CollabLearn: An Uncertainty-Aware Crowd-AI Collaboration System for Cultural Heritage Damage Assessment

Yang Zhang, *Graduate Student Member, IEEE*, Ruohan Zong, *Student Member, IEEE*, Ziyi Kou, *Student Member, IEEE*, Lanyu Shang, *Student Member, IEEE*, and Dong Wang, *Member, IEEE*

***Abstract*— Cultural heritage sites are precious and fragile resources that hold significant historical, esthetic, and social values in our society. However, the increasing frequency and severity of natural and man-made disasters constantly strike the cultural heritage sites with significant damages. In this article, we focus on a cultural heritage damage assessment (CHDA) problem where the goal is to accurately locate the damaged area of a cultural heritage site using the imagery data posted on social media during a disaster event by exploring the collective strengths of both AI and human intelligence from crowdsourcing systems. Unlike other infrastructure-based solutions, social media platforms provide a more pervasive and scalable solution to acquire timely cultural heritage damage information during disaster events. Our work is motivated by the limitation of current AI solutions that fail to accurately model the complex cultural heritage damage due to the lack of essential human cultural knowledge to differentiate various damage types and identify the actual causes of the damage. Two critical technical challenges exist in solving our problem: 1) it is challenging to effectively detect the problematic cultural heritage damage estimation of AI in the absence of ground truth labels and 2) it is nontrivial to acquire accurate cultural background knowledge from the potentially unreliable crowd workers to effectively address the failure cases of AI. To address the above-mentioned challenges, we develop *CollabLearn*, an uncertainty-aware crowd-AI collaborative assessment system that explicitly explores the human intelligence from crowdsourcing systems to identify and fix AI failure cases and boost the damage assessment accuracy in CHDA applications. The evaluation results on real-world datasets show that CollabLearn consistently outperforms both the state-of-the-art AI-only and crowd-AI hybrid baselines in accurately assessing the damage of several world-renowned cultural heritage sites in recent disaster events.**

***Index Terms*— Crowd-AI collaboration, cultural heritage damage assessment (CHDA), online social media, uncertainty quantification.**

Manuscript received 26 April 2021; revised 27 July 2021; accepted 25 August 2021. Date of publication 9 September 2021; date of current version 30 September 2022. This work was supported in part by the National Science Foundation under Grant IIS-2008228, Grant CNS-1845639, and Grant CNS-1831669; and in part by the Army Research Office under Grant W911NF-17-1-0409. 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 Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. *(Corresponding author: Dong Wang.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Notre Dame Institutional Review Board (IRB) Under Protocol No. 19-11-5657.

Yang Zhang and Ruohan Zong are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: yzhang42@nd.edu; rzong@nd.edu).

Ziyi Kou, Lanyu Shang, and Dong Wang are with the School of Information Sciences, University of Illinois at Urbana–Champaign, Champaign, IL 61820 USA (e-mail: ziyikou2@illinois.edu; lshang3@illinois.edu; dwang24@illinois.edu).

Digital Object Identifier 10.1109/TCSS.2021.3109143

I. Introduction

CULTURAL heritage sites (e.g., historical buildings, monuments, archeological sites, and landscapes) are precious and fragile resources that hold significant historical, esthetic, and social values in our society [1]. However, the increasing frequency and severity of natural and man-made disasters (e.g., earthquakes, hurricanes, and vandalism) constantly strike the cultural heritage sites with significant damages [2]. For example, thousands of heritage places in Syria have suffered significant damage from conflict, looting, and the cessation of official protection since 2011. This article focuses on an emerging application, *cultural heritage damage assessment* (CHDA), that aims to protect and conserve cultural heritage sites. The objective of CHDA applications is to accurately locate the damaged areas of a cultural heritage site by exploring the imagery data posted on social media during a disaster event. Unlike other infrastructure-based solutions (e.g., using surveillance cameras, drones, and satellites), the social media platforms provide an *infrastructure-free* solution that is more pervasive and scalable to acquire timely damage information of the cultural heritage sites during disaster events [3]–[5]. The assessment information can then be leveraged by the government agencies and organizations to provide conservation and recovery actions to the sites and save them from further damages.

