Deep Learning With Weak Supervision for Disaster Scene Description in Low-Altitude Imagery

Maria Presa-Reyes, *Student Member, IEEE*, Yudong Tao, *Graduate Student Member, IEEE*, Shu-Ching Chen, *Fellow, IEEE*, and Mei-Ling Shyu, *Fellow, IEEE*

***Abstract*—Pictures or videos captured from a low-altitude aircraft or an unmanned aerial vehicle are a fast and cost-effective way to survey the affected scene for the quick and precise assessment of a catastrophic event's impacts and damages. Using advanced techniques, such as deep learning, it is now possible to automate the description of disaster scenes and identify features in captured images or recorded videos to gain situational awareness. However, building a large-scale, high-quality dataset with annotated disaster-related features for supervised model training is time-consuming and costly. In this article, we propose a weakly supervised approach to train a deep neural network on low-altitude imagery with highly imbalanced and noisy crowd-sourced labels. We further make use of the rich spatiotemporal data obtained from the pictures and its sequence information to enhance the model's performance during training via label propagation. Our approach achieves the highest score among all the submitted runs in the TRECVID2020 Disaster Scene Description and Indexing (DSDI) Challenge, indicating its superior capabilities in retrieving disaster-related video clips compared to other proposed methods.**

***Index Terms*—Convolutional neural networks (CNNs), deep learning, disaster scene description, weak supervision.**

I. Introduction

A CATASTROPHIC event or accident may have disastrous implications, such as making some of the most impacted regions completely inaccessible due to outages in communication lines and disruptions to street-network infrastructures. Situational awareness in disaster-affected areas is critical for the safety and effectiveness of first responders. Remote sensing technology, such as aerial photography, is a viable solution to rapidly collect situational awareness information across the impacted regions while the regions remain inaccessible. Furthermore, the availability of trustworthy and accurate information is a crucial challenge for emergency management. However, the large volume of collected data and the limited time under a disaster scenario make it extremely challenging for a human to quickly identify regions that should be prioritized. Thus, it becomes crucial to develop automated systems to assist the emergency responders in analyzing the collected data and obtaining immediate situation awareness about the impacted regions.

Civil Air Patrol (CAP) supports U.S. communities going through an emergency response by taking pictures or recording videos from the low-altitude aircraft, which is a crucial and inexpensive method for the Federal Emergency Management Agency (FEMA) to quickly and effectively obtain the imagery to survey the affected region. The footage is often captured from military aircraft, primarily cargo aircraft, tankers, or helicopters [1], and, more recently, drones as well [2]. Given the massive volume of data being gathered, developing sophisticated tools and systems to curate all the information is also vital. To this end, several large-scale disaster datasets, including the Incidents Dataset [3], Low Altitude Disaster Imagery (LADI) [4], and so on, have been recently released to stimulate the development of new research and technologies in this field. However, the bird's eye view of the gathered data and the disaster-related application provides several challenges that must be appropriately addressed. Because individuals are unaccustomed to seeing and interpreting images taken at low altitudes, producing high-quality LADI annotations needs specific expertise. Meanwhile, nonprofessionals are not necessarily familiar with disaster-related concepts. As a result, collecting high-quality annotations to create an appropriate training dataset will be very costly.

While deep learning has greatly accelerated the advancement of image recognition capabilities, most of the existing techniques require large amounts of high-quality annotations to build high-performance and reliable models to properly automate image processing and concept detection [5]. Hence, they cannot meet the expectations for disaster situation awareness due to the lack of adequate training data [4]. Many techniques have been recently proposed to reduce the requirement of deep learning models on the quantity and quality of the training data. Such techniques include the semisupervised-based frameworks (such as deep cotraining [6]) that enable one to train the model with a partially annotated dataset and weakly supervised techniques, such as deep collaborative embedding model [7] that can handle mislabeled data. However, it remains challenging to train a deep learning model when both the quantity and quality of annotations are limited.

In this study, a weakly supervised deep learning framework is proposed that can manage noisy, limited, and inaccurate inputs while detecting descriptive features in connection to damage and the captured environment. Since the low-altitude imagery dataset, such as LADI, is partly labeled, the soft labels

Manuscript received June 13, 2021; revised September 22, 2021 and October 31, 2021; accepted November 15, 2021. Date of publication November 18, 2021; date of current version February 17, 2022. The work of Shu-Ching Chen was supported in part by NSF under Grant CNS-1952089 and Grant CNS-2125165. *(Corresponding author: Mei-Ling Shyu.)*

Maria Presa-Reyes and Shu-Ching Chen are with the Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199 USA (e-mail: mpres029@cs.fiu.edu; chens@cs.fiu.edu).

Yudong Tao and Mei-Ling Shyu are with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146 USA (e-mail: yxt128@miami.edu; shyu@miami.edu).

Digital Object Identifier 10.1109/TGRS.2021.3129443

defining the likelihood of an image possessing a given feature are propagated to the unlabeled data in the training dataset to enhance the training process. Furthermore, the proposed work demonstrates how spatiotemporal information obtained from the image's metadata can be leveraged to enrich the training dataset and improve model robustness. Spatiotemporal data, including the time and location of the picture taken, are used to query open-source databases for further context details in regards to the image. The main contributions of this article are summarized as follows.

1) We present a new semisupervised training method that utilizes labeled and unlabeled data by further employing label propagation and weak supervision to acquire knowledge from noisy, restricted, and inaccurate labels.
2) Multimodal information (such as geospecific tags, historical events, and weather) is merged with sequence-based information retrieved from low-altitude photographs to enrich the training dataset.
3) The proposed method is evaluated on the LADI dataset as one of the submitted runs in the TRECVID2020 [8] Disaster Scene Description and Indexing (DSDI) Challenge. Our proposed solution achieved the best score among all the participants and other appropriate methods from the literature.

The next sections of this article are structured in the following order. Section II reviews related studies that apply deep learning methods to low-altitude imagery. Section III introduces our proposed weakly supervised framework, specifically label propagation and feature fusion. In Section IV, the effectiveness of our proposed framework is shown through both quantitative and qualitative experimental results. Finally, Section V covers some of the potential future work and concludes this article.

II. Related Work

The advent of deep learning tools and techniques, especially the convolutional neural network (CNN) [5], [9], has revolutionized image and video recognition and greatly improved object detection accuracy and robustness. Considering how images and videos are a prevalent way for emergency responders to quickly survey affected areas after a natural disaster, it is no surprise that deep learning methods, such as CNNs, are being applied to automate the curation and retrieval of such images. Nonprofessionals may not be acquainted with disaster-related concepts or rarely come across low-altitude images. As a consequence, gathering enough high-quality annotations to build a good training dataset will be very expensive. However, most existing techniques require large volumes of high-quality annotations to develop reliable models that can help automate image processing and concept detection properly.

