Multi-Source Weak Supervision Fusion for Disaster Scene Recognition in Videos

Maria Presa-Reyes[1], Yudong Tao[2], Rui Ma[2], Shu-Ching Chen[1], and Mei-Ling Shyu[2]

[1]Knight Foundation School of Computing and Information Sciences,
Florida International University, Miami, Florida 33199
[2]Department of Electrical and Computer Engineering,
University of Miami, Coral Gables, Florida 33146

Abstract

Images or video recordings assist emergency responders in quickly inspecting the damage after a disaster event. New techniques are needed to help responders organize and find important information at the right time. However, most existing methods don't meet public safety standards mainly due to a lack of training data. We propose a multi-source weak supervision fusion technique to train on a highly imbalanced dataset annotated with noisy labels. Using a Confident Learning technique, we reduce the effect of the noise while boosting the quality of the class labels. We combine the predictive power from models trained on large-scale visual datasets using Differential Evolution. This research demonstrates a fully-automatic approach with great potential to reduce required time and resources while delivering exceptional results. In the TRECVID2021 Disaster Scene Description and Indexing (DSDI) Challenge, our technique achieved the top score among all the submitted runs, independent of the training data utilized.

1 Introduction

Image and video recognition algorithms are rapidly advancing, growing faster and with better precision, and are expected to become a critical component of incident and disaster response [10]. Using advanced technology and deep learning methodologies such as Convolutional Neural Networks (CNN), it is possible to deploy a drone ahead of the search team to swiftly identify the most damaged areas that should be prioritized during a disaster. The automated content-based analysis and classification of observed disaster-related features in recorded videos will allow better curation and retrieval of critical information for situational awareness. Due to insufficient training data and standards, the bulk of existing methods does not fulfill public safety demands.

Civil Air Patrol (CAP) has the technical capability to function even when severe weather disrupts power, the internet, phones, and airplane takeoffs, making it a critical and cost-effective tool for the Federal Emergency Management Agency (FEMA) to survey the impacted region swiftly and efficiently. CAP offers aerial pictures of flooded areas, collapsed dams, and other natural disaster-related events. To this end, several large-scale disaster imagery datasets, including the Incidents Dataset [18], LADI (Low Altitude Disaster Imagery) [12], xBD [8], etc., have been recently released to stimulate the development of new research and technologies in this field. Given the volume of data being collected, it is also critical to develop sophisticated tools and systems for curating all of the information.

The images taken by low-altitude planes are challenging to analyze because they have a low height perspective, an oblique angle, and many disaster-related parts that image recognition systems don't usually take into account. We propose a weakly-supervised learning technique that incorporates data from a range of sources, many of which are of low quality or have been trained on subjects significantly different from the target classification task. The proposed fully-automatic solution would significantly decrease the time and expense associated with classification jobs while delivering superior outcomes.

The main contributions of this paper are summarized as follows.

- We propose a new semi-supervised training technique that is robust to noisy, limited, and erroneous annotations and class labels from multiple sources.

- For multi-source weak supervision fusion framework, a unique approach for recognizing and merging the relevant predictions from various pre-trained networks is proposed.
- The proposed method is evaluated on the LADI dataset achieving top score among all the submitted runs in the TRECVID2021 [2] Disaster Scene Description and Indexing (DSDI) Challenge, independent of the training data utilized.

The next sections of the paper are structured in the following order. Section 2 examines works that use deep learning techniques to analyze low-altitude images. Section 3 introduces our proposed weakly-supervised framework, including confident learning and multi-source weak-supervision fusion. In Section 4, the effectiveness of our proposed framework is shown through quantitative experimental results. Finally, Section 5 outlines the paper and recommends future research.