Recent progress in AI and image processing has been made toward addressing the disaster damage assessment (DDA) problem [3], [6]–[11]. In particular, the deep learning-based DDA solutions significantly reduce the labeling costs while providing a reasonable assessment accuracy compared with the traditional domain experts-based solutions [7]. Compared with the DDA problem that primarily focuses on identifying disaster-related damages from social media images, the CHDA problem is more challenging due to the high complexity of cultural heritage damage and the lack of cultural background knowledge of AI-based DDA solutions [12]. Fig. 1 shows

a few examples of failure scenarios when current AI-based solutions are applied to assess the damaged areas of the cultural heritage sites. For example, the damage areas detected by AI algorithms for the tower in Fig. 1(A), the stone lions in Fig. 1(B), and the castle in Fig. 1(C) is actually part of the cultural and artistic design that is often observed at the cultural heritage sites [13]. Meanwhile, the damage areas detected by AI algorithms for the stair flight in Fig. 1(D), the stone wall in Fig. 1(E), and the tiles in Fig. 1(F) are caused by long-term aging effects, which are often confused with the damages caused by recent disasters for AI algorithms. In contrast, humans are often observed to perform better at identifying the damages of cultural heritage sites where AI solutions fail. The reason is intuitive: humans normally have certain cultural background knowledge and a reasonable understanding of the complex scenes in cultural heritage sites, which together help them make a better judgment in CHDA applications. However, the solutions that fully depend on human efforts are expensive in terms of both time and cost and not scalable to address our problem with a large amount of social media data inputs during disaster events [6].

In this article, we develop an integrated crowd-AI collaboration system to solve the CHDA problem by exploring the collective strength of both AI and human intelligence. In particular, our goal is to achieve a win–win objective between AI and human intelligence by effectively leveraging the high detection efficiency of AI solutions to automatically process the vast amount of cultural heritage site images and explicitly exploring the human intelligence to identify and fix the failure cases of AI in CHDA applications. To obtain timely and scalable human intelligence, we leverage the widely adopted open crowdsourcing platforms [e.g., Amazon Mechanical Turk (AMT)], which offer a large amount of 24/7 available crowd workers with reasonable costs [14]. We refer to the human intelligence acquired from the crowdsourcing platform as *crowd intelligence (CI)*. The design of such a crowd-AI collaboration system is a nontrivial task due to two critical technical challenges that are elaborated in the following.

A. Identification of AI Failure Cases

The first challenge lies in how to accurately identify the failure cases of AI damage assessment solutions without knowing the ground truth labels of images *a priori*. One straightforward solution to address this problem is to directly ask the crowd workers to examine every output of AI solutions to identify and fix the failure cases, as shown in Fig. 1. However, such an approach is impractical due to the heavy labor costs and low efficiency, especially in the context of the massive social media data inputs. Some initial efforts were made to address this issue by only selecting the imagery data with complicated image properties (e.g., images with complex contents and color distribution) for crowd labeling under the assumption that the AI solutions are more likely to fail when the image is complex [15]. However, such an assumption does not always hold for the cultural heritage damages as AI may also fail when the input image is relatively simple [e.g., the color distributions in Fig. 1(D) are quite simple]. Recent work