To reduce the reliance on the quality and quantity of training and testing data, researchers have developed a variety of deep learning techniques. Previous studies have proposed methods to integrate visual and text tag features into a common space based on deep canonical correlation analysis (DCCA) [10], [11]. The DCCA method has been extended to handle noisy labels and improve the annotation quality using techniques, such as the weakly supervised deep matrix factorization framework [12] and the unified deep collaborative embedding [7]. Many of the previously described techniques used sparse line reconstruction, sparse coding, and dictionary learning to recover textual tags, which takes a long time and takes up much space, making it unsuitable for large-scale applications.

Although research on automatic disaster scene description from images has become more prevalent in recent years, most of the existing approaches are confined to one disaster type. In addition, they are often incomparable due to the lack of well-curated datasets and benchmarks. This has recently changed with the introduction of large-scale disaster datasets, such as the Incidents Dataset [3] and LADI [4]. The Incidents Dataset is well-curated; however, it focuses on a ground-level perspective, which does not assist with the large intraclass variances shown on low-altitude images. Most of the previously proposed methods in the disaster scene description from both image and video data also focused on the ground-level point of view. These methods often aim to address challenges commonly found in the disaster image data by developing sophisticated models, such as adversarial data augmentation to deal with the limited data [13] or common techniques that put a higher penalty to errors on the minority class to address the class-imbalance issue [14], [15].

The LADI dataset features a wide range of low-altitude images and has presented a number of challenges, including noisy annotations with imbalanced samples per class and the fact that some objects and features are shown at different sizes and angles depending on the altitude at which the picture was taken, making some of these features difficult to detect. The earlier research on the disaster scene description from low-altitude images tested different supervised methods by considering the image's optical properties [16]–[18]. More recent studies started to explore an ensemble learning approach to tackle the class-imbalance and noisy-label issues [19]–[21].

Moreover, the previously proposed methods seldom leverage the rich spatiotemporal information from data and have yet to exploit the sequential-based information of the low-altitude images to improve the model performance during training. Our proposed framework leverages the weakly supervised deep learning approach with a unique label propagation model that enhances the training data as the model learns and uses the spatiotemporal information to improve the contextual awareness of the model.

III. Proposed Methods

In this article, a weakly supervised learning framework for disaster scene description, as illustrated in Fig. 1, is proposed to address the challenges that data labels are limited in both quantity and quality. In light of these limitations, classifiers pretrained on other well-curated benchmarks are leveraged to supply supervision signals by connecting their predicted concepts to the target features at the semantic level. Soft labels are created initially from human workers' annotations, and the more workers who annotate an image with a target feature, the greater the weight that is allocated to the image

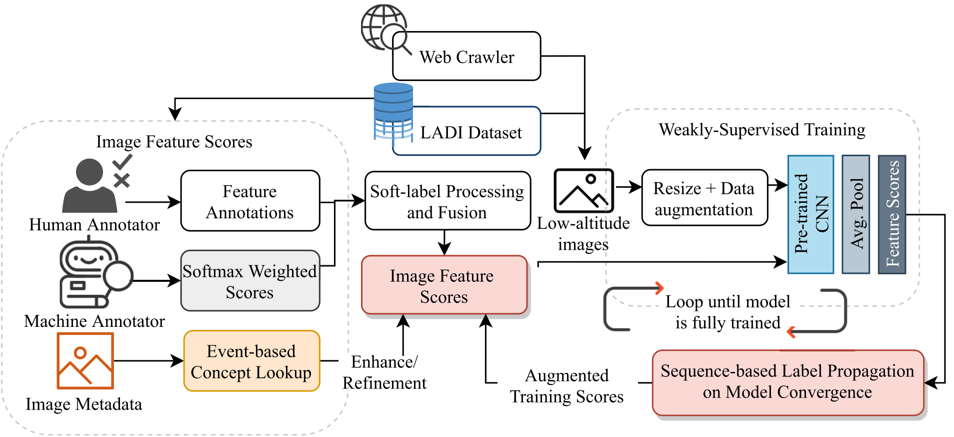


Fig. 1. Proposed weakly supervised deep learning framework implements feature score fusion and automatic score propagation based on the spatiotemporal information obtained from the metadata of low-altitude images.

under that target feature. These soft-label features are then fused with SoftMax weights of well-known deep-learning methods that have been pretrained with well-curated large image datasets. The final soft labels defining the likelihood of an image possessing a given feature are propagated throughout the training dataset to enrich the training data appropriately. While the deep learning-based model learns, it helps to identify more samples of a certain feature and expand the label set. The proposed method reduces the difficulty of obtaining well-curated expert hand-labeled datasets, which may be either expensive or unfeasible. Instead, low-cost weak labels are utilized with the goal that, though they may be flawed, they can still be leveraged to build a robust predictive model. The aim of the proposed approach is to estimate the likelihood of a certain disaster or environment-related feature being present inside a low-altitude image or video.

A. Feature Score Engineering

1) Worker Annotations: The LADI dataset uses a hierarchical labeling approach featuring five general categories, including *damage*, *environment*, *infrastructure*, *water*, and *vehicle*. Within each category, features of more specific categories are annotated. Using the Amazon Mechanical Turk (MTurk) service [22], a subset of the LADI dataset, representing more than forty thousand images, was hand-annotated by human annotators.

Assuming that the data are either labeled by nonexpert human workers through crowd-sourcing or obtained from a web crawler, label engineering is a critical initial step in reducing label noise and preventing erroneous labels from deceiving the model. Given an image i in the dataset, it may be labeled by one or more workers as containing a feature $\mathtt{f}$. However, not all the worker's labels can be treated with an equal level of confidence. Let $C_{i,\mathtt{f}}$ be the number of workers who labeled the image i as containing feature $\mathtt{f}$. Each image's feature score is $S_{i,\mathtt{f}} = (C_{i,\mathtt{f}} - C_{\mathtt{f}}^{\mathtt{min}})/(C_{\mathtt{f}}^{\mathtt{max}} - C_{\mathtt{f}}^{\mathtt{min}})$, where $C_{\mathtt{f}}^{\mathtt{min}}$ and $C_{\mathtt{f}}^{\mathtt{max}}$ are the minimum and maximum counts of workers for all the annotated images with feature $\mathtt{f}$. The soft-score function is formulated under the assumption that there will be at least one human worker annotating an image for target features under the same category. The assumption is that $C_{\mathtt{f}}^{\mathtt{max}} > 0$. Under this assumption, $C_{\mathtt{f}}^{\mathtt{max}} = C_{\mathtt{f}}^{\mathtt{min}}$ implies that all the positive samples are annotated by the same amount of workers. Thus, we assign the labels of all these samples as 1, i.e., $\forall i, S_{i,f} = 1$.

In our investigations, we employ these normalized soft-label vectors as the ground-truth confidence. Soft labels provide a model with more information about the relevance of each target feature. Such a strategy works well in a ranking problem scenario, provided that the goal is to help index the most relevant images. However, these crowd-sourced human labels are highly imbalanced. Because of the extreme disparity between different labeled samples, the calculated $S_{i,\mathtt{f}}$ may be imprecise for extremely underrepresented features. In addition, some images also bear incorrect or ambiguous labels. Hence, the dataset requires further enhancements through the addition of new data and new information.