2 Related Work

Most current solutions rely on high-quality annotations to build reliable models that can sufficiently automate image processing and concept detection. Non-experts are likely to have only seen low-altitude photos on rare occasions. Consequently, it will be too costly to get enough high-quality annotations to build a good training dataset. Numerous researchers have developed a variety of deep learning algorithms that are less reliant on the quality of training data. The weakly-supervised tag and visual information are used to train semantic-aware hash functions [17]. Previously, deep canonical correlation analysis (DCCA) [1] was used to combine visual and text tag data. Many previously reported techniques rely on sparse line reconstruction, sparse coding, and dictionary learning to recover textual tags, which costs time and space and is not suited for large-scale applications. Research into automated disaster scene descriptions from images has grown in popularity. Newly-released disaster datasets such as xBD [8] and the Incidents Dataset [18] feature a top-down and a ground-level view of the damages. However, LADI [12] is unique in the low-altitude and oblique views found in its images. More recent studies explore an ensemble learning approach to tackle the class-imbalance and noisy-label issues [11, 14]. The incorporation of spatio-temporal information to increase the model's contextual awareness has also been investigated [15]. Our proposed framework aims to improve the quality of noisy labels in the LADI training data through a Confident Learning (CL) [13] strategy. Furthermore, a novel multi-source information fusion is proposed to improve the performance of target features that are underrepresented in LADI.

3 Proposed Framework

Figure 1 illustrates the full flow of our proposed framework. CL is used to improve the quality of the noisy labels in the crowdsourced annotated training set, which is the first step in our multi-source architecture for combining weak supervision from different sources. Given the scores of numerous semantic concepts obtained from different machine annotators, several semantically related predictions are used to improve the performance of a target feature. The text that describes the target feature is turned into high-dimensional vectors, which are then used to look for semantic similarity and pick relevant concepts from other networks. We optimize a weighted average that incorporates all of the models' relevant predictions into a single scalar that serves to rank the video clip using Differential Evolution (DE).

3.1 Denoising with Confident Learning

According to the LADI researchers [12], annotations are organized as Human Intelligence Task (HIT) which asks the human worker whether any of the target features in each of the five categories (i.e., *damage*, *environment*, *infrastructure*, *water*, and *vehicle*) are correct. Each HIT is allocated to up to five workers (asking just one category at a time) in order to reach agreement on label quality. Namely, for an image i, and a target feature F that belongs to a specific category C (i.e. $F_C \in C$), the initial soft score $S_{i,F}$ is calculated as follows.

$$S_{i,F_C} = \frac{\#Positive\ Votes_{i,F_C}}{Total\ Votes_{i,C}} \tag{1}$$

To calculate the $S_{i,F}$, we assume a particular image must have at least one vote from an annotated who was assigned a specific category C. Then, we employ cross-validation confident intervals [13] to derive out-of-sample prediction probabilities from improving the label quality further.

3.2 Multi-source Weak Supervision Fusion

3.2.1 Machine Annotators

This study employs four CNN network configurations (i.e., ResNet50, DenseNet161, YOLOv4, ViT-B/16, and InceptionV3) pre-trained on four open-source datasets (i.e., Places365, Incidents Dataset, MS COCO, ImageNet21k, and LADI+Others). ResNet50 [9] and DenseNet161 are pre-trained on Places365 dataset which contains 1.8 million training images taken from 365 scene categories [19]. Another ResNet50 network is also pre-trained on the Incidents Dataset containing 446,844 manually annotated images covering 43 incidents across various scenes [18]. YOLOv4 (You Only Look Once) [4] pre-trained on Microsoft Common Objects in Context (MS COCO) is one of the leading

