Project RISE: Recognizing Industrial Smoke Emissions

Yen-Chia Hsu[1], Ting-Hao (Kenneth) Huang[2], Ting-Yao Hu[1], Paul Dille[1], Sean Prendi[1], Ryan Hoffman[1], Anastasia Tsuhlares[1], Jessica Pachuta[1], Randy Sargent[1], Illah Nourbakhsh[1]

[1]Carnegie Mellon University, Pittsburgh, PA, USA [2]Pennsylvania State University, State College, PA, USA
[1]{yenchiah,tingyaoh,pdille,sprendi,ryanhoff,anastt,jpachuta,rsargent,illah}@andrew.cmu.edu [2]txh710@psu.edu

Abstract

Industrial smoke emissions pose a significant concern to human health. Prior works have shown that using Computer Vision (CV) techniques to identify smoke as visual evidence can influence the attitude of regulators and empower citizens to pursue environmental justice. However, existing datasets are not of sufficient quality nor quantity to train the robust CV models needed to support air quality advocacy. We introduce **RISE**, the first large-scale video dataset for **R**ecognizing **I**ndustrial **S**moke **E**missions. We adopted a citizen science approach to collaborate with local community members to annotate whether a video clip has smoke emissions. Our dataset contains 12,567 clips from 19 distinct views from cameras that monitored three industrial facilities. These daytime clips span 30 days over two years, including all four seasons. We ran experiments using deep neural networks to establish a strong performance baseline and reveal smoke recognition challenges. Our survey study discussed community feedback, and our data analysis displayed opportunities for integrating citizen scientists and crowd workers into the application of Artificial Intelligence for Social Impact.

Introduction

Air pollution has been associated with adverse impacts on human health (Kampa and Castanas 2008; Pope III and Dockery 2006; Dockery et al. 1993). According to the United States Environmental Protection Agency (US EPA), air pollutants emitted from industrial sources pose a significant concern[1]. Currently, citizens who wish to advocate for better air quality rely on a manual approach (US EPA Visual Opacity Reading[2]) to determine if smoke emissions violate the permit issued to the facility. This laborious approach requires certification every six months and involves taking multiple field measurements. Prior works have shown that using Computer Vision to identify industrial smoke emissions automatically can empower citizens to pursue environmental justice and urge regulators to respond to local concerns publicly (Hsu et al. 2016, 2017). This type of data-

[1]Link to US EPA Air Pollution Sources:
https://www.epa.gov/stationary-sources-air-pollution

[2]Link to US EPA Visual Opacity Reading:
https://www.epa.gov/emc/method-9-visual-opacity

Figure 1: Dataset samples and the deployed camera system.

driven evidence, when integrated with community narratives, is essential for citizens to make sense of environmental issues and take action (Ottinger 2017; Hsu et al. 2017).

However, it is tough to collect large-scale data required to build a practical industrial smoke recognition model. Recent state-of-the-art models based on deep learning typically have millions of parameters and are data-hungry. Training these models with insufficient data can lead to overfitting. Table 1 lists publicly available datasets for smoke recognition. Compared to object and action recognition datasets, such as ImageNet (14M images, Russakovsky et al. 2015) and Kinetics (650K videos, Kay et al. 2017), existing ones for smoke recognition are small. There have been successes in prior works using deep learning for smoke recognition (Yuan et al. 2019a; Ba et al. 2019; Xu et al. 2017, 2019a,b; Filonenko, Kurnianggoro, and Jo 2017; Liu et al. 2019; Yin et al. 2017; Yang et al. 2018; Lin et al. 2019; Hu and Lu 2018; Zhang et al. 2018), but these models were trained and evaluated on relatively small datasets. In response to data sparsity, some works have attempted to create artificial data, where smoke with a transparent background were synthesized with various views (Yuan et al. 2019b; Zhang et al. 2018; Xu et al. 2019a). But such synthetic data cannot capture the rich behavior and appear-

Table 1: Comparison of publicly available datasets for recognizing smoke. This table does not include unlabeled or synthetic data. Symbol "..." means not applicable, where the corresponding datasets are image-based. The * symbol means that the dataset provides labels on image patches that do not provide sufficient context for the surroundings. The † symbol means that the dataset uses satellite images, and therefore the sources of smoke emissions are unidentifiable. We used the `ffprobe` command in FFmpeg to count video frames. Some datasets treat steam as smoke, and we counted the number of smoke frames within these datasets based only on videos that do not involve steam. Moreover, some datasets contain videos for fire detection, and we did not count them in this table, since fire detection was not our focus in this research.

