On Optimizing Model Generality in AI-based Disaster Damage Assessment: A Subjective Logic-driven Crowd-AI Hybrid Learning Approach

Yang Zhang, **Ruohan Zong**, **Lanyu Shang**, **Huimin Zeng**, **Zhenrui Yue**, **Na Wei**, **Dong Wang**
University of Illinois Urbana-Champaign
{yzhangnd, rzong2, lshang3, huiminz3, zhenrui3, nawei2, dwang24}@illinois.edu

Abstract

This paper focuses on the AI-based damage assessment (ADA) applications that leverage state-of-the-art AI techniques to automatically assess the disaster damage severity using online social media imagery data, which aligns well with the "disaster risk reduction" target under United Nations' Sustainable Development Goals (UN SDGs). This paper studies an *ADA model generality* problem where the objective is to address the limitation of current ADA solutions that are often optimized only for a *single* disaster event and lack the *generality* to provide accurate performance across *different* disaster events. To address this limitation, we work with domain experts and local community stakeholders in disaster response to develop *CollabGeneral*, a subjective logic-driven crowd-AI collaborative learning framework that integrates AI and crowdsourced human intelligence into a principled learning framework to address the ADA model generality problem. Extensive experiments on four real-world ADA datasets demonstrate that CollabGeneral consistently outperforms the state-of-the-art baselines by significantly improving the ADA model generality across different disasters.

1 Introduction

The increasing frequency and severity of natural disaster events (e.g., hurricanes, earthquakes, wildfires) have posed serious challenges to human society with significant casualties and enormous economic losses [McEntire, 2021]. For example, the recent Turkey–Syria earthquake has directly impacted 23 million people across Turkey and Syria along with over $50,000$ deaths and \$80 billion loss.[1] Damage assessment is an essential process during disaster response that aims to acquire accurate and timely information about the damages caused by the disaster, assist local/federal authorities (e.g., FEMA, public health agencies, police departments), and civil society stakeholders (e.g., regional red cross society, community disaster experts, local NGOs) in their decision making process, and prevent further damages [Kankanamge *et al.*, 2020]. The recent advances in AI have enabled more effective and scalable AI-driven solutions for timely disaster damage assessment, which aligns well with the "disaster risk reduction" objective under "Sustainable Cities and Communities" (i.e., Goal 11 of United Nations' Sustainable Development Goals (UN SDGs)). Meanwhile, the proliferation of social media also provides a pervasive data source to obtain real-time situation awareness of disaster events from common citizens [Wang *et al.*, 2019a]. In this paper, we focus on the *AI-based damage assessment (ADA)* application, where the goal is to leverage the advanced AI techniques (e.g., deep convolutional network, graph neural network, transformer) to automatically assess the disaster damage severity using social media imagery data [Nguyen *et al.*, 2017]. Specifically, we focus on addressing the limitation of current ADA models that are often optimized only for a *single* disaster event and lack the *generality* to provide accurate performance across *different* disaster events. We refer to such a knowledge gap as the *ADA model generality* problem.

Recent progress in AI and deep learning have been made to improve the performance of ADA applications [Imran *et al.*, 2022; Li *et al.*, 2019; Mouzannar *et al.*, 2018; Zhang *et al.*, 2021]. Current solutions often focus on designing *tailored* AI models that can accurately identify the event-specific damage visual characteristics for a *specific* disaster event to ensure accurate ADA performance [Li *et al.*, 2019]. We refer to the optimized performance of a customized ADA model for a specific disaster event as the model's *specificity* to that event. However, we observe that current ADA solutions often lack *model generality*, which leads to poor performance when the model is applied to a disaster event that is different from the one on which the model was trained [Zhang *et al.*, 2021]. For example, in Figure 1, we observe that (A) and (B) share similar visual features of grey sky but end up with completely different damage severity levels. An ADA model trained for the wildfire event mistakenly identifies the image in (B) as severe damage, and an ADA model trained for the hurricane event incorrectly classifies the image in (A) as no damage due to the lack of generality of the AI models. A possible solution to address the ADA model generality problem is to simultaneously train multiple ADA models, one for each specific disaster event [Li *et al.*, 2019]. However, a critical problem is that such a solution requires a good

[1] https://www.redcross.org.uk/stories/disasters-and-emergencies/world/turkey-syria-earthquake

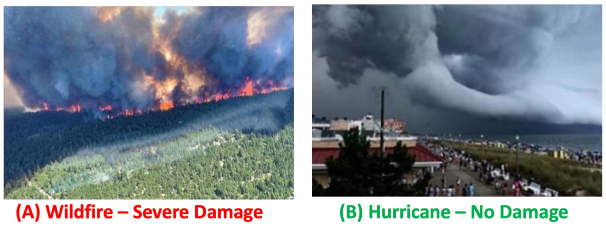

Figure 1: Illustrations of Lacking Generality for ADA Model

amount of high-quality training data from each event, which is not always available [Kumar *et al.*, 2020]. The lack of event-specific training data often leads to an overfitting issue where the trained ADA models fail to learn the event-specific damage visual features for the studied event, leading to an undesirable performance loss of the events that lack training data [Saunders, 2022].

In this paper, we jointly explore the different yet complementary AI and human intelligence from crowdsourcing systems such as Amazon Mechanical Turk[2] (i.e., *crowd intelligence*) to address the ADA model generality problem. Unlike the AI models that often rely on the training data available in each disaster event to generate the estimated ADA labels, humans can often reasonably estimate the ADA labels across different disaster events without the need for training data from such events [Fuchs, 2022; Zhang *et al.*, 2022a]. For example, in Figure 1, we can clearly understand that the grey sky in (A) indicates large smokes caused by a wildfire but reflects a heavy cumulonimbus in (B). As a result, crowd intelligence is often more generalizable in identifying disaster damages across different disaster events. However, unlike the AI models that can maintain a certain level of consistency and accuracy once they are trained for a specific disaster event, crowd workers often make inadvertent mistakes on the imagery data from a disaster event [Draws *et al.*, 2021]. Motivated by the above observations, this paper develops a hybrid crowd-AI collaborative learning framework that jointly leverages the *specificity of AI* and the *generality of human intelligence* to address the ADA model generality problem. However, two technical challenges exist in designing our framework.