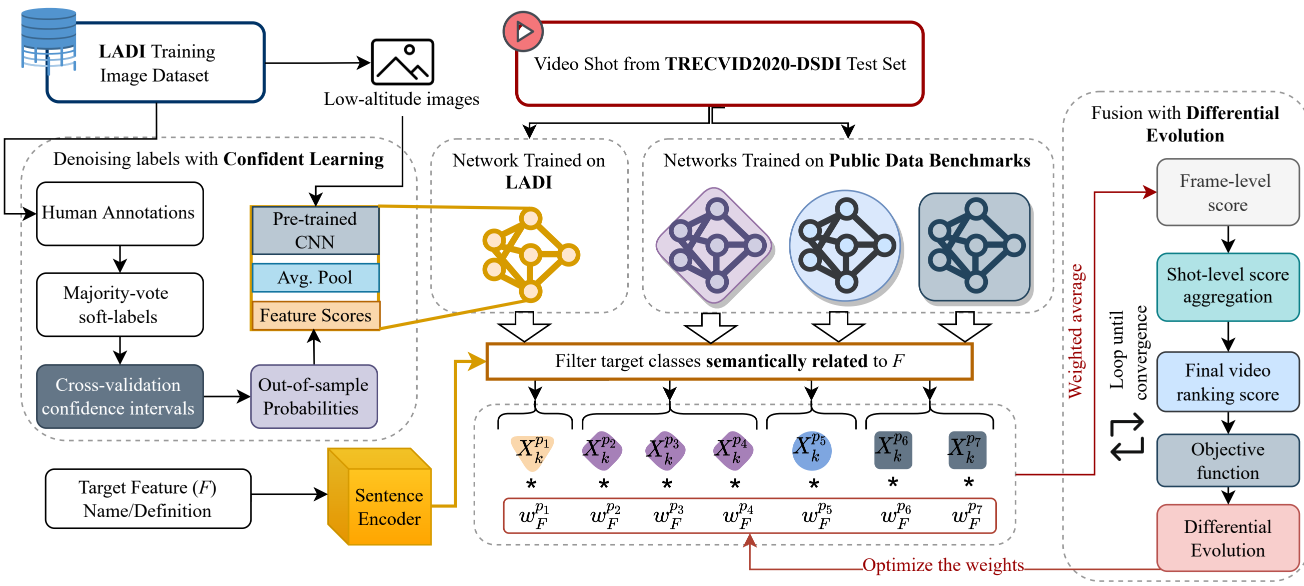

Figure 1. The proposed weakly-supervised deep learning framework implements a confident learning approach to denoise crowdsourced annotations along with a multi-modality fusion framework to search and combine relevant target features predicted by multiple networks.

deep learning-based object detection frameworks. The ViT-B/16 [7] model pre-trained on the ImageNet21K dataset is proven to be a key component in our proposed framework. Last but not least, an InceptionV3 model trained on LADI plus other sources by Presa-Reyes *et al.* [15] have also been incorporated.

3.2.2 Multi-Source Concept Fusion

Given the predicted scores X^p of many machine annotators' semantic concepts (i.e., target classes), numerous of these related concepts may help identify a target feature F. A Universal Sentence Encoder based on the Deep Averaging Network (DAN) [5] converts text describing the target feature into high-dimensional vectors T that is then utilized to the obtained semantic similarity among different concepts using the cosine distance θ of the vectors. To fuse multi-source concepts, the high-dimensional vectors of the target feature F and the semantic concept P are first matched, and the weighted average score of those closely correlated concepts are fused, i.e.,

$$S_F(k, w_F) = \sum_{p \in O} w_F^p \cdot X_k^p \tag{2}$$

where $O = \{P|\theta(T_F, T_P) > \vartheta\}$, and $w_F^p \in \mathbb{Q}$ is the set of optimized weights representing the contributing power of each pre-trained model's predicted score X_k^p for a key frame k. Moreover, the values for w_F^p bounds and the ϑ threshold are empirically decided based on validation performance. Furthermore, there exist multiple key frames inside a given video shot v. Therefore, the average score over all the key frames in v is computed as the shot-level feature score, which can be formally written as

$$S_V(V, w_F) = \frac{1}{||V||} \sum_{k \in V} S_F(k, w_F) \tag{3}$$

Then, for a given dataset of video shots $\mathcal{V}$ and a target feature F, the the top-N shot with F can be defined an ordered sequence $V_F = [V_1, V_2, \dots, V_N]$, where $V_i \in \mathcal{V}$ and $\forall i > j, S_V(V_j, w_F) > S_V(V_i, w_F)$.

3.2.3 Weight Optimization based on Differential Evolution

The remaining problem is to determine the optimal weights w_F for each target feature F. Differential Evolution (DE) is a kind of evolutionary optimization technique that works with a population of candidate solutions. It uses genetic operators like mutation and recombination to repeatedly enhance the population. The objective function G determines each candidate's fitness. If $G(s_1) < G(s_2)$, candidate s_1 is judged superior to candidate s_2. The objective function seeks to improve the average precision for a specific target feature (i.e., minimize $1 - AP^N$) by measuring the performance of a collection of retrieved results using the precision and recall metrics. Assuming the solution contains N video shots ordered by final aggregated confidence score, our objective is to minimize the following error formula:

$$\hat{w}_F = \arg\min_{w_F} G(w_F) = \arg\min_{w_F} \left[1 - AP^N(V_F)\right] \quad (4)$$

Semantically relevant predictions are combined into a single scalar, which is then used to score the video clip.