	# of views	# of labeled clips	# of frames (images)	Average # of frames per clip	Ratio of smoke frames	Has temporal data?	Has context?	Is from industrial sources?	Appearance change level
This Work	19	12,567	452,412	36	41%	yes	yes	yes	high
Bugaric et al. 2014	10	10	213,909	21,391	100%	yes	yes	no	low
Ko et al. 2013	16	16	43,090	1,514	37%	yes	yes	no	low
Dimitropoulos et al. 2014	22	22	17,722	806	56%	yes	yes	no	low
Toreyin et al. 2005	21	21	18,031	820	98%	yes	yes	no	low
Filonenko et al. 2017*	...	396	100,968	255	61%	yes	no	no	low
Xu et al. 2019b	...	...	5,700	...	49%	no	yes	no	low
Xu et al. 2019a	...	...	3,578	...	100%	no	yes	no	low
Xu et al. 2017	...	...	10,000	...	50%	no	yes	no	low
Ba et al. 2019*†	...	...	6,225	...	16%	no	no	unknown	medium
Lin et al. 2017*	...	...	16,647	...	29%	no	no	no	low
Yuan et al. 2019a*	...	...	24,217	...	24%	no	no	no	low

ance of smoke under real-world conditions, such as changing weather. Moreover, existing smoke-recognition datasets lack sufficient quality for our task. No existing ones contain smoke emissions from industrial sources. Most of them are from burning materials (*e.g.*, in labs) and fire events (*e.g.*, wildfire), which have low temporal appearance changes. Thirty-six percent of them are imbalanced (ratio of smoke frames higher than 80% or less than 20%). The average number of frames per labeled clip is high in them, indicating weak temporal localization.

This paper introduces **RISE**, the first large-scale video dataset for **Recognizing Industrial Smoke Emissions**. We built and deployed camera systems to monitor activities of various industrial coke plants in Pittsburgh, Pennsylvania (Figure 1). We collaborated with air quality grassroots communities to install the cameras, which capture an image about every 10 seconds. These images were streamed back to our servers, stitched into panoramas, and stacked into timelapse videos. From these panorama videos, we cropped clips based on domain knowledge about where smoke emissions frequently occur. Finally, these clips were labeled as yes/no ("whether smoke emissions exist") by volunteers using a web-based tool that we developed (Figure 3 and 4).

Our contributions include: (1) building a dataset to facilitate using AI for social impact, (2) showing a way to empower citizens through AI research, and (3) making practical design challenges explicit to others. RISE contains clips from 30 days spanning four seasons over two years, all taken in daylight. The labeled clips have 19 views cropped from three panoramas taken by cameras at three locations. The dataset covers various characteristics of smoke, including opacity and color, under diverse weather and lighting conditions. Moreover, the dataset includes distractions of var-

ious types of steam, which can be similar to smoke and challenging to distinguish. We use the dataset to train an I3D ConvNet (Carreira and Zisserman 2017) as a strong baseline benchmark. We compare the labeling quality between citizen scientists and online crowd workers on Amazon Mechanical Turk (MTurk). Using a survey study, we received community feedback. The code and dataset[3] are open-source, as is the video labeling system[4].

Related Work

A set of prior work relied on physics or hand-crafted features to recognize smoke. Kopilovic et al. (2000) computed the entropy of optical flow to identify smoke. Celik et al. (2007) used color to define smoke pixels. Toreyin et al. (2005) combined background subtraction, edge flickering, and texture analysis. Lee et al. (2012) used change detection to extract candidate regions, computed features based on color and texture, and trained a support vector machine classifier using these features. Tian et al. (2015) presented a physical-based model and used sparse coding to extract reliable features for single-image smoke detection. Gubbi et al. (2009) and Calderara et al. (2008) applied texture descriptors (such as a wavelet transform) on small image blocks to obtain feature vectors and train a classifier using these features. These works relied on heuristics to tune the hand-crafted features, which makes it tough to be robust (Hsu et al. 2016).

Another set of prior work developed or enhanced various deep learning architectures for smoke recognition. For example, Yuan et al. (2019b) trained two encoder-decoder net-

[3]Link to the Dataset and Code: https://github.com/CMU-CREATE-Lab/deep-smoke-machine
[4]Link to the Video Labeling System: https://github.com/CMU-CREATE-Lab/video-labeling-tool

Table 2: The number and ratio of video clips for all 19 camera views filtered by various temporal conditions.

	# of labeled clips	Ratio
Has smoke	5,090	41%
Winter (Dec to Feb)	7,292	58%
Spring (Mar to May)	1,057	8%
Summer (Jun to Aug)	2,999	24%
Fall (Sep to Nov)	1,219	10%
6 am to 10 am	4,001	32%
11 am to 3 pm	6,071	48%
4 pm to 8 am	2,495	20%

works to focus on global contexts and fine details, respectively, for smoke region segmentation. Hu and Lu (2018) trained spatial-temporal ConvNets using multi-task learning. Liu et al. (2019) classified smoke by fusing ResNet and ConvNet trained with the original RGB and Dark Channel images (He, Sun, and Tang 2010), respectively. Other work applied or enhanced object detectors, such as SSD (Liu et al. 2016), MS-CNN (Cai et al. 2016), Faster R-CNN (Ren et al. 2015), and YOLO (Redmon and Farhadi 2017), to identify regions that have smoke (Xu et al. 2019a; Zhang et al. 2018; Yang et al. 2018; Lin et al. 2019). These works were evaluated on small datasets (Table 1), and none of them collaborated with local communities in air quality advocacy.