The first challenge is how to effectively optimize the ADA model generality without sacrificing its specificity on an individual disaster event. A possible solution to tackle the ADA model generality problem is to train an ADA model using the training data from *all* studied disaster events so that the ADA model instance can learn the disaster-related visual features from all trained events. However, such a one-size-fits-all solution can lose the sensitivity on the event-specific visual features and lead to undesirable performance loss on specific events of interest [Ghifary *et al.*, 2016]. On the other hand, recent efforts have been made to tackle the AI model generality problem [Zhang *et al.*, 2020; Sankaranarayanan *et al.*, 2018]. Those solutions often leverage the divergence-based or adversarial-based neural network designs to transfer or adapt the ADA model learned from a source event (with sufficient training data) to a target event (with little training data) that shares similar damage visual characteristics with the source event. However, the actual ADA performance largely depends on the level of similarity between the source and target events, and an appropriate source event is not guaranteed to exist [Zhang *et al.*, 2021].

The second challenge is how to effectively integrate the complementary yet different AI and crowd intelligence to address the ADA model generality problem. In particular, a few recent crowd-AI collaborative systems have been developed to leverage crowd intelligence to troubleshoot the AI failure cases or retrain the AI models to boost the application performance [Sener and Savarese, 2018; Zhang *et al.*, 2019]. However, those solutions are not designed to address the AI model generality problem and can lead to suboptimal performance when they are directly applied to different disaster events [Risi and Togelius, 2020]. In addition, we observe that the imperfect crowd labels could confuse the AI models to identify incorrect visual features that might further impair the model generality across different events [Rolnick *et al.*, 2017; Zhang *et al.*, 2022b]. Moreover, there also exist efforts in active learning and label aggregation that can be applied to fuse the inputs from both AI and crowd intelligence [Gemalmaz and Yin, 2021; Hube *et al.*, 2019]. Those solutions often leverage analytical approaches (e.g., Bayesian optimization, maximum likelihood estimation) to boost the aggregated label accuracy. However, those approaches do not jointly model the specificity of AI and the generality of crowd intelligence to achieve the optimal ADA performance on individual disaster events.

To address the above challenges, this paper develops *CollabGeneral*, a subjective logic-driven crowd-AI collaborative learning framework that exploits AI and crowd intelligence to address the ADA model generality problem. To address the first challenge, we develop a novel deep model optimization framework that designs a generality-aware network optimization function design to optimize the trade-off between the ADA model's generality and specificity. To address the second challenge, we model the crowd intelligence and AI through a novel subjective logic framework to fuse the intelligence from both humans and AI. We work closely with domain experts and local community stakeholders in disaster response to provide the in-the-field know-how to validate our framework. To the best of our knowledge, CollabGeneral is the first crowd-AI hybrid approach to tackle the AI model generality problem in ADA applications. We also envision that our framework can be applied to address the AI model generality problem in a much broader set of AI-driven applications beyond ADA (e.g., misinformation detection, intelligent transportation, smart health). We evaluate CollabGeneral through four real-world ADA datasets and the results demonstrate that CollabGeneral consistently outperforms state-of-the-art deep neural networks, crowd-AI models, and AI model generality frameworks by improving ADA model generality under a rich set of evaluation scenarios.

2 Related Work

AI-based Disaster Informatics: AI-based disaster informatics received a good amount of attention in recent years due to

[2] https://www.mturk.com/

its efficiency, scalability, and effectiveness in providing accurate and timely information during the disaster events [Sun *et al.*, 2020; Zhang *et al.*, 2023]. From disaster risk and damage identification [Metaxa-Kakavouli *et al.*, 2018] to emergency response and recovery [Soden and Owen, 2021], AI-based disaster informatics plays a vital role in reducing the negative impacts of natural disasters [Soden and Palen, 2018]. Disaster damage assessment using social media data is an important application in AI-based disaster informatics, where timely observations of the disaster damages from social media posts are leveraged to obtain the situational awareness (e.g., damage severity, casualties and injuries) during devastating natural disasters [Zade *et al.*, 2018]. For example, Ning *et al.* designed a deep convolutional network based context information extraction framework that explores the real-time Twitter posts to identify highly impacted areas and track the dynamic damage severity during flood disasters [Ning *et al.*, 2020]. Mangalathu *et al.* developed a recurrent neural network model that analyzes earthquake-related social media posts to estimate building damage severity in earthquake events [Mangalathu and Burton, 2019]. Barmpoutis *et al.* designed a convolutional neural network framework that leverages the multimodal social media posts to detect fire regions [Barmpoutis *et al.*, 2019]. However, current ADA models often focus on optimizing the performance for a single disaster event and lack model generality when performing ADA tasks on different events. In contrast, we design a crowd-AI hybrid learning solution to tackle the ADA model generality problem in AI-based disaster informatics.

AI Model Generality: The lack of generality is a fundamental issue in AI applications, and recent efforts have been made to improve the generality for AI models [Li *et al.*, 2020; Sankaranarayanan *et al.*, 2018; Kini *et al.*, 2021; Wang *et al.*, 2019b; Chan *et al.*, 2018; Zong *et al.*, 2023]. For instance, Li *et al.* proposed an unsupervised domain adaptation solution that incorporates a cluster-based regularization to improve the image classification performance across different domains [Li *et al.*, 2020]. Kini *et al.* designed a vector-scaling optimization framework that leverages a multiplicative logit adjustment mechanism to improve the cross-domain model generality in domain-sensitive image classification [Kini *et al.*, 2021]. Wang *et al.* developed a symmetric cross-entropy optimization framework that leverages a counterpart reverse optimization design to minimize the domain-wise overfitting problem and improve the model generality in natural scene classification [Wang *et al.*, 2019b]. However, current AI model generalization solutions often sacrifice the AI model's specificity when optimizing the generality of the model. To the best of our knowledge, CollabGeneral is the first crowd-AI hybrid learning framework that explicitly leverages the complementary nature of AI and crowd intelligence to optimize the ADA model generality without sacrificing the ADA performance in each studied disaster event.