4 Experiment Results

4.1 Experimental Setup

4.1.1 Dataset

We test our methods using the LADI dataset, which comprises images acquired by CAP from a low-flying aircraft and maintained by FEMA. The LADI training dataset consists of images captured from an airplane, and the LADI test dataset consists of brief video clips captured from a UAV. The DSDI track's test dataset in 2021 comprises 2,802 video shots with a maximum duration of 60 seconds per shot, focusing on the devastation wrought by an earthquake tragedy. The test set supplied in TRECVID2020-DSDI [3] is used as validation during the DE processing in our case. The Mean Average Precision (MAP) metric is used to examine and compare the performance of different approaches.

4.1.2 Competing Methods

To determine the effectiveness of the proposed technique, we compare it to other competing methods, such as BUPT_MCPR [2] and VCL_CERTH [6]. Both competing methods are trained solely on the LADI-dataset. In particular, the problem was approached by VCL_CERTH as a panoptic segmentation problem with additional instance and semantic segmentation annotations for 300 LADI images. The method proposed by Presa-Reyes *et al.* [15] trained on LADI plus other datasets was also included. Two baseline fusion techniques using the average of the best performing model (Ours-CL-BA) and aggregated predictive scores after z-score normalization (Ours-CL-ZS) are also explored to compare against our proposed DE fusion.

4.1.3 Feature Score Model and Fusion

Two feature score models, EfficientNet-B5 [16] and ResNet50 [9], are trained on the LADI's confident labels generated by the CL-based approach. Using transfer learning, we fine-tune the network's weights on ImageNet. The network's final classification head is replaced with a fully-connected layer followed by a sigmoid activation for multi-class soft-label classification. With a starting learning rate of ($\eta = 1e-4$), we use the Adam solver to optimize our model. For the differential evolution search, we employed the DE/best/1/bin technique which generates new candidate solutions by randomly picking solutions from the population, subtracting one from the other, and adding a scaled version of the difference to the population's best candidate solution.

4.2 Results and Discussion

The proposed framework is compared to competing methods mainly categorized as LADI-based (L) the LADI + Others (O) track submission—where "Others" in our proposed approach refers to the inclusion of models pre-trained on open-source data benchmarks. Table 1 summarizes the performance comparison across different methods. The excellent results obtained by the panoptic segmentation approach proposed by VCL_CERTH on the LADI-based (L) track underline the necessity to integrate additional information about the image other than the noisy labels.

Our proposed technique achieves impressive results on the LADI + Others (O) track, particularly compared to other competing methods. The high recall rate illustrates our classification model's ability to detect and recover the majority of positive examples within a relevant target feature. By comparing the baseline methods Ours-CL-BA and Ours-CL-ZS, we demonstrate our proposed Ours-CL-DE approach can find better weights when aggregating predictions of different models. Furthermore, compared to simply training on LADI-based (L), the proposed method introduced on LADI + Others (O) exhibits a considerable improvement of about 34% in MAP score, indicating the effectiveness of our strategy of fusing the weak supervision from multiple sources.

In Figure 2, the average precision at the target feature level shows that our suggested approach has obtained the greatest performance for target features such debris, rock, snow/ice, building, utility-line, boat, river, and road. Figure 3 depicts the performance contribution of each additional dataset used to train the machine annotators previously described in Section 3.2.1. Starting from our proposed CL technique trained on LADI only, each additional dataset is added to the ensemble as depicted by a checkmark in the figure. The ResNet50 pre-trained on the Incidents Dataset contributed a performance boost for the environment category, detecting concepts such as 'snow covered' and 'field' and improving on features snow/ice and grass. Damage features, on the other hand, did not improve as expected given the damage concepts from the Incidents Dataset, necessitating further investigation. Environment features achieve better performance because they are simpler to discern from long distances and show lower inter-class variation than other categories. YOLOv4 network pre-trained on MS COCO contributes a performance

Table 1. Performance comparison among our proposed technique and competing methods.