We employed citizen science (Shirk et al. 2012) to simultaneously empower laypeople and build a large dataset. Citizen science has been successfully applied in science projects, especially when the research scale makes it infeasible for experts to tackle alone. For example, PlantNet invites volunteers to submit images to develop a plant species classification model (Joly et al. 2016). In PlantNet, experts define research questions and invite citizens to participate. On the other hand, we utilized data-driven evidence to address concerns of local communities. Specifically, we applied a civic-oriented framework, *Community Citizen Science* (Hsu and Nourbakhsh 2020), to empower citizens affected by industrial pollution to advocate for better air quality.

RISE Dataset

Unlike static data annotation projects, RISE is powered by local communities, where multiple factors can influence the data distribution (explained in the Discussion section). The RISE dataset has 12,567 labeled clips from industrial sources, including those emitted from stacks and escaped from facilities. Each clip has 36 frames (resolution 180x180 pixels), representing about six minutes in real-world time. These clips contain 19 views (Figure 2), where 15 are cropped from the panorama timelapse at one site, and four from two other sites. These clips span 30 days from two years, including all four seasons. They were taken in the daytime and include different weather and lighting conditions. Using weather archives and checking the videos, we manually selected these 30 days to include various conditions, balance the number of videos that have smoke, skip days when cameras were down, and add hard cases (*e.g.*, snow

days). Of the 30 days, 20 days are cloudy, four days are fair, six days are a mixture of fair and cloudy, six days have light rain, four days have light snow, two days are windy, and one day has a thunderstorm. Using domain knowledge, we manually picked the 19 views to cover the locations of smoke emission sources. Since the length of day is different across seasons, fewer videos are in the morning (6 am to 10 am) and evening (4 pm to 8 pm). Table 2 summarizes the dataset.

System for Data Collection

We built and deployed camera systems to monitor pollution sources (Figure 1). Each system had a Nikon 1 J5 camera controlled by Raspberry Pi. It had a heater and cooler to control temperature and a servo motor to enable remote power cycling. The cost per camera was about $2,000 US dollars. The camera took a photo about every 5 to 10 seconds, and the photos were streamed to a server for stitching and stacking into panorama timelapse videos. Areas in videos that look inside house windows were cropped or blocked.

Deploying cameras relied heavily on citizen engagement. To build relationships and mutual trust between our organization and the community, we canvassed the local region, met with affected residents, and attended local events (*e.g.*, community meetings) to explain our mission. We installed the equipment and offered complimentary internet service to those residents willing to host cameras.

We chose to use timelapse instead of video capturing due to real-world deployment constraints. We deployed 10 camera systems. To zoom into locations of emission sources, our camera captured each photo in 5K resolution. Using video capturing meant streaming 10 high-resolution videos back to our server, which would be impractical since many camera locations did not have the high-speed internet infrastructure to enable such large data-transmission bandwidth. Moreover, our 10 cameras produced about 600 GB data per day. Recording videos in 20 frames per second would generate 120 TB data per day, which was beyond our system's storage space and computational power.

System for Data Annotation

We applied citizen science (Shirk et al. 2012; Irwin 2002) to engage volunteers in this research and develop a web-based tool that allows volunteers to annotate data. Citizen science opens opportunities to collaborate with residents and advocacy groups who have diverse expertise in local concerns. Through two workshops with air quality advocates, three presentations during community events, and two guest lectures at universities, we recruited volunteers to help label smoke emissions. The design and use of this tool were iteratively refined based on community feedback. For example, many volunteers mislabeled smoke initially during the workshops, suggesting that the task can be challenging for those unfamiliar with the US EPA Visual Opacity Reading. Thus, we implemented an interactive tutorial to introduce the task with step-by-step guidance. Users are first presented with simplified tasks that explain concepts, such as the characteristics of smoke and the difference between smoke and steam. After performing each small task, the system shows the answer, explanation, and the location of smoke.

Figure 2: All views of videos in the RISE dataset. The rightmost four views are from different sites pointing at another facility.

Figure 3: The individual mode of the smoke labeling system. Users can scroll the page and click or tap on the video clips to indicate that the video has smoke with a red border-box.
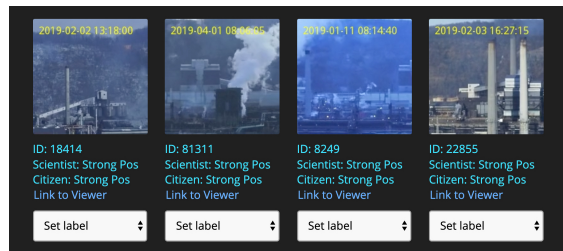

Figure 4: The collaborative mode of the smoke labeling system. Researchers can confirm citizen-provided labels by using the dropdown menu below each video.

In our tool, there are two modes for data labeling. The first one (individual mode) asks volunteers or researchers to label 16 randomly-chosen video clips at once (Figure 3). Users can scroll the interface and click or tap on clips to indicate the presence of smoke emissions. The second one (collaborative mode) asks researchers to confirm the labels contributed by volunteers, where citizens' answers are shown as prior information, and the researcher can choose to agree with or override these answers (Figure 4). This mode is designed to reduce researcher's mental load and increase the speed of gathering labels at the early stage of system deployment when the number of users is few.