3 Problem Description

Definition 1. Disaster Event ($\boldsymbol{D}$): We define $\boldsymbol{D} = \{D_1, D_2, ..., D_T\}$ to be a set of studied disaster events where D_t represent t^{th} studied disaster event, and T is the total number of disaster events in the studied ADA application.

Definition 2. Disaster-related Imagery Data ($\boldsymbol{X}$): We define $\boldsymbol{X} = \{X_1, X_2, ..., X_I\}$ to be a set of disaster-related social media imagery data posted during different disaster events for the ADA application (e.g., Figure 1). In particular, X_i indicates the i^{th} image sample, and I represents the total number of studied image samples. In addition, we define $\boldsymbol{X_{D_t}}$ to be the subset of image samples in $\boldsymbol{X}$ collected from the t^{th} studied disaster event D_t.

Definition 3. Class Label Estimated by AI ($\widehat{\boldsymbol{Y^A}}$): This work focus on the physical status based disaster scene classification in ADA applications. For instance, in a prior ADA study [Nguyen *et al.*, 2017], the disaster damage severity is categorized into three different classes: no/minor damage, medium damage, and severe damage. In particular, we define $\widehat{\boldsymbol{Y^A}}$ as the set of class labels estimated by the AI model for all imagery data $\boldsymbol{X}$, where $\widehat{Y_i^A}$ represents the estimated class label for X_i.

In our paper, we focus on exploring both AI and crowd intelligence to tackle the ADA model generality problem. Therefore, we further define a few key definitions on acquiring the crowdsourcing-based human intelligence.

Definition 4. Crowd Intelligence Query (Q): We define Q as a crowdsourcing task to acquire human intelligence from crowd workers. In particular, our CollabGeneral framework focuses on identifying a subset of image samples in $\boldsymbol{X}$ that the AI models fail to provide accurate estimation results due to their lack of generality [Ren *et al.*, 2021]. The identified image samples are forwarded to a crowdsourcing platform where each image in Q is annotated by a set of B crowd workers. We define $\boldsymbol{W} = \{W_1, W_2, ..., W_B\}$ as the set of crowd workers participating in Q. W_b indicates the b^{th} crowd worker. We present the details of crowd intelligence query tasks in Section 4.

Definition 5. Class Label Annotated by Crowd Workers ($\widehat{\boldsymbol{Y^W}}$): We define $\widehat{\boldsymbol{Y^W}}$ as the set of class labels annotated by crowd workers for the imagery data that are selected in Q. In addition, we define $\widehat{\boldsymbol{Y^{W_b}}}$ as the set of class labels contributed by a crowd worker W_b where $\widehat{Y_i^{W_b}}$ is the class label contributed by W_b for image sample X_i in Q.

Definition 6. Class Label Identified by CollabGeneral ($\widehat{\boldsymbol{Y}}$): We define $\widehat{\boldsymbol{Y}}$ as the final outputs of the CollabGeneral framework by leveraging the class labels returned by AI (i.e., $\widehat{\boldsymbol{Y^A}}$) and crowd workers (i.e., $\widehat{\boldsymbol{Y^W}}$). Specifically, $\widehat{Y_i}$ represents the final identified class label for image sample X_i.

The goal of our ADA model generality problem is to utilize the collaborative strengths of AI and crowd intelligence to achieve the optimal ADA performance in each studied disaster event as follows:

$$\underset{\widehat{\boldsymbol{Y_{D_t}}}}{\arg\max} \left(\Pr(\widehat{\boldsymbol{Y_{D_t}}} = \boldsymbol{Y_{D_t}} \mid \boldsymbol{X}, Q)\right), \forall\, D_t \in \boldsymbol{D} \tag{1}$$

where $\widehat{\boldsymbol{Y_{D_t}}}$ and $\boldsymbol{Y_{D_t}}$ indicate the *estimated* and *ground-truth* class labels for imagery data $\boldsymbol{X_{D_t}}$ from disaster event D_t, respectively. Note that, instead of learning an individually tailored ADA model for each studied disaster event, our CollabGeneral framework leverages the image samples $\boldsymbol{X}$ from all events to learn a holistic and accurate ADA model that generates the optimized results for each event.

4 Solution

CollabGeneral is a crowd-AI hybrid learning framework that integrates AI and crowd intelligence to optimize model generality in AI-based disaster damage assessment applications.

1) Generality-aware Deep Optimization (GDO): it designs a novel deep model optimization scheme that effectively learns a set of ADA model instances to achieve a good trade-off between AI model generality and specificity through a novel generality-aware network optimization design. The learned ADA model instances are then used to identify the subset of image samples for crowd intelligence query.

2) Subjective logic-driven Crowd-AI Fusion (SCF): it develops a principled subjective logic-driven crowd-AI fusion framework to effectively integrate the class labels generated by the ADA model instances from GDO module and the crowd labels returned by crowd intelligence query to derive accurate ADA results for each studied disaster event.

4.1 Generality-aware Deep Optimization

We first present the generality-aware deep network optimization design to learn a set of ADA model instances that have a high likelihood of achieving an optimized trade-off between the model generality and specificity in ADA applications. We first introduce a key definition for our GDO module.

Definition 7. Deep Estimation Network (Φ): We define Φ to be the deep estimation network (i.e., AI model) in the GDO module that estimates the class labels from the input image samples. Rather than reinventing the wheel, we set Φ to be a representative convolutional neural network (e.g., ResNet, VGG, DenseNet) that is designed to perform the image-based multi-class classification tasks.