2) Machine Annotations: The LADI dataset includes several machine-generated feature scores from commercial and open-source image recognition platforms to provide additional knowledge for various features found in the images. These feature scores are in the form of SoftMax weights that can be defined as follows:

$$\sigma(\hat{x})_i = \frac{\exp(x_i)}{\sum_{k=1}^{K} \exp(x_k)} \qquad (1)$$

where, given the input vector $\hat{x} = (x_1, \ldots, x_K) \in \mathbb{R}^K$, the equation applies a normalization term to output a probability distribution for K classes. The SoftMax weights are, thus, numerical scores indicating the relative confidence of the pretrained model in detecting the existence of a certain feature, and these machine annotations contain the names of the detected features, allowing us to match them with the features in LADI via semantic similarity.

The first machine annotator is a ResNet50 model [23] pretrained on Places365 [24], which includes 365 categories of common scenes. It is crucial to improve the efficacy of detecting the scenes and environments in the LADI imagery. Many features present in LADI are broad terms and can be

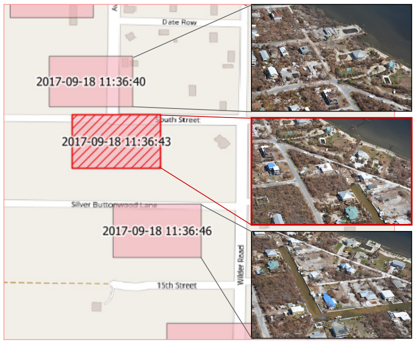


Fig. 2. Visualization of the polygons and an indicated timestamp of when the picture was captured following the sequence. The image highlighted in red has been annotated by the human worker for three features under the damage category (namely, flood water, misc., and rubble). Images taken before and after the annotated image are shown to also contain these same features.

matched with many available features in Places365 according to the semantic meaning of a feature. For instance, the concept for "building" has a close semantic meaning with the Places365 concepts, including *apartment building outdoor*, *basement*, *beach house*, *building facade*, *construction site*, *downtown*, *residential neighborhood*, *roof garden*, *skyscraper*, and so on. Therefore, many of the matched concepts could be safely regarded as "building" for disaster scene description.

Machine-generated annotations from Google Cloud Vision (GCV) [25] are also part of the machine annotators. The GCV API provides robust pretrained machine learning models for instantly assigning labels to images and classifying them into millions of preset categories. The scores from GCV *label detection* and *web entity detection* services are available for a subset of the LADI dataset.

We further made use of the predictions from the YOLOv4 [26] model pretrained on the COCO dataset [27]. The annotations supplied by the YOLOv4 model trained on COCO contain significant characteristics, such as car and truck, and have shown to be significant in improving the *vehicle* category model.

Information from other sources was retrieved for a subset of the features that were highly underrepresented. Crawling for more data helps to alleviate some of the training datasets' imbalanced problems and mistakes found within the labels. A small number of sample images under these features are crawled using an image search engine, such as Microsoft Bing Image[1], while also making sure the queries contain words, such as *drone* and *aerial* along with the target feature name. The noise from the crawled images is reduced by applying the CNN model trained on the human workers' limited labeled data and selecting the images that are most relevant to their target feature (i.e., score > 0.5). The scores from all relevant crawled images are then set to 1.

3) Metadata Concept Lookup: To include more concepts relevant to real-life events, we further utilize time and location metadata obtained from each image. The focal length (F), altitude (A), latitude, longitude, and camera type are all contained in the data that may be retrieved from an image's metadata, provided that the image's format supports the exchangeable image file format (Exif). This information is useful to approximate the geographical region covered by the image taken from an airplane. Although there is no direct access via Exif to the height $\mathtt{H}$ and width $\mathtt{W}$ of the camera sensor, the camera model provided by the metadata was used to acquire this information from other sources. The simple trigonometry can specify the current footprint through the computation of width $= (A \times \mathtt{W})/F$ and height $= (A \times \mathtt{H})/F$ of the geographic region. The measured area is only a rough approximation of the area photographed, as illustrated in Fig. 2, and is limited by its assumption that the camera takes the picture while being pointed directly downward. The angle from which the image was taken is a required parameter in order to be able to determine the exact geographical boundaries covered by the image. Nonetheless, several drones on the market today feature more detailed information regarding location captured and camera angle.

The image's time and location metadata offer multiple useful indications in relation to the image's contextual content. Additional information about the photographed region can be accessed from open datasets by considering the particular incidents, locations, and special weather conditions that may have been recorded at the time and place where the picture was taken. Open databases used for the retrieval of the images' contextual data are summarized as follows.

1) *OpenStreetMaps OSM:* The computed geographical region represented as a polygon is used to index open-source geodatabases OpenStreetMaps [28] that provide a valuable location-based context of the aerial images captured. OSM tags are crowd-sourced and describe specific features of map elements. The more area covered by the image, the more tags it contains in terms of buildings, roads, and so on. Bringing in the OSM data starts with collecting the total number of tags that fall inside the image's capture region and using a min–max normalization to provide more confidence over the images containing a higher number of a particular tag. Nonetheless, as illustrated in Fig. 2, there is a noticeable shift between the computed region and the actual area that is shown in the image. Despite the limitations of the OSM's geographic information-based approach, our proposed methods of combining several modalities help to generate more reliable scores. In addition, Fig. 3 demonstrates the logistic regression fit on the relationship between the LADI soft labels and the matching OSM scores, further supporting our hypothesis that there is a positive relationship between the relevant OSM tags and the target features that can be found in an image. The plot illustrates the most relevant target features that are semantically similar to the OSM tags. Section III-B details the matching procedure and aggregation method for the OSM scores.
2) *FEMA:* FEMA Disaster Declarations are an excellent resource for records of historical disasters in the United States, such as coastal storms, earthquakes, fires, floods, hurricanes, tornadoes, and volcanic activities. The FEMA data are used to confirm that the images

[1] Microsoft Bing Image: https://www.bing.com/images

TABLE I

SUMMARY OF THE CONFLICTS BETWEEN LADI'S INITIAL LABELS SPECIFIED BY HUMAN WORKERS AND THE LABELS RECTIFIED UTILIZING EXTERNAL DATABASES, NAMELY, FEMA AND NOAA CDO

Feature	P→P	P→N	U→P
(1) Damage (misc)	5146 (12.7%)	7839 (19.3%)	-
(2) Flood/water	12554 (30.9%)	4236 (10.4%)	-
(6) Smoke/fire	4 (0%)	1614 (4.0%)	100 (0.2%)
(12) Snow/ice	0 (0%)	115 (0.3%)	198 (0.5%)

in the LADI dataset annotated with the specific reported damage correspond to the actual real-life incident. If an annotator or pretrained model declares that a label contains certain damage caused by a type of disaster in contradiction with the FEMA records, the score for that feature is set to 0. Otherwise, the score's confidence is increased to 1. Table I summarizes the concordance and conflicts between the original worker labels and the labels rectified using the external databases, including FEMA. The positive-to-positive (P→P) category demonstrates the agreement that an image under a relevant target feature was annotated as positive. On the other hand, the conflicts between actual and rectified labels are shown through the positive-to-negative (P→N) category, meaning that the actual label was set to positive in conflict with the rectified negative label. In the unknown-to-positive (U→P) category, because of the very poor reliability of the original annotations, positive samples are drawn from the LADI unlabeled set for the target features, such as smoke/fire by using the information from the external databases. Later, in Section IV-A1, we show how a reliability measure further helps to explain the high degree of conflicts observed in the LADI label set for many target features, including smoke/fire and snow/ice.