Instead of labeling opacity, we decided to perform the simpler task of annotating whether smoke exists. Negative labels mean no smoke, and positive labels mean smoke appears at some time. This weak labeling approach is more straightforward for laypeople, which enabled broader citizen participation. When labeling, we referred to the 16 clips provided by the system as a page. For quality control, we randomly inserted four gold standards (labeled by a researcher). The system accepts a page if the volunteer answers all the gold standards correctly. Otherwise, the page is discarded. At least one negative and one positive sample are included in the gold standards to prevent uniform guessing. Also, each clip was reviewed by at least two volunteers (identified using Google Analytics tracker). If their answers did not agree, a third volunteer was asked to review the clip. The system takes the majority vote of the answers as the final label.

Analysis of User Contribution and Data Quality

The tool was launched in early February 2019. Through February 24, 2020 (12 months), we had 12,567 labeled clips. Among them, 42% (5,230) and 20% (2,560) were labeled by researchers and citizens in the individual mode, and 38% (4,777) were labeled in the collaborative mode. During the 12-month launch period, we attracted 60 volunteers who contributed at least one page (with 16 clips) that passed the system's quality check. Most of our volunteers received in-person training from a researcher in a 3-hour workshop. Of the entire labeled videos, only 193 (less than 2%) had volunteer label disagreement. Table 3 showed that 12% of the volunteers contributed 86% of the data (the top enthusiast group). Volunteers in this group had a higher acceptance rate (≥ 0.5) and a higher number of accepted pages ($\geq$ the average of all participants). This skewed pattern of contribution is typical among citizen science projects such as Zooniverse (https://zooniverse.org), where many volunteers participate only a few times (Sauermann and Franzoni 2015).

We compared the labels produced by citizens, researchers, and MTurk workers. For MTurk workers, we randomly sampled 720 clips from 10,625 ones that had been labeled by both citizens and researchers between February 2019 and November 2019. We then divided those clips into 60 tasks. For quality control, we added four randomly-sampled gold standards to each task. The user interface was identical to the one used for citizens. Differently, the interactive tutorial is required for MTurk workers before labeling. We first posted the tutorial task with 50 assignments to Amazon Mechanical Turk ($1.50 per task). We then posted 60 labeling tasks, where each task collected five assignments from different

Table 3: Analysis of volunteers (excluding the researchers) who contributed at least one page (with 16 clips) that passed the quality check and was accepted by the system. The format for the 4th and 6th columns is "mean±standard deviation".

User group	# of users	Page acceptance rate ∀ group	Page acceptance rate ∀ user	# of accepted pages ∀ group	# of accepted pages ∀ user
Top Enthusiasts	7 (12%)	.86	.76±.10	1,491 (86%)	213±328
Other Enthusiasts	...	...	...	...	...
Top Volunteers	41 (68%)	.69	.74±.19	218 (13%)	5±5
Other Volunteers	12 (20%)	.26	.28±.08	18 (1%)	2±1
All	60 (100%)	.81	.65±.25	1,727 (100%)	29±125

Table 4: The data quality (simulated 100 times) of citizens and MTurk workers, using researcher labels as the ground truth based on 720 labeled videos with 392 positive labels. Filtered MTurk workers' page acceptance rate is larger than 0.3. The reported format is "mean±standard deviation."

User group	Precision	Recall	F-score
Citizen	.98	.83	.90
Filtered MTurk workers	.94±.01	.89±.01	.91±.01
All MTurk workers	.93±.01	.83±.01	.88±.01

Table 5: The inter-rater agreement (simulated 100 times) between pairs of MTurk workers, citizen, and researcher groups. The reported format is "mean±standard deviation."

	Cohen's kappa
Researcher v.s. Citizen	.80
Researcher v.s. Filtered MTurk workers	.81±.01
Researcher v.s. All MTurk workers	.75±.02
Citizen v.s. All MTurk workers	.72±.02
Citizen v.s. Filtered MTurk workers	.75±.01

Table 6: Ratio of the number of videos for each split (rounded to the nearest second digit). Split S_0, S_1, S_2, S_4, S_5 is based on views. Split S_3 is based on time sequence.

	S_0	S_1	S_2	S_3	S_4	S_5
Training	.62	.62	.62	.61	.62	.62
Validation	.13	.12	.12	.14	.11	.13
Test	.25	.26	.26	.25	.26	.25

workers. Only workers who finished the tutorial task could perform these labeling tasks. The estimated time to complete a labeling task was 90 seconds. We paid $0.25 per labeling task, yielding an hourly wage of $15. Fourteen workers were recruited and accomplished all tasks in about 12 hours.

The data quality between citizens and filtered MTurk workers is similar. Filtered workers' page acceptance rates are better than 0.3 (random guessing is 0.07). To match the labeling tool's quality-control mechanism, we randomly selected three assignments for majority voting and simulated 100 times. Using researcher labels as the ground truth, Table 4 indicates similar strong performance of the positive labels (with smoke). The strong Cohen's kappa in Table 5 shows high inter-rater agreement.