Given the deep estimation network Φ, our next step is to learn the optimal network instance of Φ for accurate event-wise ADA performance. To that end, our GDO module introduces two sets of loss functions to explicitly supervise the network optimization process and derive the optimal network instance that can achieve a good trade-off between ADA model generality and specificity. We first define the accuracy-aware loss function for Φ as:

$$\mathcal{L}_1 = \sum_{\forall D_t \in \boldsymbol{D}} \sum_{k=1}^{K} || \Pr(\widehat{Y_{D_t}^{\Phi}} \neq k | Y_{D_t} = k) ||_2 \tag{2}$$

where $\mathcal{L}_1$ denotes the accuracy-aware loss function for Φ. D_t is a disaster event from the set of studied events $\boldsymbol{D}$. K denotes the number of unique classes in the ADA application of interest. $\widehat{Y_{D_t}^{\Phi}}$ and Y_{D_t} indicate the *estimated* class labels from Φ and *ground-truth* class labels for all imagery data from disaster event D_t, respectively. $|| \cdot ||_2$ is the L2-norm of a matrix. The objective of the accuracy-aware loss is to supervise Φ to accurately estimate the class labels from all input imagery data. However, a limitation of $\mathcal{L}_1$ loss function is that $\mathcal{L}_1$ only focuses on the overall ADA performance but may not supervise Φ to achieve optimized ADA performance on each individual disaster event. Therefore, we further define the generality-aware loss function for Φ to address such a limitation as:

$$\mathcal{L}_2 = \sum_{\forall D_{t_1}, D_{t_2} \in \boldsymbol{D}, D_{t_1} \neq D_{t_2}} \sum_{k=1}^{K} || \Pr(\widehat{Y_{D_{t_1}}^{\Phi}} = k | D_{t_1}, Y_{D_{t_1}} = k) - \Pr(\widehat{Y_{D_{t_2}}^{\Phi}} = k | D_{t_2}, Y_{D_{t_2}} = k) ||_2 \tag{3}$$

where $\mathcal{L}_2$ is the generality-aware loss function for Φ. D_{t_1}, D_{t_2} represent any two different disaster events from the set of studied disaster events $\boldsymbol{D}$. $\widehat{Y_{D_{t_1}}^{\Phi}}$ and $\widehat{Y_{D_{t_2}}^{\Phi}}$ indicate the *estimated* class labels for all imagery data from event D_{t_1} and D_{t_2}, respectively. $Y_{D_{t_1}}$ and $Y_{D_{t_2}}$ indicate the *ground-truth* class labels for all imagery data from event D_{t_1} and D_{t_2}, respectively. We then combine the two loss functions to derive the overall loss function for Φ to learn the optimal network instance of Φ as:

$$\mathcal{L}_{\mathit{Overall}} = \mathcal{L}_1 + \mathcal{L}_2 \tag{4}$$

Using the overall loss function above, the optimal network instances of Φ can be learned by investigating the trade-off between the exploitation and exploration during the network optimization process through a budget-constrained multi-armed bandit learning process [Feurer and Hutter, 2019]. On the one hand, we keep tuning the same network instance that achieves the low value for $\mathcal{L}_{\mathit{Overall}}$. On the other hand, we take action to attempt new network instances to prevent the model from being trapped into a local optimum. Such a optimization strategy could jointly explore the large network instance space while finding the optimal network instance for Φ.

After performing the budget-constrained multi-armed bandit learning process, one possible solution to obtain the optimal network instance is to use the network instance with the lowest value of $\mathcal{L}_{\mathit{Overall}}$ as the optimal network instance. However, the optimized network instance could be overfitted to the training/validation data and lead to non-negligible performance degradation when it is applied to the testing data due to the potential feature discrepancy between the training/validation and testing sets [Saunders, 2022]. To address such an issue, our GDO module not only exploits the network instances with the lowest value of $\mathcal{L}_{\mathit{Overall}}$ but also explores other candidate network instances with low values of $\mathcal{L}_{\mathit{Overall}}$. We formally define the network instances generated by our GDO module as follows.

Definition 8. Optimized Network Instance Set ($\boldsymbol{M}$): We define $\boldsymbol{M} = \{M_1, M_2, ..., M_J\}$ as a set of network instances learned by the GDO module, which includes network instances with top J lowest values in $\mathcal{L}_{\mathit{Overall}}$. In addition, M_j indicate the j^{th} learned network instance.

Note that all network instances in $\boldsymbol{M}$ are the instances of the deep estimation network Φ (Definition 7). To generate different network instances in $\boldsymbol{M}$, our GDO module keeps tracking the $\mathcal{L}_{\mathit{Overall}}$ of different network instances generated during *one* budget-constrained multi-armed bandit learning process. Our GDO module then adds the network instances with top J lowest values in $\mathcal{L}_{Overall}$ to $\boldsymbol{M}$. The above design avoids the low computational efficiency of performing the budget-constrained multi-armed bandit learning process J times to generate different network instances in $\boldsymbol{M}$.

Our CollabGeneral then jointly leverages the identified network instances and crowd intelligence to derive accurate ADA results for all studied disaster events, which will be discussed in the next subsection.

4.2 Subjective Logic-driven Crowd-AI Fusion

In this module, we design a novel subjective logic-driven crowd-AI fusion framework to fuse the AI and crowd intelligence to derive the accurate ADA results for all studied disaster events to address the ADA model generality problem.

We first discuss how to perform crowd intelligence query Q to collect crowd intelligence for the SCF module. We observe that it is impractical to query the crowd intelligence for all studied image samples due to the budget and resource constraints, which is especially challenging in ADA applications with massive social media data inputs [Li *et al.*, 2019]. Therefore, our SCF module samples a subset of image samples for Q in which different network instances in $\boldsymbol{M}$ (Definition 8) cannot reach a consensus on. We first measure the *divergence* of the class labels estimated by all network instances in $\boldsymbol{M}$ for each image sample X_i using Shannon entropy [Lin, 1991]. The divergence indicates the degree of disagreement between different network instances in $\boldsymbol{M}$ on the estimated class label for X_i. We then select the image samples with top $\delta \times I$ highest divergence for Q. Here, δ indicates the percentage of studied disaster-related imagery data that are sampled for Q. δ is determined by the trade-off between the ADA model performance and the crowdsourcing cost in the ADA application of interest. I is the total number of studied images.