Method	Training Data	Precision@k			Recall@k			F1@k			MAP
		k=10	k=100	k=1000	k=10	k=100	k=1000	k=10	k=100	k=1000	
BUPT_MCPRL [2]	L	0.271	0.225	0.228	0.271	0.232	0.405	0.271	0.227	0.244	0.159
VCL_CERTH [6]	L+	**0.510**	0.367	0.245	**0.511**	0.415	0.378	**0.510**	0.377	0.255	0.282
Presa-Reyes *et al.* [15]	O	0.413	0.392	0.285	0.413	0.448	0.682	0.413	0.404	0.316	0.298
Ours-CL-BA	L	0.394	0.346	0.279	0.394	0.383	0.648	0.394	0.351	0.307	0.254
Ours-CL-ZS	O	0.471	0.409	0.296	0.471	**0.522**	0.789	0.471	0.425	0.332	0.339
Ours-CL-DE (proposed)	L	0.384	0.351	0.286	0.384	0.395	0.683	0.384	0.360	0.315	0.268
	O	0.481	**0.425**	**0.310**	0.481	0.502	**0.793**	0.481	**0.439**	**0.345**	**0.359**

L+ LADI-based (L) training data plus additional human annotations (i.e., instance and segmentation).

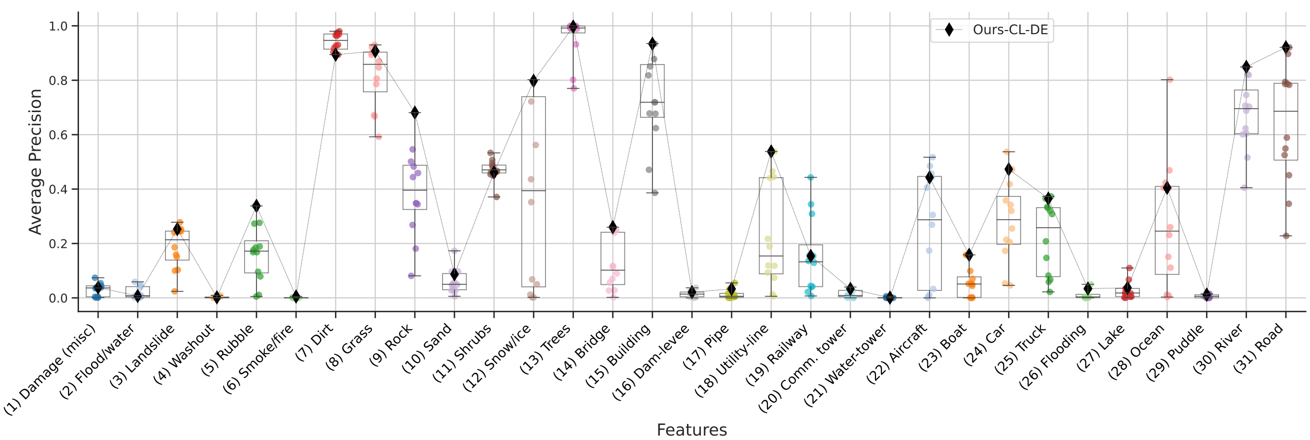

Figure 2. The boxplot shows the distribution for a feature's precision score compared across all submissions to TRECVID2021-DSDI, independent of which training dataset was used to train each technique. The placement of our proposed method's performance is demonstrated using a black diamond.

boost for vehicle categories, detecting concepts for 'aeroplane,' 'boat,' 'car,' and 'truck.'

We employ a weighted average ensemble achieving better performance thanks to the integration of human and machine-generated annotations. Since it is clear the relationship between the relevant features is not linear to their semantic similarity, our proposed technique has been proven to be a viable approach to identifying the best predictions based on the performance of each machine annotator. Because our proposed technique outperforms existing methods with minimal training, they are an excellent means of leveraging and transferring information from the methods that have already been presented in previous research into any emerging topic.

5 Conclusion and Future Work

Due to a lack of appropriate training data, most present-day picture recognition algorithms fail to meet public safety requirements. As part of our multi-source weak supervision fusion architecture, we apply the CL technique to enhance the quality of noisy labels in the crowdsourced annotated training set. Semantic similarity is used to identify relevant concepts predicted by other networks. We use DE to rank the video clip based on a weighted average of all relevant model predictions. Combining many classifiers pre-trained on well-known data standards improves overall performance, but only the best and most relevant predicted score towards a particular target feature should be used. Overall, the study shows how this framework has great potential to save a significant amount of time and resources while still achieving outstanding results in the disaster scene description task. Although this work focuses on disaster scene description, the proposed methods have been developed with extendability in mind. Our approaches are effective for leveraging and transferring knowledge from past study into any new topic. As a potential future work, we will explore more advanced techniques of incorporating other multi-modality sources using our proposed technique, such as spatio-temporal data.