Experiments

We split the dataset into training, validation, and test sets in six ways (S_0 to S_5 in Table 6). Most of the splits (except S_3) are based on camera views, where different views are used for each split. Besides S_3, the splitting strategy is that each view is in the test set at least once, where 15 views from one site are distributed among the three sets, and four from two other sites are always in the test set. In this way,

we can estimate whether the model is robust across views, as classification algorithms can suffer from overfitting in static cameras (Beery, Van Horn, and Perona 2018). Split S_3 is based on time sequence, where the farthermost 18 days are used for training, the middle two days for validation, and the nearest 10 days for testing. Thus, we can evaluate whether models trained with data in the past can be used in the future.

We treat our task as an action recognition problem and establish a baseline by using I3D ConvNet architecture with Inception-v1 layers (Carreira and Zisserman 2017), a representative model for action recognition. The inputs of this baseline are RGB frames. The model is pretrained on ImageNet (Russakovsky et al. 2015) and Kinetics (Kay et al. 2017) datasets, and then fine-tuned on our training set. During training, we apply standard data augmentation, including horizontal flipping, random resizing and cropping, perspective transformation, area erasing, and color jittering. We refer to this baseline as RGB-I3D. The validation set is used for hyper-parameter tuning. Our baseline models are optimized using binary cross-entropy loss and Stochastic Gradient Descent with momentum 0.9 for 2,000 steps. Table 8 shows the details of hyper-parameters. All models and scripts are implemented in PyTorch (Paszke et al. 2017). Model selection is based on F-score and validation error.

In order to understand the effectiveness of our baseline, we also train five other models for comparison. RGB-I3D-ND is the same baseline model without data augmentation. RGB-SVM exploits Support Vector Machine (SVM) as the classifier, which takes the pretrained I3D features (without fine-tuning on our dataset) as input. RGB-I3D-FP is trained using video clips with frame-wise random permutation. Flow-I3D has the same network architecture as our baseline, but processes precomputed TVL1 optical flow frames. It also conducts the same data augmentation pipeline, except color jittering. Flow-SVM, similar to RGB-SVM, uses raw I3D features extracted with optical flow frames.

Table 7: F-scores for comparing the effect of data augmentation and temporal models on the test set for each split. Abbreviation ND and FP means no data augmentation and with frame perturbation, respectively.

Model	S_0	S_1	S_2	S_3	S_4	S_5	Mean
RGB-I3D	.80	.84	.82	.87	.82	.75	.817
RGB-I3D-ND	.76	.79	.81	.86	.76	.68	.777
RGB-SVM	.57	.70	.67	.67	.57	.53	.618
RGB-I3D-FP	.76	.81	.82	.87	.81	.71	.797
Flow-I3D	.55	.58	.51	.68	.65	.50	.578
Flow-SVM	.42	.59	.47	.63	.52	.47	.517
RGB-TSM	.81	.84	.82	.87	.80	.74	.813
RGB-LSTM	.80	.84	.82	.85	.83	.74	.813
RGB-NL	.81	.84	.83	.87	.81	.74	.817
RGB-TC	.81	.84	.84	.87	.81	.77	.823

Table 8: Hyper-parameters. Symbol η is the initial learning rate, and i is the number of iterations to accumulate gradients before backward propagation. Milestones are the steps to decrease the learning rate by a factor of 0.1.

Model	η	Weight decay	i	Batch size	Milestones
RGB-I3D	0.1	10^{-6}	2	40	(500, 1500)
RGB-TSM	0.1	10^{-10}	1	40	(1000, 2000)
RGB-LSTM	0.1	10^{-4}	1	32	(1000, 2000)
RGB-TC	0.1	10^{-6}	1	32	(1000, 2000)

Implementations of these five models are the same as mentioned. The ConvNets not noted in Table 8 use the same hyper-parameters as RGB-I3D. Table 7 shows that I3Ds outperform SVMs by a large margin, and data augmentation can improve the performance. Also, permuting frame ordering does not degrade the performance much, and the flow-based models perform worse than their RGB counterparts.To further understand the challenge of using temporal information, we train the other five variations based on RGB-I3D with different temporal processing techniques. RGB-NL wraps two Non-Local blocks (Wang et al. 2018) in the last Inception layer (closest to the output) around the 3D convolution blocks with a kernel size larger than one. RGB-LSTM attaches one Long Short-Term Memory layer (Hochreiter and Schmidhuber 1997) with 128 hidden units after the last Inception layer. RGB-TSM wraps Temporal Shift modules (Lin, Gan, and Han 2019) around each Inception layer. RGB-TC attaches one Timeception layer (Hussein, Gavves, and Smeulders 2019) after the last Inception layer, using a 1.25 channel expansion factor. This variation is our best baseline model (Table 9). We fine-tune the LSTM and TC variation from the best RGB-I3D model with the I3D layers frozen. Table 7 shows that these temporal processing techniques do not outperform the baseline model, which confirms the challenge of using the temporal information.

To check how our baseline model makes decisions, we visualize the semantic concept by applying Gradient-weighted

Table 9: Evaluation of the best baseline model (RGB-TC) on the test set for each split. ROC/AUC means area under the receiver operating characteristic curve.