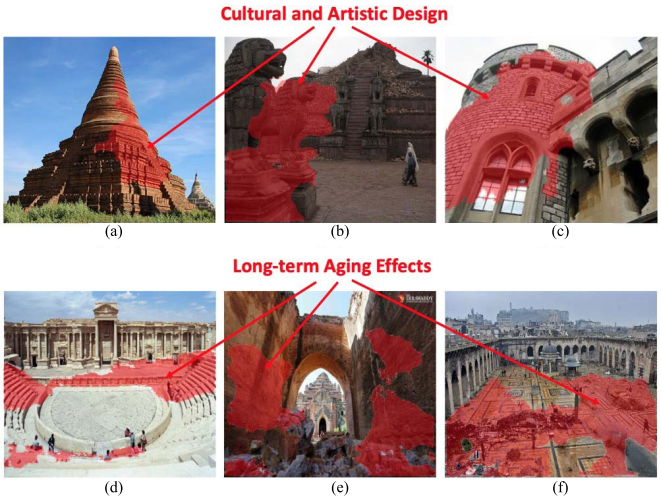


Fig. 1. Examples of failure cases of AI solutions for CHDA. (a) Tower. (b) Stone lions. (c) Castle. (d) Stair flight. (e) Stone wall. (f) Tiles.

on uncertainty-aware AI solutions (e.g., query-by-committee, dropout) could also potentially be applied to detect the failure cases of AI [11], [16]. Those approaches often leverage a committee of different AI models or different instances of the same model to identify the problematic cases based on the consensus from the outputs of the committee members. However, those approaches will fail when all members in the committee happen to make similar mistakes on the same input [17]. Therefore, it remains to be a challenging question on how to effectively detect the failure cases of AI in the absence of ground truth labels in CHDA applications.

B. Imperfect Crowd Intelligence

The second challenge lies in how to acquire accurate CI from the potentially unreliable crowd workers to fix the failure cases of AI. Unlike the labels annotated by domain experts, the labels from the crowd workers can be uncertain and inconsistent [18]. Such inconsistency is especially salient in CHDA applications due to the intricate nature of the cultural heritage site damage. For example, in Fig. 2, we observe that the damage areas identified by different crowd workers are not always consistent. In particular, workers 1 and 2 in Fig. 2(B) and (C) believe the stone pillar is damaged, while worker 3 in Fig. 2(D) thinks the stone pillar is intact during a disaster event. Such uncertain and inconsistent crowd labels present a critical challenge to the current active learning-based AI systems that rely on accurate human labels to troubleshoot and retrain the AI models to optimize the model performance [19], [20]. In particular, the imperfect CI could potentially collapse the AI model during the model retraining process [21]. Several recent efforts have been made on training the AI models with imperfect labels [22], [23]. However, those models are designed for structural image processing tasks (e.g., segmenting structural magnetic resonance imaging data) with limited errors in the training labels annotated by domain experts, which cannot be directly applied to handle the complex social media images with the uncertain and inconsistent crowd labels. Therefore, it remains to be a nontrivial question on how to

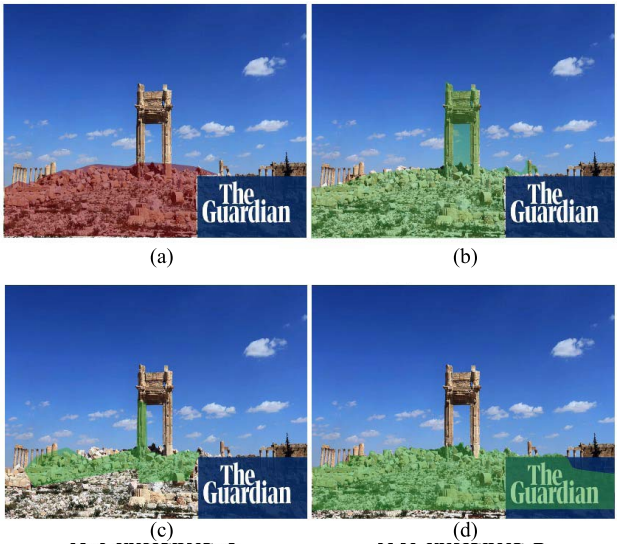


Fig. 2. Example of imperfect CI in CHDA applications. (a) Ground truth. (b) Worker 1. (c) Worker 2. (d) Worker 3.

leverage the imperfect CI to effectively address the failure cases of the AI model in CHDA applications.