Our next step is to effectively fuse the crowd labels returned by Q with the estimated labels generated by different network instances in $\boldsymbol{M}$. In particular, we define:

Definition 9. Crowd-AI Fusion Committee ($\boldsymbol{S}$): We define $\boldsymbol{S} = \{S_1, S_2, ..., S_C\}$ as a crowd-AI fusion committee, which contains all J different optimized network instances $\boldsymbol{M}$ learned by the GDO module and all B different crowd workers $\boldsymbol{W}$ in an ADA application. In particular, we have $\boldsymbol{S} = \boldsymbol{M} \cup \boldsymbol{W}$, where $C = J + B$. C is the size of committee $\boldsymbol{S}$, and S_c is a committee member in $\boldsymbol{S}$ (i.e., either an AI network instance or a crowd worker.

The goal of our SCF module is to effectively fuse the inputs from all members in $\boldsymbol{S}$ to derive the accurate ADA labels for the studied disaster events. To that end, we first define the "opinion" of each committee member towards the class label of each image sample through subjective logic, a probabilistic logic that models the epistemic uncertainty and source trust when combining the opinions from different sources [Jøsang, 2016]. In our paper, we leverage the subjective logic to explicitly model each committee member's uncertainty and reliability in estimating the ADA labels for all disaster events.

Definition 10. Committee Member Opinion Entity ($\boldsymbol{E}$): We define $E^k_{S_c} = \{T^k_{S_c}, F^k_{S_c}, U^k_{S_c}\}$ to represent the opinion of a member S_c on whether an image sample belongs to a particular class k or not. In particular, we have:

$$T^k_{S_c}, F^k_{S_c}, U^k_{S_c} \in [0,1], T^k_{S_c} + F^k_{S_c} + U^k_{S_c} = 1 \tag{5}$$

where $T^k_{S_c}$ and $F^k_{S_c}$ indicates S_c's belief and disbelief in the class label of an image sample to be k, respectively. $U^k_{S_c}$ indicates S_c's uncertainty in determining if the class label of an image sample to be k or not.

Given the opinion entity of each committee member, we can utilize the consensus operation from subjective logic to combine the opinions from different committee members. Consensus operation is a key operation in subjective logic that is used to determine the shared belief and uncertainty of two sources by considering the individual belief and uncertainty of each source. In particular, we can use the consensus operation $\oplus$ to combine the opinions from any two committee member S_p and S_q as follows:

$$E^K_{S_p,S_q} = \{T^k_{S_p,S_q}, F^k_{S_p,S_q}, U^k_{S_p,S_q}\} = E^k_{S_p} \oplus E^k_{S_q} \tag{6}$$

where $E^K_{S_p,S_q}$ indicates the opinion entity after combining the opinions from both S_p and S_q, which indicates their collective opinions on whether an image sample belongs to a particular class k or not.

Then, we can recursively adopt the consensus operation $\oplus$ to combine the opinions from all committee members in the crowd-AI fusion committee as follows:

$$E^k_{\boldsymbol{S}} = \{T^k_{\boldsymbol{S}}, F^k_{\boldsymbol{S}}, U^k_{\boldsymbol{S}}\} = E^k_{S_1} \oplus E^k_{S_2} \oplus, ..., \oplus E^k_{S_C} \tag{7}$$

Given the combined opinion $E^k_{\boldsymbol{S}}$ from all committee member in the crowd-AI fusion committee $\boldsymbol{S}$, we can leverage it to derive the accurate class label for each image sample. In particular, we set the class label estimated by our CollabGeneral framework to be the one that has the highest belief value $T^k_{\boldsymbol{S}^{i,k}}$ among all possible class labels k for each studied image sample X_i as follows:

$$\arg\max_{k^*} T^k_{\boldsymbol{S}^{i,k}}, \text{ where } k \in \{1, 2, ..., K\}, \text{ set } k^* \text{ as } \widehat{Y}_i \tag{8}$$

where $\boldsymbol{S}^{i,k}$ indicates the set of committee members who estimates the class label for X_i as k.

However, $E^k_{S_c}$ for each committee member S_c in $\boldsymbol{S}$ is unknown *a priori* and we need to infer the value for each $E^k_{S_c}$ before estimating the accurate class label for each image sample. To that end, we further design an iterative learning framework in our SCF module to obtain the accurate value for each $E^k_{S_c}$. In particular, we first introduce two important concepts in our iterative learning framework.

Definition 11. Committee Member Reliability ($\boldsymbol{R}$): We define R^k_c to be the probability of a committee member S_c in correctly estimating the class label of an image from class k.

Definition 12. Image Sample Discriminative Score ($\boldsymbol{Z}$): We define Z^k_i as the discriminative score of an image sample X_i in terms of identifying the reliable committee member that can correctly estimate the label for image samples of class k.

Given the above two definitions, we note that the values of both committee member reliability $\boldsymbol{R}$ and the image sample discriminative score $\boldsymbol{Z}$ are unknown and depend on each other. Therefore, we optimize $\boldsymbol{R}$ and $\boldsymbol{Z}$ alternately as follows.

First, we optimize the image sample discriminative score $\boldsymbol{Z}$ given the committee member reliability $\boldsymbol{R}$ as follows:

$$Z_i^k = \frac{\sum_{S_p, S_q \in \boldsymbol{S^{i,k}}} R_p^k \times R_q^k \times \frac{N_{S_p,S_q}^k}{N_{S_q}^k}}{\sum_{S_p, S_q \in \boldsymbol{S^{i,k}}} R_p^k \times R_q^k} \tag{9}$$

where $\boldsymbol{S^{i,k}}$ is the set of crowd-AI committee members who estimate the class label of X_i to be k. S_p and S_q are any two committee members in $\boldsymbol{S^{i,k}}$. R_p^k and R_q^k are the reliability of S_p and S_q, respectively. N_{S_p,S_q}^k is the number of image samples where both S_p and S_q estimate the class label to be k. $N_{S_q}^k$ is the number of image samples where S_q estimates the class label to be k. In addition, Z_i^k is set to be 0 if there is only 1 or no committee member label X_i to be k. Intuitively, a high value of Z_i^k indicates a high likelihood that the estimated class label for X_i is to be k, and vice versa.