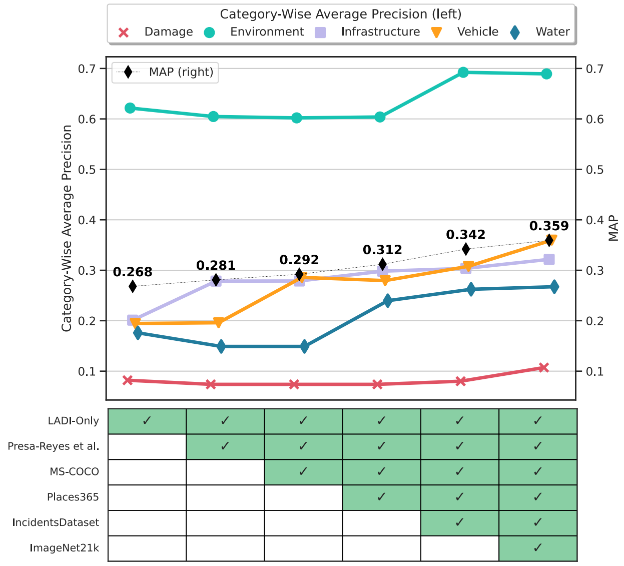

Figure 3. Ablation research demonstrating how the performance of the proposed Ours-CL-DE improves with the inclusion of each machine annotator's dataset.

Acknowledgment

This research is partially supported by NSF CNS-1952089 and CNS-2125165.

References

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[2] G. Awad, A. A. Butt, K. Curtis, J. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. Diduch, J. Liu, Y. Graham, G. J. F. Jones, and G. Quénot. Evaluating multiple video understanding and retrieval tasks at trecvid 2021. In *TRECVID*. NIST, USA, 2021.

[3] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, and G. Quénot. TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *TRECVID*. NIST, USA, 2020.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[5] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.

[6] E. Christakis, Z. Batzos, K. Konstantoudakis, K. Christaki, P. Boutis, D. Sainidis, D. Tsiakmakis, G. Almpanis, T. Dimou, and P. Daras. Low altitude image analysis using panoptic segmentation. In *TRECVID*. NIST, USA, 2021.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] R. Gupta, B. Goodman, N. N. Patel, R. Hosfelt, S. Sajeev, E. T. Heim, J. Doshi, K. Lucas, H. Choset, and M. E. Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. *ArXiv*, abs/1911.09296, 2019.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] M.-F. R. Lee and T.-W. Chien. Artificial intelligence and internet of things for robotic disaster response. *2020 International Conference on Advanced Robotics and Intelligent Systems (ARIS)*, pages 1–6, 2020.

[11] Y. Li, H. Wang, S. Sun, and B. Buckles. Integrating multiple deep learning models to classify disaster scene videos. In *2020 IEEE High Performance Extreme Computing Conference*, 2020.

[12] J. Liu, D. Strohschein, S. Samsi, and A. Weinert. Large scale organization and inference of an imagery dataset for public safety. In *IEEE High Performance Extreme Computing Conference*, pages 1–6, Sep. 2019.

[13] C. Northcutt, L. Jiang, and I. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[14] S. Okazaki, Q. Kong, M. Klinkigt, and T. Yoshinaga. Hitachi at TRECVID DSDI 2020. In *TRECVID*. NIST, USA, 2020.

[15] M. Presa-Reyes, Y. Tao, S.-C. Chen, and M.-L. Shyu. Deep learning with weak supervision for disaster scene description in low-altitude imagery. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2021.

[16] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[17] J. Tang and Z. Li. Weakly supervised multimodal hashing for scalable social image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 28:2730–2741, 2018.

[18] E. Weber, N. Marzo, D. P. Papadopoulos, A. Biswas, A. Lapedriza, F. Ofli, M. Imran, and A. Torralba. Detecting natural disasters, damage, and incidents in the wild. In *The European Conference on Computer Vision*, August 2020.

[19] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.