Metric	S_0	S_1	S_2	S_3	S_4	S_5	Average
Precision	.87	.84	.92	.88	.88	.78	.862
Recall	.76	.83	.77	.87	.76	.76	.792
F-score	.81	.84	.84	.87	.81	.77	.823
ROC/AUC	.90	.94	.94	.95	.92	.91	.927

Figure 5: True positives in the test set from split S_0. The top and bottom rows show the original video frame and the overlaying heatmap of Class Activation Mapping, respectively.

Class Activation Mapping (Selvaraju et al. 2017). It uses the gradients that flow into the last convolutional layer to generate a heatmap, highlighting areas that affect the prediction. Ideally, our model should focus on smoke emissions instead of other objects in the background (*e.g.*, stacks or facilities). Figure 5 shows true positives (sampled from 36 frames) for smoke and co-existence of both smoke and steam.

Survey Study

We conducted a survey to learn how and why volunteers participate. The survey was online (Google Forms), anonymous, voluntary, and had no compensation. We distributed the survey through community leaders and via email lists that we curated during the design workshops. We collected 11 responses. Table 10 shows the demographics of the survey participants. Most of them labeled smoke emissions (82%) and discussed the project in person with others (82%).

We asked an open-ended question to identify motivations. Selected quotes are in parentheses. Three participants noted the importance in advocacy (*"It is an essential tool for improving regional air quality"*), and two mentioned the desire to help research (*"To help provide more data for researchers to work with [...]"*). Others referred to their background (*"I have been interested in air pollution from local steel mills since childhood and have wanted to do something about it"*), a desire to push regulators (*"[...] it is my hope that we can use that information to force the regulatory agencies to do a better job of enforcing clean air regulations"*), and a wish to adopt our system (*"[...] hopes that this technology can help in our monitoring of other industries [...]"*).

Although participants believed our system can raise public awareness of pollution (*"[...] wider marketing to commu-*

Table 10: Demographics of 11 survey responses. Four are female. Eight have college degrees (or above).

Age	25-34	35-44	45-54	55-64	65+	Unknown
	1	2	2	1	4	1

nity members not already involved/concerned with AQ initiatives"), there was a social-technical gap (Ackerman 2000) between community expectations and the system's capability. For example, in another open-ended question about community feedback, participants express curiosity about the correlation between smoke emissions and air quality complaints (*"[...] when there are more visible emissions, are there also more reports of smells?"*). Also, participants expect the information to be useful to policy-makers (*"The videos may even provide motivation for local politicians to revise emissions regulations in a way that makes them more strict"*). Based on the survey study, more co-design workshops are needed to create data-driven evidence for making social impact, including policy changes.

Discussion

Community Engagement Challenges. Citizen science can empower residents to address air pollution (Hsu et al. 2019, 2017), but engaging communities is challenging. For example, setting up cameras and recruiting volunteers requires substantial community outreach efforts. Communities who suffer from air pollution are often financially impoverished, whose top priority might not be improving air quality. Moreover, air pollution issues are not as enjoyable as the topics of other citizen science projects in astronomy (Lintott et al. 2008) or bird watching (Sullivan et al. 2009). Also, labeling industrial smoke emissions requires volunteers to understand smoke behavior and is thus harder than recognizing generic objects. We tackled the challenges by adopting community co-design, where the people who are impacted the most participate in the design process. Understanding community expectations and concerns through conversations (*e.g.*, local meetings or workshops) is the key to building mutual trust and sustaining community-powered projects.

Wicked Problems in Data Collection. Community-powered projects like RISE suffer from the dilemma of "Wicked Problems" (Rittel and Webber 1973): *they have no precise definition, cannot be fully observed initially, depend on context, have no opportunities for trial and error, and have no optimal or provably correct solutions.* In this case, multiple uncontrollable factors can affect the data and label distribution. For instance, the availability of human labeling power within citizen groups varies over time, as community agenda changes often. Moreover, our system co-design and data labeling processes happened simultaneously and iteratively, which meant the user interface changed at each design stage, and the amount of available data was incremental. Also, we have no prior knowledge about the distribution of smoke emission across time since obtaining such information requires having profound knowledge about the pollution sources' working schedule, which is not available. The

community dynamics and underlying uncertainty can lead to imbalanced datasets. One could pursue crowdsourcing for better data quality, but crowdsourcing relinquishes the opportunity to build informed communities that can take action for social change. Making AI systems work under such real-world constraints remains an open research question.

Integration of Citizen Science and Crowdsourcing. Our analysis finds performance between volunteers and MTurk workers to be similar, suggesting opportunities for combining citizen science and crowdsourcing due to their comparable reliability. For example, at the early stage of development, one could recruit MTurk workers to label an initial small-scale dataset. Once the model is trained with the initial dataset, one can use it to produce visual evidence that can attract active citizen scientists. At this stage, the model accuracy may not reach the desired level, and human intervention may be required. However, as the community's attention starts to increase, it would be possible to collect more labels and improve model performance over time. Also, as the uncertainty in community dynamics can lead to imbalanced datasets, crowdsourcing can be a way to improve the label distribution. We leave this to future work.