Then, we compute the committee member reliability $\boldsymbol{R}$ using the updated image sample discriminative score $\boldsymbol{Z}$ as:

$$R_c^k = \frac{\sum_{\forall i \in \boldsymbol{\Delta_k^{S_c}}} \left(Z_i^k \times \sum_{S_p \in \boldsymbol{S^{i,k}}} \frac{N_{S_p,S_c}^k}{N_{S_c}^k}\right)}{\sum_{\forall i \in \boldsymbol{\Delta_k^{S_c}}} Z_i^k} \tag{10}$$

where $\boldsymbol{\Delta_k^{S_c}}$ indicates the set of all image samples where S_c estimates the class label to be k. Intuitively, a high value of R_c^k indicates that the class labels estimated by S_c are more likely to be correct.

Given the above two definitions, we can obtain the optimal value for all ${Z_i^k}^*$ and ${R_c^k}^*$ by iteratively updating all Z_i^k and R_c^k until their values convergence (e.g., the values of Z_i^k and R_c^k remains unchanged between two consecutive iterations). We then leverage ${Z_i^k}^*$ and ${R_c^k}^*$ to derive the optimal opinion entity ${E_{S_c}^k}^*$ for each committee member as follows:

$${U_{S_c}^k}^* = 1 - \Omega(\mathcal{Z}), {T_{S_c}^k}^* = \Omega(\mathcal{Z}) \times {R_c^k}^*, {F_{S_c}^k}^* = 1 - {T_{S_c}^k}^* - {U_{S_c}^k}^* \tag{11}$$

where $\Omega(\cdot)$ is a normalization function to normalize the input between 0 and 1. $\mathcal{Z} = \sum_{\forall i \in \boldsymbol{\Delta_k^{S_c}}} {Z_i^k}^*$ indicates the likelihood that S_c is certain about estimated labels for images of class k.

The learned opinion entity ${E_{S_c}^k}^*$ is then plugged in Equation (7) to derive the accurate class labels for all imagery data in each studied disaster event.

5 Evaluation

5.1 Datasets and Crowdsourcing Settings

Disaster Damage Assessment Datasets: In the experiments, we use four publicly available real-world ADA datasets.[3] The datasets consist of social media images collected from four different disaster events: Hurricane Irma (2017), Ecuador Earthquake (2016), Nepal Earthquake (2015), and Sri Lanka Flooding (2017). Images in each dataset reflect disaster-specific visual characteristics of a disaster (e.g., structure damage vs. flooding damage, urban layout vs. rural layout, plateau landscape vs. coastal landscape). Following the standard practice in ADA applications [Nguyen *et al.*, 2017], we classify the disaster damages into three classes including *severe damage*, *medium damage*, and *no/minor damage*. Each image is annotated by three independent annotators, with the majority voting as the aggregated label. We invited our domain expert to cross-validate the aggregated label to obtain the final ground-truth annotation. A summary of all datasets is presented in Table 1. Additionally, we split the training and testing sets with a ratio of 7:3 and use the training sets to train all compared schemes for ADA tasks and evaluate their performance on the testing sets.

Crowdsourcing Settings: We leverage the widely-used Amazon Mechanical Turk (AMT) to acquire crowd intelligence in our experiments. AMT is one of the largest crowdsourcing platforms offering 24/7 crowdsourcing services from a massive amount of crowd workers around the world. For each task on AMT, we recruit crowd workers who have finished at least 1000 approved tasks with an overall task approval rate of 95% or above to ensure the crowdsourcing label quality. We pay $0.05 per image to the crowd workers and follow the IRB protocol approved for this project.

Event	Images	No/Minor	Medium	Severe
Hurricane Irma	893	34.6%	39.6%	25.8%
Ecuador Earthquake	670	41.0%	5.7%	53.3%
Nepal Earthquake	666	41.9%	13.5%	44.6%
Sri Lanka Flooding	144	40.4%	40.3%	19.3%

Table 1: Statistics of Four ADA Datasets

5.2 Baseline Methods and Experiment Settings

In our evaluation, we compare CollabGeneral with a set of state-of-the-art baselines, including (1) Deep Neural Network (DNN): **ResNet** [Targ *et al.*, 2016], **DenseNet** [Huang *et al.*, 2017], and **VGG** [Li *et al.*, 2018]; (2) AI Model Generalization: **GTA** [Sankaranarayanan *et al.*, 2018], **VS** [Kini *et al.*, 2021], **SL** [Wang *et al.*, 2019b]; (3) Crowd-AI Collaboration: **Deep Active** [Sener and Savarese, 2018], **CrowdLearn** [Zhang *et al.*, 2019], **SL** [Wang *et al.*, 2019b].

In the experiments, to ensure the fairness of comparison, we use the same inputs for all compared methods. In particular, the inputs to each scheme include: 1) the social media images for all studied disaster events in both training and testing datasets; 2) the ground-truth labels for social media images in the training dataset, where the number of training images from each disaster event is proportional to the total number of images from that event as shown in Table 1; and 3) the labeled social media images returned by the crowd workers. In particular, we use the crowd labels to retrain the DNN and AI model generalization baselines to ensure all baselines have the same inputs and the performance of compared baselines is optimized. For the DNN baselines, we consider two different training settings: 1) training a single DNN model for *all* studied disaster events, which is referred as *DNN-A* (e.g., ResNet-A for ResNet); 2) training four DNN models, one for each *specific* disaster event, which is referred as *DNN-S* (e.g., ResNet-S for ResNet).

