Question: Does the paper report error bars suitably and correctly defined or other appropriate
776 information about the statistical significance of the experiments?

777 Answer: [No]

778 Justification: The results are represented in by the intersection over union values, there are
779 no error bars.

780 Guidelines:

- 781 • The answer NA means that the paper does not include experiments.
- 782 • The authors should answer "Yes" if the results are accompanied by error bars, confi-
783 dence intervals, or statistical significance tests, at least for the experiments that support
784 the main claims of the paper.
- 785 • The factors of variability that the error bars are capturing should be clearly stated (for
786 example, train/test split, initialization, random drawing of some parameter, or overall
787 run with given experimental conditions).
- 788 • The method for calculating the error bars should be explained (closed form formula,
789 call to a library function, bootstrap, etc.)
- 790 • The assumptions made should be given (e.g., Normally distributed errors).

- 801 • It should be clear whether the error bar is the standard deviation or the standard error
802 of the mean.
803 • It is OK to report 1-sigma error bars, but one should state it. The authors should
804 preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis
805 of Normality of errors is not verified.
806 • For asymmetric distributions, the authors should be careful not to show in tables or
807 figures symmetric error bars that would yield results that are out of range (e.g. negative
808 error rates).
809 • If error bars are reported in tables or plots, The authors should explain in the text how
810 they were calculated and reference the corresponding figures or tables in the text.

811 **8. Experiments Compute Resources**

812 Question: For each experiment, does the paper provide sufficient information on the com-
813 puter resources (type of compute workers, memory, time of execution) needed to reproduce
814 the experiments?

815 Answer: [Yes]

816 Justification: We mention the A100 GPU, 10GB of memory and 80 hours of run time.

817 Guidelines:

- 818 • The answer NA means that the paper does not include experiments.
819 • The paper should indicate the type of compute workers CPU or GPU, internal cluster,
820 or cloud provider, including relevant memory and storage.
821 • The paper should provide the amount of compute required for each of the individual
822 experimental runs as well as estimate the total compute.
823 • The paper should disclose whether the full research project required more compute
824 than the experiments reported in the paper (e.g., preliminary or failed experiments that
825 didn't make it into the paper).

826 **9. Code Of Ethics**

827 Question: Does the research conducted in the paper conform, in every respect, with the
828 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

829 Answer: [Yes]

830 Justification: There are no conflicts between the research and the NeurIPS Code of Ethics.

831 Guidelines:

- 832 • The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
833 • If the authors answer No, they should explain the special circumstances that require a
834 deviation from the Code of Ethics.
835 • The authors should make sure to preserve anonymity (e.g., if there is a special consid-
836 eration due to laws or regulations in their jurisdiction).

837 **10. Broader Impacts**

838 Question: Does the paper discuss both potential positive societal impacts and negative
839 societal impacts of the work performed?

840 Answer: [Yes]

841 Justification: There are no negative, but there are positive that are mentioned in the paper
842 such as better tools for public health decision making.

843 Guidelines:

- 844 • The answer NA means that there is no societal impact of the work performed.
845 • If the authors answer NA or No, they should explain why their work has no societal
846 impact or why the paper does not address societal impact.
847 • Examples of negative societal impacts include potential malicious or unintended uses
848 (e.g., disinformation, generating fake profiles, surveillance), fairness considerations
849 (e.g., deployment of technologies that could make decisions that unfairly impact specific
850 groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: There are no risks for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The raw NOAA datasets used to create SmokeViz do not have licenses while the python packages used do, we list these in the appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- 904 • If this information is not available online, the authors are encouraged to reach out to
905 the asset's creators.

906 **13. New Assets**

907 Question: Are new assets introduced in the paper well documented and is the documentation
908 provided alongside the assets?

909 Answer: [Yes]

910 Justification: The dataset, supporting code and user-friendly Notebooks to play with the
911 dataset/model all support the assets accessibility.

912 Guidelines:

- 913 • The answer NA means that the paper does not release new assets.
914 • Researchers should communicate the details of the dataset/code/model as part of their
915 submissions via structured templates. This includes details about training, license,
916 limitations, etc.
917 • The paper should discuss whether and how consent was obtained from people whose
918 asset is used.
919 • At submission time, remember to anonymize your assets (if applicable). You can either
920 create an anonymized URL or include an anonymized zip file.

921 **14. Crowdsourcing and Research with Human Subjects**

922 Question: For crowdsourcing experiments and research with human subjects, does the paper
923 include the full text of instructions given to participants and screenshots, if applicable, as
924 well as details about compensation (if any)?

925 Answer: [NA]

926 Justification: The paper does not involve crowdsourcing nor research with human subjects.

927 Guidelines:

- 928 • The answer NA means that the paper does not involve crowdsourcing nor research with
929 human subjects.
930 • Including this information in the supplemental material is fine, but if the main contribu-
931 tion of the paper involves human subjects, then as much detail as possible should be
932 included in the main paper.
933 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
934 or other labor should be paid at least the minimum wage in the country of the data
935 collector.

936 **15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human
937 Subjects**

938 Question: Does the paper describe potential risks incurred by study participants, whether
939 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
940 approvals (or an equivalent approval/review based on the requirements of your country or
941 institution) were obtained?

942 Answer: [NA]

943 Justification: The paper does not involve crowdsourcing nor research with human subjects.

944 Guidelines:

- 945 • The answer NA means that the paper does not involve crowdsourcing nor research with
946 human subjects.
947 • Depending on the country in which research is conducted, IRB approval (or equivalent)
948 may be required for any human subjects research. If you obtained IRB approval, you
949 should clearly state this in the paper.
950 • We recognize that the procedures for this may vary significantly between institutions
951 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
952 guidelines for their institution.
953 • For initial submissions, do not include any information that would break anonymity (if
954 applicable), such as the institution conducting the review.