Limitations. Project RISE tackles Wicked Problems, which means we can only make practical design decisions based on available information at each design stage. These practical but not optimal decisions lead to several limitations. Comparing the reliability between citizens and MTurk workers may be unfair due to the difference in their label-aggregation logic. Also, models trained on our 19 views from three facilities might not generalize well to other industrial facilities. Furthermore, RISE does not include night-time videos, which are difficult to label due to insufficient light. Moreover, our dataset does not offer bounding box labels. We applied domain knowledge to define the locations that smoke emissions were likely to occur, making smoke recognition a classification rather than a detection problem. Finally, there are other ways for aggregating labels provided by researchers and citizens, such as EM-based methods (Raykar et al. 2010). In our case, researchers always overrode the decisions made by citizens. We leave the expansion of different label types and the methodology for aggregation decisions from various user groups to future work.

Conclusion

Project RISE shows that, besides model performance, AI systems' social impact and ethics are also critical. We hope to reinforce citizens' voices and rebalance power relationships among stakeholders through system design and deployment. We have deployed the AI model to recognize smoke. Community activists and health department officers are working with our system to curate a list of severe pollution events as evidence to conduct air pollution studies. We envision that our work can encourage others to keep communities in the center of every AI system design stage. Communities affected most by social or environmental problems know their needs best and have active roles in our project. By adopting community co-design, our work demonstrates a way to forge sustainable alliances and shared prosperity between academic institutions and local citizen groups.

Acknowledgments

References

Ackerman, M. S. 2000. The Intellectual Challenge of CSCW: The Gap between Social Requirements and Technical Feasibility. *Hum.-Comput. Interact.* .

Ba, R.; Chen, C.; Yuan, J.; Song, W.; and Lo, S. 2019. SmokeNet: Satellite Smoke Scene Detection Using Convolutional Neural Network with Spatial and Channel-Wise Attention. *Remote Sensing* .

Beery, S.; Van Horn, G.; and Perona, P. 2018. Recognition in Terra Incognita. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, 472–489. Cham: Springer International Publishing. ISBN 978-3-030-01270-0.

Bugarić, M.; Jakovčević, T.; and Stipaničev, D. 2014. Adaptive estimation of visual smoke detection parameters based on spatial data and fire risk index. *Computer vision and image understanding* .

Cai, Z.; Fan, Q.; Feris, R. S.; and Vasconcelos, N. 2016. A unified multi-scale deep convolutional neural network for fast object detection. In *european conference on computer vision*. Springer.

Calderara, S.; Piccinini, P.; and Cucchiara, R. 2008. Smoke Detection in Video Surveillance: A MoG Model in the Wavelet Domain. In *Computer Vision Systems*, Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Çelik, T.; Özkaramanli, H.; and Demirel, H. 2007. Fire and smoke detection without sensors: Image processing based approach. In *European Signal Processing Conference*.

Dimitropoulos, K.; Barmpoutis, P.; and Grammalidis, N. 2014. Spatio-temporal flame modeling and dynamic texture analysis for automatic video-based fire detection. *IEEE transactions on circuits and systems for video technology* .

Dockery, D. W.; Pope, C. A.; Xu, X.; Spengler, J. D.; Ware, J. H.; Fay, M. E.; Ferris, B. G. J.; and Speizer, F. E. 1993. An Association between Air Pollution and Mortality in Six U.S. Cities. *New England Journal of Medicine* .

Filonenko, A.; Kurnianggoro, L.; and Jo, K.-H. 2017. Smoke Detection on Video Sequences Using Convolutional and Recurrent Neural Networks. In *International Conference on Computational Collective Intelligence*. Springer.

Gubbi, J.; Marusic, S.; and Palaniswami, M. 2009. Smoke detection in video using wavelets and support vector machines. *Fire Safety Journal* .

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence* .

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* .

Hsu, Y.-C.; Cross, J.; Dille, P.; Tasota, M.; Dias, B.; Sargent, R.; Huang, T.-H. K.; and Nourbakhsh, I. 2019. Smell Pittsburgh: Community-empowered Mobile Smell Reporting System. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19. ACM.

Hsu, Y.-C.; Dille, P.; Cross, J.; Dias, B.; Sargent, R.; and Nourbakhsh, I. 2017. Community-Empowered Air Quality Monitoring System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17. ACM.

Hsu, Y.-C.; Dille, P. S.; Sargent, R.; and Nourbakhsh, I. 2016. Industrial Smoke Detection and Visualization. Technical Report CMU-RI-TR-16-55, Carnegie Mellon University, Pittsburgh, PA.

Hsu, Y.-C.; and Nourbakhsh, I. 2020. When human-computer interaction meets community citizen science. *Communications of the ACM* .

Hu, Y.; and Lu, X. 2018. Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features. *Multimedia Tools and Applications* .

Hussein, N.; Gavves, E.; and Smeulders, A. W. 2019. Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Irwin, A. 2002. *Citizen science: A study of people, expertise and sustainable development*. Routledge.