3) *NOAA Climate Data Online (CDO):* NOAA allows for public access to the National Climatic Data Center's (NCDC) [29] data, which is an archive of global historical meteorological and climatic data. For features related to the climate, such as snow, weather details were obtained from the NCDC API using the time and location information. Similar to FEMA, the NCDC data are used to confirm that the annotation given to the image does not contradict the real-life event. For example, to ensure that a given image's snow feature is valid, the image's time and location are compared to the NCDC data, and the feature score is adjusted to 1 if snow has been reported or 0 otherwise. As shown in Table I, similar to the smoke/fire target feature, snow/ice was further enhanced through the addition of positive samples from the unlabeled set.

B. Final Score Fusion

The model is trained to identify a particular feature inside an image and its confidence level by employing soft labels. This also makes it easier to integrate soft labels generated from human annotations with SoftMax weights supplied by various pretrained classifiers assessed on the LADI dataset. Prior to the actual fusion of the scores, a semantic match of the feature's name in both LADI and the pretrained model's own feature list is formed. Semantic similarity borrows techniques from natural language processing (NLP), such as word embedding, to determine how similar two words are, even when they are not exact matches. The feature's word vectors were first generated using the pretrained model from spaCy [30] followed by a pairwise computation of the vector's cosine similarity [31]. Since some features are composed of two or more words, spaCy helps to identify the unique words (or tokens) within the feature name and generate the word vector. Considering that the features found in the LADI dataset are very broad in concepts, we take advantage of the variety provided by the scores generated from both the pretrained models and OSM tags. Selecting the relevant OSM tags is also done through semantic nearness by finding the similarity between two word vectors for the OSM tags and the target feature name in the vector space.

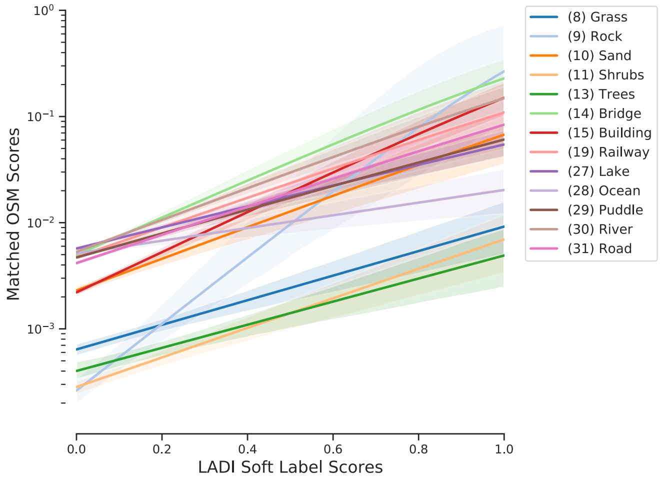


Fig. 3. Logistic regression plot and a 95% confidence interval of the relationship between the LADI soft labels and the matching OSM scores. Only the target features found to be semantically similar to the OSM tags are included in this plot.

Because the SoftMax weights are a probability distribution that awards the highest score to the best-detected classification in each image, a min–max normalization is applied before working on the final score fusion. Let $\hat{\omega}_i^{\mathtt{f}}$ represent the word vector of the ith word in the name that describes the target feature $\mathtt{f}$, and $\hat{\omega}_j^{\mathtt{p}}$ represents the word vector of the jth word in the name that describes the auxiliary feature $\mathtt{p}$ in either the pretrained classifiers or the OSM tags. All the word vectors are generated by spaCy [30]. The final score fusion is decided based on the distance of the word vectors, as shown in the following:

$$S_{\mathtt{f}}^{*} = \max\left(S_{\mathtt{f}}', \max_{\mathtt{p}\in\mathbb{P}_f} S_{\mathtt{p}}\right) \tag{2}$$

$$\mathbb{P}_f = \left\{ p \,\middle|\, \max_{i\in[1,N_{\mathtt{f}}],\, j\in[1,N_{\mathtt{p}}]} \frac{\hat{\omega}_i^{\mathtt{f}} \cdot \hat{\omega}_j^{\mathtt{p}}}{\left\|\hat{\omega}_i^{\mathtt{f}}\right\| \left\|\hat{\omega}_j^{\mathtt{p}}\right\|} > \vartheta \right\} \tag{3}$$

where $S_{\mathtt{f}}^{*}$ is the final score assigned to a certain image for feature $\mathtt{f}$, $S_{\mathtt{f}}' \in [0, 1]$ is the score of the image for target

Algorithm 1 Weakly Supervised Model Training Implementing Label Propagation

```
1: $M_0^1 \leftarrow M_S, r \leftarrow 100, \sigma \leftarrow \infty, \rho \leftarrow 5, i \leftarrow 1$
2: repeat
3:     $\rho_i \leftarrow 0$
4:     while $\rho_i < \rho$ do                              ▷ Train until convergence
5:         $M^* \leftarrow Train(M_{i-1}^1; X_t, Y_t)$
6:         $\rho_i \leftarrow \rho_i + 1$
7:         $\sigma_i \leftarrow Loss(M^*; X_v, Y_v)$          ▷ Using Equation 4
8:         if $\sigma_i < \sigma$ then                       ▷ Keep the best model
9:             $M_i^1 \leftarrow M^*, \sigma \leftarrow \sigma_i, \rho_i \leftarrow 0$
10:        end if
11:    end while
12:    for each feature $\mathtt{f}_j \in F$ do
13:        $\hat{j} \leftarrow TopScores(y_{i,j}, r)$         ▷ Fetch the top samples
14:        $X_{\hat{U}} \leftarrow NearestSamples(X, \hat{j})$
15:    end for
16:    $Y_{\hat{U}} \leftarrow Predict(M_i^1; X_{\hat{U}})$     ▷ Propagate the scores
17:    $X_t \leftarrow Merge(X_t, X_{\hat{U}}), Y_t \leftarrow Merge(Y_t, Y_{\hat{U}})$
18:    $i \leftarrow i + 1$
19: until Stop Condition is met                          ▷ Model is fully trained
```

feature $\mathtt{f}$ integrating the human annotations and machine annotations, $S_{\mathtt{p}} \in [0, 1]$ is the score integrating pretrained classifiers and OSM tags for the auxiliary feature $\mathtt{p}$, and $N_{\mathtt{f}}$ and $N_{\mathtt{p}}$ are the numbers of words that describe the target feature $\mathtt{f}$ and auxiliary feature $\mathtt{p}$, respectively. The final score fusion first checks that the cosine distance between $\hat{\omega}_i^{\mathtt{f}}$ and $\hat{\omega}_j^{\mathtt{p}}$ must be greater than a given threshold ϑ (0.5 in this study) for the concepts to be considered semantically similar. The final score for the target feature $\mathtt{f}$, i.e., $S_{\mathtt{f}}^*$, is thus the largest score among the original score $S'_{\mathtt{f}}$ and the scores of any auxiliary features whose names contain at least a word semantically similar to any word in the name of target feature $\mathtt{f}$, as illustrated in (2) and (3).