[3] https://crisisnlp.qcri.org/

		Ecuador Earthquake			Sri Lanka Flooding			Overall		
Category	Algorithm	F1-Score	MCC	$\mathcal{K}$-Score	F1-Score	MCC	$\mathcal{K}$-Score	F1-Score	MCC	$\mathcal{K}$-Score
DNN-A	ResNet-A	0.8032	0.6658	0.6513	0.5214	0.4582	0.3732	0.7326	0.6138	0.5851
	DenseNet-A	0.7999	0.6529	0.6444	0.5519	0.4607	0.3942	0.7384	0.6098	0.5904
	VGG-A	0.8023	0.6469	0.6319	0.7105	0.5647	0.5323	0.7785	0.6505	0.6330
DNN-S	ResNet-S	0.8315	0.6849	0.6833	0.6434	0.5507	0.4850	0.7779	0.6518	0.6426
	DenseNet-S	0.7975	0.6527	0.6399	0.6913	0.5796	0.5372	0.7758	0.6539	0.6363
	VGG-S	0.8132	0.6569	0.6493	0.4837	0.4232	0.3264	0.7270	0.5875	0.5639
AI Model Generalization	GTA	0.7117	0.4523	0.4516	0.5545	0.4106	0.3733	0.6666	0.4498	0.4488
	VS	0.7724	0.5457	0.5334	0.7502	0.6212	0.5950	0.7561	0.5940	0.5820
	SL	0.8309	0.7058	0.7034	0.6406	0.4897	0.4622	0.7870	0.6668	0.6607
Crowd-AI	Deep Active	0.7986	0.6524	0.6452	0.4347	0.4184	0.3329	0.7112	0.5912	0.5695
	CrowdLearn	0.8145	0.6574	0.6552	0.5263	0.4796	0.3955	0.7425	0.6074	0.5938
	LL4AL	0.7886	0.6133	0.6123	0.4177	0.3830	0.3110	0.7018	0.5565	0.5459
Ours	**CollabGeneral**	**0.8574**	**0.7267**	**0.7266**	**0.8024**	**0.6864**	**0.6791**	**0.8436**	**0.7388**	**0.7384**

Table 2: Evaluation Results (Different Types of Events)

We use three evaluation metrics that are commonly used to quantify the performance of multi-class text classification: 1) *F1-score*, and 2) *Matthews Correlation Coefficient (MCC)*, 3) *kappa score ($\mathcal{K}$-Score)*. We use MCC and $\mathcal{K}$-Score in our evaluation because our datasets are imbalanced, and these two metrics are known to be reliable on imbalanced data [Chicco and Jurman, 2020]. The higher values of the above metrics demonstrate better ADA performance.

5.3 Evaluation Results

Model Generality on Different Types of Disaster Events

Firstly, we study the ADA model generality with a challenging evaluation setting, where the studied disaster events are of *completely different types*: Ecuador Earthquake and Sri Lanka Flooding. We summarize the evaluation results in Table 2. We observe that CollabGeneral consistently outperforms all compared baselines in terms of the ADA performance on each individual event and the overall performance across two different types of events. For example, the performance gains of CollabGeneral compared to the best-performing baseline (i.e., VS) on the Sri Lanka Flooding event on F1-Score, MCC, and $\mathcal{K}$-Score are 5.22%, 6.52%, and 8.41%, respectively.

Model Generality on Different Number of Events

Secondly, we evaluate the ADA performance of CollabGeneral when there exist more than two disaster events. In particular, we evaluate CollabGeneral up to four different disaster events by leveraging all possible disaster events available in our datasets. In our experiments, we evaluate the ADA performance by comparing CollabGeneral with the best-performing baselines in each category. The results are presented in Figure 2. Note that we only show the evaluation results on the F1-Score due to the page limit. The evaluation results on other metrics are similar. We observe that our CollabGeneral continuously outperforms all compared baselines on both individual events and overall performance when the number of studied disaster events increases. This is because our subjective logic-based crowd-AI framework design effectively improves the ADA model generality without sacrificing the model's specificity on each studied disaster event.

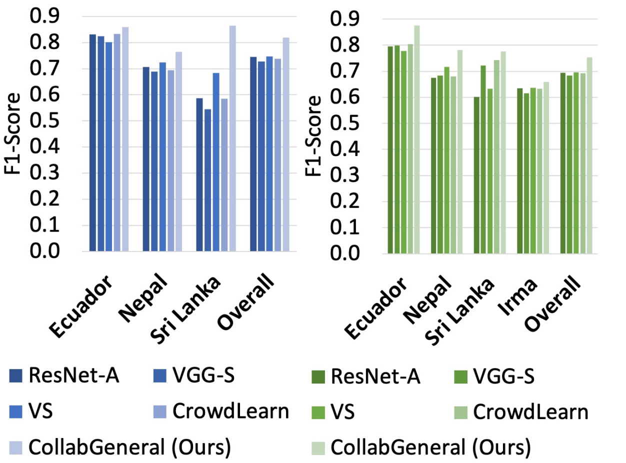

Figure 2: Performance Comparisons on Different Number of Events

6 Conclusion

The paper presents a CollabGeneral framework to address the ADA model generality problem. In particular, we design a generality-aware crowd-AI collaborative framework that integrates the complementary AI and crowd intelligence to achieve an optimized trade-off between generality and specificity of ADA models. Our CollabGeneral is shown to significantly improve the ADA model generality by achieving the highest ADA accuracy in each studied disaster event compared to a rich set of deep neural networks, AI model generality, and crowd-AI baselines in four different real-world ADA datasets. We believe CollabGeneral provides useful insights to address the AI model generality problem in many real-world AI-driven applications (e.g., intelligent transportation, smart health, AIoT) for future research in this domain.

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105005, IIS-2008228, CNS-1845639, CNS-1831669. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[Barmpoutis *et al.*, 2019] Panagiotis Barmpoutis, Kosmas Dimitropoulos, Kyriaki Kaza, and Nikos Grammalidis. Fire detection from images using faster r-cnn and multidimensional texture analysis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8301–8305. IEEE, 2019.

[Chan *et al.*, 2018] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21, 2018.

[Chicco and Jurman, 2020] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

[Draws *et al.*, 2021] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 48–59, 2021.

[Feurer and Hutter, 2019] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

[Fuchs, 2022] Thomas Fuchs. Human and artificial intelligence: A critical comparison. In *Intelligence-Theories and Applications*, pages 249–259. Springer, 2022.

[Gemalmaz and Yin, 2021] Meric Altug Gemalmaz and Ming Yin. Accounting for confirmation bias in crowdsourced label aggregation. In *IJCAI*, pages 1729–1735, 2021.