To address the above-mentioned challenges, we develop *CollabLearn*, an uncertainty-aware crowd-AI collaboration system that explicitly explores the imperfect CI to identify and fix the AI failure cases in CHDA applications. In particular, our CollabLearn jointly models the uncertainty from both AI and CI under a *unified* framework to solve the CHDA problem. To address the first challenge, we develop an uncertainty-aware deep damage assessment (UDDA) model to quantify the uncertainty of the estimated damage areas and detect the failure cases of AI. To address the second challenge, we design a novel crowd-AI fusion model that integrates the uncertainty of both AI models and crowd responses into a holistic estimation framework that addresses the failure cases of AI and improves the overall damage assessment accuracy in CHDA. To the best of our knowledge, our CollabLearn is the first integrated crowd-AI collaboration system that explicitly explores the collective power of *uncertain* AI models and *imperfect* CI under the same analytical framework to address the CHDA problem. We evaluate the CollabLearn using a set of real-world CHDA datasets from seven world-renowned cultural heritage sites that were recently damaged. The evaluation results show that CollabLearn consistently outperforms both state-of-the-art AI approaches and crowd-AI baselines in correctly identifying the cultural heritage damages under diversified types of cultural heritage sites and evaluation scenarios. We summarize our main contributions as follows.

1) We study an important *CHDA* problem that aims to protect and conserve cultural heritage sites by exploring the collective power of uncertain AI models and imperfect CI.
2) We develop CollabLearn, the first uncertainty-aware crowd-AI collaboration system in CHDA applications to address two important technical challenges, i.e., *identification of AI failure cases* and *imperfect CI*, under a unified analytical framework.
3) We perform extensive experiments to evaluate CollabLearn through real-world case studies from seven world-renowned and recently damaged cultural heritage sites and the results demonstrate clear performance gains of our CollabLearn scheme compared with state-of-the-art baselines.

The rest of this article is organized as follows. We first review the related work in Section II. In Section III, we formally define our crowd-AI CHDA problem. The proposed CollabLearn framework is elaborated in Section IV. Experiments and evaluation results are presented in Section V. Finally, we conclude this article in Section VI.

II. Related Work

A. Crowdsourcing

Crowdsourcing has emerged as a new application paradigm, where individual workers work collaboratively to address some challenging problems [24], [25]. Examples of crowdsourcing applications include enhancing driver situation awareness using participatory sensing [26], monitoring infectious disease outbreaks using real-time mobile crowdsensing [27], detecting ongoing cyber-attacks using social media feeds [28], and obtaining situational awareness in disaster response using social sensing [29]. A comprehensive summary of crowdsourcing applications can be found in [30]. Several key challenges exist in current crowdsourcing applications, including data reliability, incentive design, data scarcity, human–computer interaction, and privacy protection [31]–[34]. However, leveraging the imperfect CI to identify and fix the AI failure cases in CHDA applications remains to be a challenging problem in crowdsourcing applications. In this article, we developed the CollabLearn scheme to address this problem by designing a novel crowd-AI collaboration system to boost the CHDA performance. Our work is also related to the recent efforts in obtaining reliable information from unreliable crowd-sourced data [35], [36]. However, those solutions primarily focus on fusing the labels from different crowd workers and do not explore the collaboration between CI and AI, which often leads to suboptimal system performance [37]. In contrast, our CollabLearn develops a unified analytical framework to explicitly model the uncertainty of both AI models and crowd responses to address the failure cases of AI and optimize the performance of CHDA applications.