C. Weakly Supervised Training

Weak supervision discussed in this article is a strategy that learns from the partially annotated and noisy labels and the low-quality information from various data sources. Our proposed system combines a novel label propagation approach with a weakly supervised deep learning framework to improve the data quality as the deep learning model trains. The suggested technique aims to make acquiring well-curated expert hand-labeled datasets easier by using low-cost weak labels. Algorithm 1 illustrates the steps to train one of the categorical models in a weakly supervised approach via label propagation. The proposed approach extracts deep features from the final convolutional layer. It then outputs a feature vector corresponding to the 2-D picture using the InceptionV3 architecture, pretrained using the ImageNet weights. The pretrained weights of the networks have been completely fine-tuned to the new low-altitude image dataset. The model's original classification head is replaced with a dense layer followed by a sigmoid activation function to enable multifeature score prediction, indicating the likelihood that an image includes a certain feature.

The training process starts with the model M_0^1 initialized to the ImageNet weights M_S. From line 4, the model trains on the training dataset composed of the images X_t and scores Y_t, where X is the entire set of low-altitude images whether labeled or unlabeled, and $X_t \subseteq X$. In each epoch, the total loss $\mathcal{L}$ is calculated by aggregating the binary cross-entropy (BCE) loss across all the individual features as follows:

$$\mathcal{L}(p_{\mathtt{f}}, q_{\mathtt{f}}) = -\frac{1}{\mid F \mid}\sum_{\mathtt{f}\in F}(p_{\mathtt{f}}\log(q_{\mathtt{f}}) + (1 - p_{\mathtt{f}})\log(1 - q_{\mathtt{f}})) \tag{4}$$

where $p_{\mathtt{f}}$ is the probability (or soft label) of the image containing the target feature $\mathtt{f}$, and $q_{\mathtt{f}}$ is the predicted probability of the image containing $\mathtt{f}$ as calculated by the model. As the model trains, it is validated on the validation samples X_v in which its predicted values are compared to the target scores Y_v using the loss function. The variable σ keeps track of the lowest validation loss such that only the best model is kept at the end of the training process.

Once the model reaches a point where it is no longer improving for ρ consecutive epochs, the algorithm starts the label propagation process at line 12 by first sampling the top scores from each feature in X_t to later acquire the nearest samples. For r unique observations identified from X_t under a certain feature $\mathtt{f}$, the algorithm finds the nearest unclassified samples and stores them in $X_{\hat{U}}$. Scores are propagated throughout the identified image's neighboring images. The idea is that, if a picture taken at a particular moment contains a certain feature, the picture taken before and after it will most likely have the same feature. The training process is terminated when the *Stop Condition* is met, i.e., the model is considered to be fully trained. The stop condition is defined as the model not being improved after two consecutive label propagations.

IV. Experiment Results

A. Experimental Setup

1) Dataset: This article uses the LADI dataset that consists of pictures captured from a low-flying aircraft by CAP and hosted by FEMA. The National Institute of Standards and Technology (NIST)'s TREC Video Retrieval Evaluation (TRECVID) competition released the dataset to participants in the middle of 2020. The LADI training dataset is a collection of pictures acquired from an aircraft, whereas the LADI test dataset is a collection of short video clips recorded from a UAV. According to the LADI developers [4], [32], each Human Intelligence Task (HIT) on the MTurk platform asks the human worker if any of the labels in each of the coarse categories is accurate—each HIT only asks about one category at a time. Consequently, each HIT is assigned to three workers to reach an agreement on the label quality. If further validation was required, the HIT was outsourced to two more workers, for a total of five workers per category and image.

With only about 6%–7% of the 500k images in the LADI training dataset being labeled by human workers, we tackle

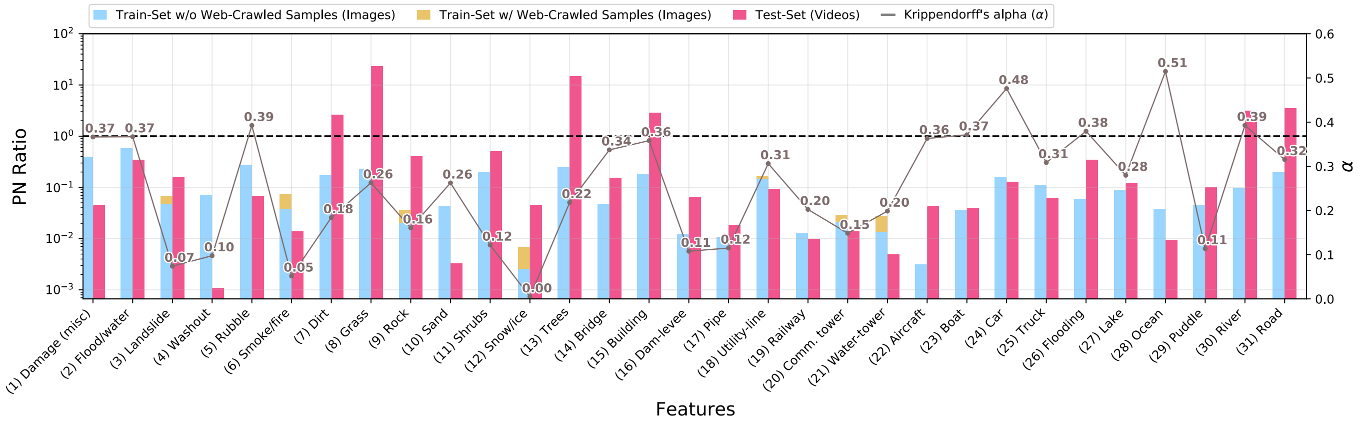


Fig. 4. PN ratios of the 31 target features from the LADI training and testing datasets before any of the proposed rectification techniques was applied. The reliability coefficient Krippendorff's alpha (α) indicates the measure of the agreement among the workers when annotating the training dataset for each target feature.

several challenges that arise from working with a highly imbalanced and noisy dataset to train a reliable model. The datasets class imbalance and label noise are illustrated in Fig. 4. The positive-to-negative (PN) ratios of the 31 target features calculated from LADI illustrate how the training dataset varies from mostly severely imbalanced (low PN ratios) to reasonably balanced (high PN ratios) representations. We further compute Krippendorff's alpha (α) [33], a well-known reliability metric used to measure interrater agreement for the annotated labeled training dataset. Unlike other reliability techniques, α handles missing data and is flexible in sample size, category, and the number of workers. The α coefficient is calculated following LADI's described annotation procedure, in which each worker will have the chance to determine whether an image contains a target feature or not. If an annotator never comes across a certain image, this is declared as a missing value. The maximum value for each target feature's α coefficient can only reach $\alpha = 0.51$, indicating the need to correct the label noise and augment the training dataset.