Joly, A.; Goëau, H.; Champ, J.; Dufour-Kowalski, S.; Müller, H.; and Bonnet, P. 2016. Crowdsourcing biodiversity monitoring: how sharing your photo stream can sustain our planet. In *Proceedings of the 24th ACM international conference on Multimedia*.

Kampa, M.; and Castanas, E. 2008. Human health effects of air pollution. *Environmental Pollution* Proceedings of the 4th International Workshop on Biomonitoring of Atmospheric Pollution (With Emphasis on Trace Elements).

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* .

Ko, B.; Park, J.; and Nam, J.-Y. 2013. Spatiotemporal bag-of-features for early wildfire smoke detection. *Image and Vision Computing* .

Kopilovic, I.; Vagvolgyi, B.; and Sziranyi, T. 2000. Application of panoramic annular lens for motion analysis tasks: surveillance and smoke detection. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*.

Lee, C.-Y.; Lin, C.-T.; Hong, C.-T.; Su, M.-T.; et al. 2012. Smoke detection using spatial and temporal analyses. *International Journal of Innovative Computing, Information and Control* .

Lin, G.; Zhang, Y.; Xu, G.; and Zhang, Q. 2019. Smoke Detection on Video Sequences Using 3D Convolutional Neural Networks. *Fire Technology* .

Lin, G.; Zhang, Y.; Zhang, Q.; Jia, Y.; Xu, G.; and Wang, J. 2017. Smoke detection in video sequences based on dynamic texture using volume local binary patterns. *KSII Transactions on Internet and Information Systems* .

Lin, J.; Gan, C.; and Han, S. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proceedings of the IEEE International Conference on Computer Vision.*

Lintott, C. J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M. J.; Nichol, R. C.; Szalay, A.; Andreescu, D.; et al. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* .

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single shot multibox detector. In *European conference on computer vision.* Springer.

Liu, Y.; Qin, W.; Liu, K.; Zhang, F.; and Xiao, Z. 2019. A Dual Convolution Network Using Dark Channel Prior for Image Smoke Classification. *IEEE Access* .

Ottinger, G. 2017. Making sense of citizen science: stories as a hermeneutic resource. *Energy research & social science* .

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic Differentiation in PyTorch. In *NIPS Autodiff Workshop.*

Pope III, C. A.; and Dockery, D. W. 2006. Health effects of fine particulate air pollution: lines that connect. *Journal of the air & waste management association* .

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* .

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition.*

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems.*

Rittel, H. W.; and Webber, M. M. 1973. Dilemmas in a general theory of planning. *Policy sciences* .

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* .

Sauermann, H.; and Franzoni, C. 2015. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences* .

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision.*

Shirk, J.; Ballard, H.; Wilderman, C.; Phillips, T.; Wiggins, A.; Jordan, R.; McCallie, E.; Minarchek, M.; Lewenstein, B.; Krasny, M.; et al. 2012. Public participation in scientific research: a framework for deliberate design. *Ecology and society* .

Sullivan, B. L.; Wood, C. L.; Iliff, M. J.; Bonney, R. E.; Fink, D.; and Kelling, S. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* .

Tian, H.; Li, W.; Ogunbona, P.; and Wang, L. 2015. *Single Image Smoke Detection*, chapter ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II. Springer International Publishing.

Töreyin, B. U.; Dedeoğlu, Y.; and Cetin, A. E. 2005. Wavelet based real-time smoke detection in video. In *2005 13th European Signal Processing Conference.* IEEE.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Nonlocal neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.*

Xu, G.; Zhang, Q.; Liu, D.; Lin, G.; Wang, J.; and Zhang, Y. 2019a. Adversarial Adaptation From Synthesis to Reality in Fast Detector for Smoke Detection. *IEEE Access* .

Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; and Wang, J. 2017. Deep domain adaptation based video smoke detection using synthetic smoke images. *Fire safety journal* .

Xu, G.; Zhang, Y.; Zhang, Q.; Lin, G.; Wang, Z.; Jia, Y.; and Wang, J. 2019b. Video smoke detection based on deep saliency network. *Fire Safety Journal* .

Yang, Z.; Shi, W.; Huang, Z.; Yin, Z.; Yang, F.; and Wang, M. 2018. Combining Gaussian Mixture Model and HSV Model with Deep Convolution Neural Network for Detecting Smoke in Videos. In *2018 IEEE 18th International Conference on Communication Technology (ICCT).* IEEE.

Yin, Z.; Wan, B.; Yuan, F.; Xia, X.; and Shi, J. 2017. A deep normalization and convolutional neural network for image smoke detection. *Ieee Access* .

Yuan, F.; Zhang, L.; Wan, B.; Xia, X.; and Shi, J. 2019a. Convolutional neural networks based on multi-scale additive merging layers for visual smoke recognition. *Machine Vision and Applications* .

Yuan, F.; Zhang, L.; Xia, X.; Wan, B.; Huang, Q.; and Li, X. 2019b. Deep smoke segmentation. *Neurocomputing* .

Zhang, Q.-x.; Lin, G.-h.; Zhang, Y.-m.; Xu, G.; and Wang, J.-j. 2018. Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images. *Procedia engineering* .