[Ghifary *et al.*, 2016] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[Hube *et al.*, 2019] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[Imran *et al.*, 2022] Muhammad Imran, Umair Qazi, Ferda Ofli, Steve Peterson, and Firoj Alam. Ai for disaster rapid damage assessment from microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12517–12523, 2022.

[Jøsang, 2016] Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016.

[Kankanamge *et al.*, 2020] Nayomi Kankanamge, Tan Yigitcanlar, Ashantha Goonetilleke, and Md Kamruzzaman. Determining disaster severity through social media analysis: Testing the methodology with south east queensland flood tweets. *International journal of disaster risk reduction*, 42:101360, 2020.

[Kini *et al.*, 2021] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34:18970–18983, 2021.

[Kumar *et al.*, 2020] Pakhee Kumar, Ferda Ofli, Muhammad Imran, and Carlos Castillo. Detection of disaster-affected cultural heritage sites from social media images using deep learning techniques. *Journal on Computing and Cultural Heritage (JOCCH)*, 13(3):1–31, 2020.

[Li *et al.*, 2018] Xukun Li, Doina Caragea, Huaiyu Zhang, and Muhammad Imran. Localizing and quantifying damage in social media images. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 194–201. IEEE, 2018.

[Li *et al.*, 2019] Xukun Li, Doina Caragea, Cornelia Caragea, Muhammad Imran, and Ferda Ofli. Identifying disaster damage images using a domain adaptation approach. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response And Management*, 2019.

[Li *et al.*, 2020] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020.

[Lin, 1991] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[Mangalathu and Burton, 2019] Sujith Mangalathu and Henry V Burton. Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions. *International Journal of Disaster Risk Reduction*, 36:101111, 2019.

[McEntire, 2021] David A McEntire. *Disaster response and recovery: strategies and tactics for resilience*. John Wiley & Sons, 2021.

[Metaxa-Kakavouli *et al.*, 2018] Danaë Metaxa-Kakavouli, Paige Maas, and Daniel P Aldrich. How social ties influence hurricane evacuation behavior. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–16, 2018.

[Mouzannar **et al.**, 2018] Hussein Mouzannar, Yara Rizk, and Mariette Awad. Damage identification in social media posts using multimodal deep learning. In *ISCRAM*, 2018.

[Nguyen **et al.**, 2017] Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 569–576, 2017.

[Ning **et al.**, 2020] Huan Ning, Zhenlong Li, Michael E Hodgson, and Cuizhen Wang. Prototyping a social media flooding photo screening system based on deep learning. *ISPRS International Journal of Geo-Information*, 9(2):104, 2020.

[Ren **et al.**, 2021] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.

[Risi and Togelius, 2020] Sebastian Risi and Julian Togelius. Increasing generality in machine learning through procedural content generation. *Nature Machine Intelligence*, 2(8):428–436, 2020.

[Rolnick **et al.**, 2017] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[Sankaranarayanan **et al.**, 2018] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

[Saunders, 2022] Danielle Saunders. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:351–424, 2022.

[Sener and Savarese, 2018] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[Soden and Owen, 2021] Robert Soden and Embry Owen. Dilemmas in mutual aid: Lessons for crisis informatics from an emergent community response to the pandemic. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–19, 2021.

[Soden and Palen, 2018] Robert Soden and Leysia Palen. Informating crisis: Expanding critical perspectives in crisis informatics. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

[Sun **et al.**, 2020] Wenjuan Sun, Paolo Bocchini, and Brian D Davison. Applications of artificial intelligence for disaster management. *Natural Hazards*, 103(3):2631–2689, 2020.

[Targ **et al.**, 2016] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*, 2016.

[Wang **et al.**, 2019a] Dong Wang, Boleslaw K Szymanski, Tarek Abdelzaher, Heng Ji, and Lance Kaplan. The age of social sensing. *Computer*, 52(1):36–45, 2019.

[Wang **et al.**, 2019b] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.

[Zade **et al.**, 2018] Himanshu Zade, Kushal Shah, Vaibhavi Rangarajan, Priyanka Kshirsagar, Muhammad Imran, and Kate Starbird. From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–18, 2018.

[Zhang **et al.**, 2019] Daniel Zhang, Yang Zhang, Qi Li, Thomas Plummer, and Dong Wang. Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1221–1232. IEEE, 2019.

[Zhang **et al.**, 2020] Yang Zhang, Ruohan Zong, and Dong Wang. A hybrid transfer learning approach to migratable disaster assessment in social media sensing. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 131–138. IEEE, 2020.

[Zhang **et al.**, 2021] Yang Zhang, Ruohan Zong, Lanyu Shang, Ziyi Kou, and Dong Wang. A deep contrastive learning approach to extremely-sparse disaster damage assessment in social sensing. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 151–158, 2021.

[Zhang **et al.**, 2022a] Yang Zhang, Ruohan Zong, Ziyi Kou, Lanyu Shang, and Dong Wang. Crowdnas: A crowd-guided neural architecture searching approach to disaster damage assessment. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29, 2022.

[Zhang **et al.**, 2022b] Yang Zhang, Ruohan Zong, Lanyu Shang, Ziyi Kou, Huimin Zeng, and Dong Wang. Crowdoptim: A crowd-driven neural network hyperparameter optimization approach to ai-based smart urban sensing. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27, 2022.

[Zhang **et al.**, 2023] Yang Zhang, Lanyu Shang, Ruohan Zong, Huimin Zeng, Zhenrui Yue, and Dong Wang. Collabequality: A crowd-ai collaborative learning framework to address class-wise inequality in web-based disaster response. In *Proceedings of the ACM Web Conference 2023*, pages 4050–4059, 2023.

[Zong **et al.**, 2023] Ruohan Zong, Yang Zhang, Lanyu Shang, and Dong Wang. Contrastfaux: Sparse semi-supervised fauxtography detection on the web using multi-view contrastive learning. In *Proceedings of the ACM Web Conference 2023*, pages 3994–4003, 2023.