The LADI training dataset was further expanded by including web-crawled images for those very underrepresented features, such as water-tower, utility-lines, communication tower, snow/ice, rock, landslide, and smoke/fire. Less than 1000 images for each feature were added from web-crawling to improve these features and avoid adding more noise into the data. While crawling for new pictures helps retrieve more relevant samples, the web-crawled pictures may add extra noise into the training data if not utilized properly. In addition, it is challenging to acquire high-quality low-altitude pictures from the image engine for many of these target features, considering that people seldom take photographs from this viewpoint. Thus, the web-crawled pictures were not used to balance the training dataset, and the impact of the additional crawled images on the PN ratio is shown in Fig. 4. Data augmentation methods were used on the training data to improve the model performance, especially for the minority classes. Specifically, the applied augmentation methods include horizontal and vertical flipping with 0.5 probability, 90° rotations with 0.1 probability, contrast change by a factor between 0 and 0.25, and the horizontal and vertical shifts within ±10% of the width and height, respectively.

2) Competing Methods: To ensure that the suggested method is effective, we compare it to a baseline and several competing methods, which are listed as follows.

1) *DCCA [10]:* The DCCA employs neural networks to exploit the nonlinear transformations and learns the representations of images and texts that maximize their correlations.
2) *DCE [7]:* The deep collaborative embedding employs a weak supervision technique for refining initial tags and assigning tags to new images via discovering the unified latent space for images and tags.
3) *SHIELD [21]:* Experimented with various CNN combinations on the LADI dataset and extended the LADI training data by labeling an unlabeled subset from LADI using Amazon MTurk.
4) *VCL [18]:* Perform a series of experiments to evaluate the roles that objects play in scene comprehension, utilizing various methods for integrating the local-level information (e.g., objects and entities).
5) *Ours-InceptionV3-Base:* The baseline model consists of five categorical models based on Inception-V3, but, different from the proposed approach, it is trained solely on the soft labels generated from the human annotators, following the method introduced in Section III-A1.

3) Feature Score Model: The feature score model consists of five categorical models based on the InceptionV3 architecture, with each model being trained on the feature scores of a particular category. The weights of these models are pretrained on ImageNet and then fine-tuned on the disaster-related dataset following the transfer learning process. The last classification head of the network replaces a dense layer implementing the sigmoid activation function for multiclass soft-label classification. The binary cross-entropy function measures the model loss during training and adjusts the model weights accordingly. Separating each model by category gives more flexibility and alleviates the high class-imbalance problem in the data.

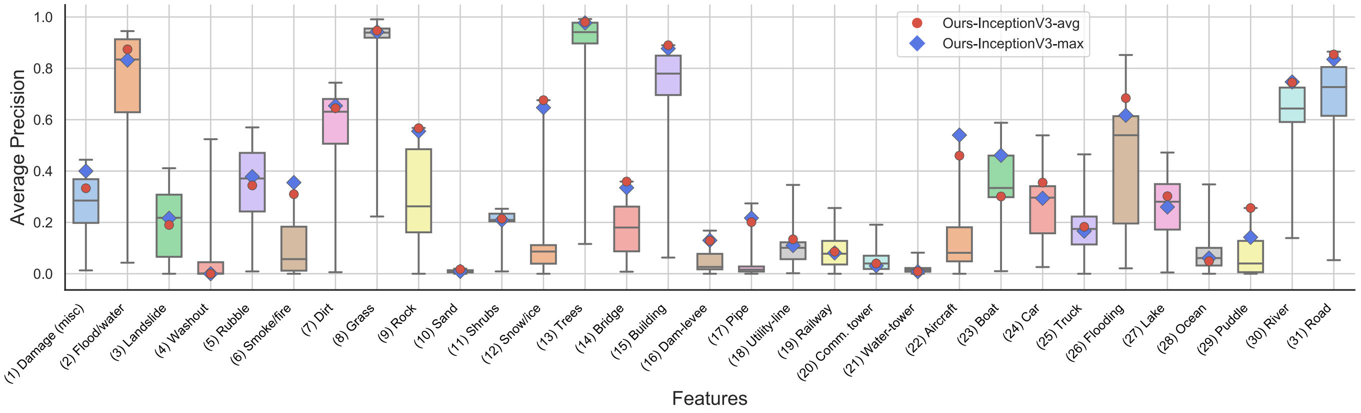


Fig. 5. Comparison of the boxplot distribution for feature's precision score among all submissions to TRECVID2020-DSDI regardless of the track. The interquartile range of the boxplot is from 25th to 75th percentiles. The red dot indicates the placement of our best run among all the submissions. The blue diamond indicates our second-best run.

TABLE II

Comparing the Mean Precision at 10, 100, and 1000 Precision Depths, Along With the MAP of Our Suggested Methodology, of Our Proposed Approach to Various Competing Methods and a Baseline

Method	P@10	P@100	P@1000	MAP
DCCA [10]	0.177	0.196	0.210	0.167
DCE [7]	0.329	0.282	0.238	0.205
SHIELD [21]	0.506	0.379	0.236	0.297
VCL [18]	0.232	0.218	0.225	0.176
	0.400	0.346	0.260	0.275
	0.355	0.369	0.264	0.285
	0.471	0.394	0.272	0.333
Ours-InceptionV3-base	0.445	0.404	0.274	0.283
Ours-InceptionV3-top	0.568	0.446	0.278	0.388
Ours-InceptionV3-max	**0.580**	0.444	0.279	0.390
Ours-InceptionV3-avg	0.561	**0.460**	**0.281**	**0.391**

The LADI data are randomly split into two parts: 80% for training and 20% for validation. Each model is trained on small batch sizes containing 16 sample images from the training set. At the end of each epoch, the model's performance is evaluated on the validation set—only the best model with the lowest validation loss is kept. With an initial learning rate of $\eta = 1e - 4$, the Adam [34] solver is employed to fine-tune the model weights. Models are trained for 100 epochs.

4) Inference and Ranking: The LADI test dataset is composed of 41 original videos segmented into 1825 short video clips ranging between two to twenty seconds. Unlike the training dataset, the test set is composed majorly of drone footage. Nonetheless, our methods prove successful in generalizing well across the different tools used to capture the low-altitude images. During inference, the test video shots are split into multiple unique keyframes and fed to the five categorical models to obtain the scores for the 31 features. In order to facilitate content-based retrieval, a shot-level aggregation of the keyframe-level scores is then introduced to rate the video shot according to its significance.