B. Social Media-Based Damage Assessment

Previous efforts have been made to address the damage assessment problem using social media data [3], [6], [7], [38]–[42]. For example, Li *et al.* [7] proposed a deep convolutional network approach to classify the severity levels of the damage based on social media images during natural disasters. Mouzannar *et al.* [38] developed a deep learning framework that utilizes heterogeneous social media data to obtain situation awareness in disaster response via multimodal convolutional neural networks. Kumar *et al.* [3] designed an end-to-end social media image processing and analytical model to identify the disaster damage images on social media using deep neural networks. However, those approaches cannot

be directly applied to solve our CHDA problem due to the complex nature of the cultural heritage damages and the lack of cultural background knowledge of those AI-based solutions. There also exist a couple of initial efforts that leverage human intelligence to identify and address the failure cases of AI in DDA [19], [20]. However, those human-assisted AI systems often rely on accurate human labels to troubleshoot and retrain the AI models to optimize the model performance. The inconsistent and uncertain crowd labels could cause a potential model collapse in the retraining process of those models. In contrast, this article explores the uncertainty of both AI models and crowd responses and integrates them into a holistic uncertainty-aware estimation framework to address the failure cases of AI in CHDA applications.

C. Crowd-AI Hybrid Systems

Our work belongs to the growing trend of designing the crowd-AI hybrid systems to solve the complex real-world problems [11], [15], [43]–[46]. For example, Jarrett *et al.* [15] developed an elastic crowd-AI learning framework that introduces a task complexity index to optimize the integration of AI and CI to improve the overall task performance in a mobile face recognition application. Sener and Savarese [43] proposed a deep core-set selection approach that collects crowd labels from a subset of representative images to retrain the AI models to improve the overall accuracy in natural scene image classification tasks. Zhang *et al.* [11] designed a crowd-AI hybrid system that leverages CI to retrain the AI models and combine crowd labels with AI outputs to troubleshoot and tune the performance of AI algorithms in DDA applications. Guo *et al.* [44] designed a crowd-AI hybrid question-answering system in smart home applications by analyzing the content from camera steam captured by smart IoT devices. Yang *et al.* [45] proposed an interactive framework to leverage crowdsourcing platforms and a deep probabilistic model to denoise the data in movie reviews and news articles. Current crowd-AI solutions often rely on a committee of different AI models to identify the problematic cases when those models do not agree with each other. However, those approaches would likely fail when all members of the committee happen to make similar mistakes on the same input due to the lack of cultural background knowledge [12]. More importantly, we observe that current crowd-AI approaches often retrain the AI models with additional labels from the crowd workers to improve their performance. However, we find that such a retraining mechanism does not work well with the imperfect labels on cultural heritage damage images obtained from the crowd due to the complex nature of cultural heritage damage, which can be easily confused with specific cultural and artistic designs and long-term aging effects that are often observed at the cultural heritage sites. In contrast, CollabLearn is the first crowd-AI collaboration system that explicitly explores the collective power of *uncertain* AI models and *imperfect* CI to boost the assessment accuracy in CHDA applications.

D. Deep Learning-Based Image Processing and Analytics

Our work also bears resemblance to the deep learning technique to automate the intelligent image processing and analytics in many real-world applications [47]. For example, Wang *et al.* [8] proposed a semantic reranking framework that leverages the deep features extracted by convolutional neural networks to improve the sketch-based image retrieval performance. Ronneberger *et al.* [48] designed a skip-connected convolutional neural network that utilizes both contracting and expanding paths to enable cross-layer information transmission for biomedical image segmentation. Xie *et al.* [9] proposed an image classification framework to classify the building damage status during a natural disaster from satellite radar images via ensemble models and deep learning networks. Zhu *et al.* [10] developed a multimodal hypergraph learning approach that leverages vertices and hyperedges in hypergraphs to capture the complex similarities between different landmarks in content-based landmark image searching. Li *et al.* [49] proposed a deep feature aggregation framework that aggregates discriminative features from different subnetworks to achieve a fast model convergence for semantic image segmentation. While the above-mentioned solutions focused on developing deep learning models to optimize the performance of specific applications, they are not designed to accurately detect the failure cases of the deep learning models in the absence of the ground truth labels. In contrast, CollabLearn designs a UDDA network to accurately quantify the uncertainty of the estimated results to detect the failure cases of the deep learning models in CHDA applications.