B. Results and Discussion

1) Quantitative Results: For each run, the mean average precision (MAP) across 1000 retrieved shots is determined as a measure of the accuracy in identifying the most relevant features in a shot. For a fair comparison among other methods tested on the LADI dataset, variants of the proposed framework are compared to the submission of the LADI + Others (O) track—where "Others" in our proposed approach involves the inclusion of data obtained from the web crawler along with the LADI data to improve the performance of some of the most underrepresented features. A summary of the result comparison is demonstrated in Table II. The proposed method significantly outperforms other tested methods under the same training type. It is worth mentioning that our best-performed approach ranks first among all the solutions in the TRECVID2020-DSDI competition, regardless of the training type.

Furthermore, the average precision (AP) per feature is summarized in Fig. 5. A boxplot is used to visualize the distribution of the feature-level performance across all competition entries, independent of the training type. The red dot and blue diamond represent our best and our second-best submissions, respectively. The comparisons to the feature-level measurements reveal that our proposed method excels in snow/ice, bridge, building, road, and puddle features. We attribute the great performance for most of these features to the effective exploitation of the contextual information derived from the image's metadata. For instance, infrastructure locations, such as roads and buildings, are well-documented in OSM, which our proposed method was able to effectively leverage to refine and enhance the soft labels used to train the model to recognize these types of target features.

The proposed approach is also compared to a baseline model. It is noteworthy that the baseline method already achieves a comparable performance to other proposed techniques, confirming that the proposed technique to calculate soft labels from human-workers' annotations effectively reduces some of the noise in the feature scores. By applying the proposed feature fusion with label propagation, we can see the improvements made compared to the final score and each feature's score, as shown in Fig. 6. Very significant

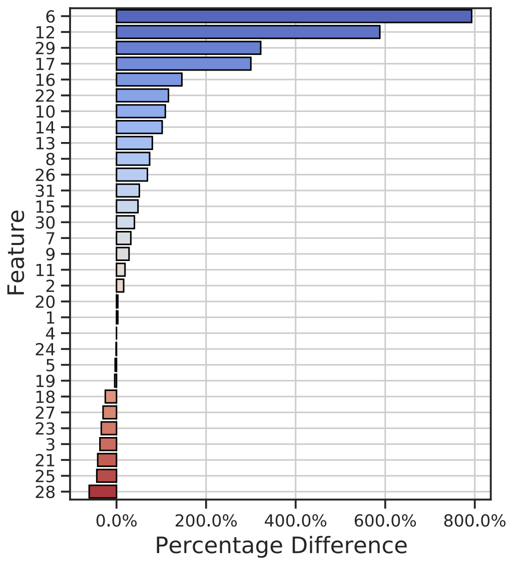
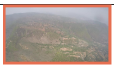
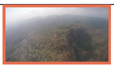
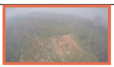
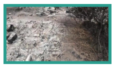
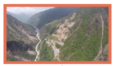
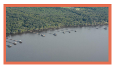
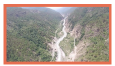
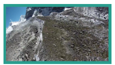
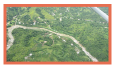
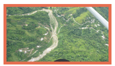
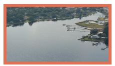


Fig. 6. Percentage difference between each feature's AP values from both the baseline and our proposed method. The feature IDs are aligned with those in Fig. 5.

TABLE III

QUALITATIVE RESULTS OF THE FIRST FIVE VIDEO CLIPS RETRIEVED BY THE BASELINE AND THE PROPOSED METHOD. BELOW EACH VIDEO'S SCREENSHOT, THE CHECK (✓) INDICATES ITS RELEVANCE TO THE FEATURE, WHILE THE CROSS (✗) MARKS THOSE VIDEOS THAT HAVE BEEN INCORRECTLY RETRIEVED AS FALSE POSITIVES

		Retrieval				
Smoke/Fire	Baseline	 ✗	 ✗	 ✗	 ✗	 ✗
	Proposed	 ✓	 ✓	 ✓	 ✓	 ✗
Snow/Ice	Baseline	 ✗	 ✗	 ✗	 ✗	 ✗
	Proposed	 ✓	 ✓	 ✓	 ✓	 ✓
Shrubs	Baseline	 ✗	 ✗	 ✗	 ✗	 ✓
	Proposed	 ✓	 ✓	 ✗	 ✗	 ✗

improvements can be observed for most of the features. More notably, smoke/fire and snow/ice demonstrated improvements of 792% and 587%. By virtue of the proposed methods, time and location data were effectively used in instances where labels were extremely limited. Certain features, including landslide and lake, performed worse due to the overlap and ambiguity in some feature definitions—further discussed in Section IV-B2.

2) Qualitative Results: Table III illustrates the qualitative results from the top five videos retrieved by the baseline and the proposed method for some features, where the video's keyframe that achieves the highest score for that particular feature is displayed. Because the LADI's target features are so broad, there are many variations in what may be regarded as valid observations within each feature target. Moreover, many of the features have vague meanings that may overlap. The false positives obtained from the retrieval further demonstrate the ambiguity and broadness that are present in some of the features' meanings and how these limitations have affected the results. As can be seen from this table, for the smoke/fire feature, the baseline method retrieved many examples of environments surrounded by "fog," which means that the model might have identified some characteristics in "fog" to be very similar to "smoke." However, we also observe that the proposed approach's fifth retrieved observation is a false positive, as the model most likely misconstrues the feature of "dust" for smoke. Despite the ambiguity and broadness of the feature, the proposed method significantly improves the retrieval performance for smoke/fire by incorporating the historical data of the relevant real-life events. The snow/ice is another feature that benefits a lot from matching the historical data in NOAA's CDO database using time and location from the training image's metadata as queries. The uncertainty and overlap in these feature definitions may be seen more clearly in the case of the feature, shrub. A shrub is a kind of foliage that might be difficult to tell apart from a tree from afar. Although our proposed framework effectively retrieves more relevant videos with shrubs, some of the videos categorized as not relevant may include shrubs as well, or it is not easy to discern.

V. CONCLUSION AND FUTURE WORK

It is now feasible, more than ever, to dispatch a drone ahead of the rescue crew to inspect the impacted region and aid responders to automatically identify those areas that are the most affected and should be prioritized to deliver a timely and appropriate response. The proposed framework aims to predict the chance of a certain catastrophe or environment-related characteristic being present inside a low-altitude snapshot or video. This article introduces a weakly supervised deep learning approach developed for automatic disaster scene description of low-altitude pictures captured from an aircraft. The proposed approach is also intended to cut down the time and effort that human annotators spend labeling images. As part of our future work, we will continue to work on improving the model's performance by further analyzing the image's sequential characteristics. The proposed approaches will also be evaluated on additional comparable datasets that may be noisy, imbalanced, or lacking ground truth.

REFERENCES

[1] B. Lewis, "Civil air patrol offers local support," *TechBeat*, pp. 10–13, Mar. 2014. [Online]. Available: https://www.ojp.gov/ncjrs/virtual-library/abstracts/civil-air-patrol-offers-local-support

[2] *Civil Air Patrol Begins Deploying Small Drones for Search and Rescue*. Accessed: Nov. 2019. [Online]. Available: https://www.airforcemag.com/civil-air-patrol-begins-deploying-small-drones-for-search-and-rescue/

[3] E. Weber *et al.*, "Detecting natural disasters, damage, and incidents in the wild," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 331–350.

[4] J. Liu, D. Strohschein, S. Samsi, and A. Weinert, "Large scale organization and inference of an imagery dataset for public safety," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Sep. 2019, pp. 1–6.

[5] S. Pouyanfar *et al.*, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–36, Sep. 2018.

[6] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proc. ECCV*, 2018, pp. 142–159.

[7] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2070–2083, Sep. 2018.

[8] G. Awad *et al.*, "TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains," in *Proc. TRECVID*. Gaithersburg, MD, USA: NIST, 2020, pp. 1–55.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2017.

[10] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[11] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," in *Proc. MLSLP*, 2012, pp. 34–37.

[12] Z. Li and J. Tang, "Weakly supervised deep matrix factorization for social image understanding," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 276–288, Jan. 2016.

[13] S. Pouyanfar *et al.*, "Unconstrained flood event detection using adversarial data augmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 155–159.

[14] S. Kim, H. Kim, and Y. Namkoong, "Ordinal classification of imbalanced data with application in emergency and disaster information services," *IEEE Intell. Syst.*, vol. 31, no. 5, pp. 50–56, Sep./Oct. 2016.

[15] S. Pouyanfar *et al.*, "Dynamic sampling in convolutional neural networks for imbalanced data classification," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 112–117.

[16] J. Mao, K. Harris, N.-R. Chang, C. Pennell, and Y. Ren, "Train and deploy an image classifier for disaster response," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Sep. 2020, pp. 1–5.

[17] E. Sava, L. Clemente-Harding, and G. Cervone, "Supervised classification of civil air patrol (CAP)," *Natural Hazards*, vol. 86, no. 2, pp. 535–556, Mar. 2017.

[18] E. Christakis, S. Demertzis, K. Stavridis, A. Psaltis, A. Dimou, and P. Daras, "Towards low-altitude image analysis: Object-enhanced concept detection," in *Proc. TRECVID*. Gaithersburg, MD, USA: NIST, 2020, pp. 1–5.

[19] Y. Li, H. Wang, S. Sun, and B. Buckles, "Integrating multiple deep learning models to classify disaster scene videos," in *Proc. IEEE High Perform. Extreme Comput. Conf.*, Jun. 2020, pp. 1–7.

[20] S. Okazaki, Q. Kong, M. Klinkigt, and T. Yoshinaga, "Hitachi at TRECVID DSDI 2020," in *Proc. TRECVID*. Gaithersburg, MD, USA: NIST, 2020, pp. 1–8.

[21] M. Zaffaroni, F. Oldani, and C. Rossi, "Independent category classifiers for emergency scene description using deep learning approaches," in *Proc. TRECVID*. Gaithersburg, MD, USA: NIST, 2020.

[22] K. Crowston, "Amazon mechanical turk: A research tool for organizations and information systems scholars," in *Shaping the Future of ICT Research. Methods and Approaches*, A. Bhattacherjee and B. Fitzgerald, Ed. Berlin, Germany: Springer, 2012, pp. 210–221.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2017.

[25] D. Mulfari, A. Celesti, M. Fazio, M. Villari, and A. Puliafito, "Using Google cloud vision in assistive technology scenarios," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2016, pp. 214–219.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[27] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[28] C. Bansal *et al.*, "Characterizing the evolution of Indian cities using satellite imagery and open street maps," in *Proc. 3rd ACM SIGCAS Conf. Comput. Sustain. Societies*, Jun. 2020, pp. 87–96.

[29] NOAA Climate Data Online. *National Climatic Data Center (NCDC)*. Accessed: Jun. 2020. [Online]. Available: www.ncdc.noaa.gov/

[30] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," Explosion, Berlin, Germany, Tech. Rep., 2020, doi: 10.5281/zenodo.1212303.

[31] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A comparison of semantic similarity methods for maximum human interpretability," in *Proc. Artif. Intell. Transforming Bus. Soc. (AITB)*, vol. 1, Nov. 2019, pp. 1–4.

[32] J. Liu and A. Weinert. (2019). *Low Altitude Disaster Imagery (LADI) Dataset*. [Online]. Available: https://github.com/LADI-Dataset/ladi-overview

[33] K. Krippendorff, "Computing Krippendorff's alpha-reliability," Univ. PA ScholarlyCommons, Harrisburg, PA, USA, Tech. Rep. 43, 2011. [Online]. Available: https://repository.upenn.edu/asc_papers/43

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–15.

Maria Presa-Reyes (Student Member, IEEE) received the B.Sc. degree in computer science from Florida International University (FIU), Miami, FL, USA, in 2015, where she is currently pursuing the Ph.D. degree in computer science.

She is currently a Research Assistant with FIU. Her research interests include geospatial data mining, machine learning, and multimedia.

Yudong Tao (Graduate Student Member, IEEE) received the B.Sc. degree in microelectronics from Fudan University, Shanghai, China, in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering (ECE) with the University of Miami (UM), Coral Gables, FL, USA.

He is currently a Research Assistant with UM. His research interests include data mining, machine learning, and multimedia.

Shu-Ching Chen (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering (ECE) from Purdue University, West Lafayette, IN, USA, in 1998.

He is currently a Professor with the Knight Foundation School of Computing and Information Sciences, Florida International University (FIU), Miami, FL, USA. He has authored or coauthored more than 360 research papers in journals and refereed conference/symposium/workshop proceedings, book chapters, and three books.

Dr. Chen is also a fellow of American Association for the Advancement of Science (AAAS), Society for Information Reuse and Integration (SIRI), and Asia-Pacific Artificial Intelligence Association (AAIA). He was named a 2011 recipient of the ACM Distinguished Scientist Award.

Mei-Ling Shyu (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering (ECE) from Purdue University, West Lafayette, IN, USA, in 1999.

She is currently a Professor with the Department of ECE, University of Miami (UM), Coral Gables, FL, USA. She has authored or coauthored two books and more than 300 research papers in journals and refereed conference/symposium/workshop proceedings, and book chapters.

Dr. Shyu is also a fellow of American Association for the Advancement of Science (AAAS), The American Institute for Medical and Biological Engineering (AIMBE), Society for Information Reuse and Integration (SIRI), and Asia-Pacific Artificial Intelligence Association (AAIA). In 2012, she was awarded the Computer Society Technical Achievement Award and the ACM Distinguished Scientist Award.