E. Social Computing

Our work is also related to the recent advances in social computing, which have been successfully applied in many application domains, such as human–robot interaction, the Internet-of-Things (IoT), public health, and information diffusion [50]–[53]. For example, Erol *et al.* [50] proposed an affection-based perception system that enables social robots to recognize human emotional states to improve the personalization in human–robot interaction. Liu *et al.* [51] introduced an edge-cloud collaborative computing system to improve energy efficiency and reduce system latency in face detection and recognition using field-programmable gate array-based CNN accelerators. Zhu *et al.* [52] developed an attentive deep recurrent framework for daily mental-state monitoring of depression patients by examining the dynamics of human blood vascular systems using Photoplethysmography. Dong *et al.* [53] designed a social media information flow model to track the information spread during disaster events and study the influence of different social media user groups on disaster information dissemination. To the best of our knowledge, our CollabLearn is the first crowd-AI collaboration system that explicitly explores the uncertainty of both AI and crowd responses in a unified analytical framework to address a real-world issue that has an important social impact—cultural heritage protection and conservation.

III. Problem Description

In this section, we formally define our crowd-AI CHDA problem. We first define a few key terms used in the problem formulation.

Fig. 3. Examples of cultural heritage damage images.

Definition 1 (Cultural Heritage Damage Images (X)): We define X to be the set of cultural heritage damage images posted on social media (e.g., Twitter), where each image captures a specific scene of a damaged cultural heritage site, as shown in Fig. 3. In particular, we collect the cultural heritage damage images from social media sites using a cultural heritage imagery data crawler tool [3], where each collected image contains the damage of a cultural heritage site from a recent damaging event (e.g., disaster and war). In addition, we define $X = \{X_1, X_2, \ldots, X_A\}$ as a set of collected cultural heritage damage images, where A represents the number of collected images.

Definition 2 (Actual Damage Area (D)): We define D as the actual damage areas in cultural heritage site images (e.g., the red color areas, as shown in Fig. 3). In particular, we define $D = \{D_1, D_2, \ldots, D_A\}$ to represent the *actual* damage areas in all collected images, where D_a represents the actual damage area of the ath image.

Definition 3 (Estimated Damage Area by AI ($\widehat{D^{\mathrm{AI}}}$)): We define $\widehat{D^{\mathrm{AI}}}$ as the damage areas estimated by the AI module of the crowd-AI collaboration system for the cultural heritage site images. In particular, we define $\widehat{D_a^{\mathrm{AI}}}$ as the *estimated* damage area of the ath image.

Definition 4 (Marked Damage Area by Crowd ($\widehat{D^{\mathrm{CI}}}$)): We define $\widehat{D^{\mathrm{CI}}}$ as the damage area annotated by the crowd workers from the crowdsourcing platforms (e.g., AMT). In particular, we define $\widehat{D_a^{\mathrm{CI}}}$ to represent the *marked* damage area by a crowd worker for the cultural heritage image X_a.

Definition 5 (Crowd Query (Q)]): We define a crowd query to be a crowdsourcing task where our crowd-AI collaboration system decides to send a set of cultural heritage damage images to the crowdsourcing platform, where each image in the crowd query is marked by a set of N crowd workers on the damaged area in the image as follows:

$$Q(X_a) = \left\{\widehat{D_a^{\mathrm{CI}}}(1), \widehat{D_a^{\mathrm{CI}}}(2), \ldots, \widehat{D_a^{\mathrm{CI}}}(N)\right\} \tag{1}$$

where $\widehat{D_a^{\mathrm{CI}}}(n)$ indicates the damage area marked by the nth crowd worker for the image X_a. We note that the damage areas marked by different crowd workers in each crowd query could be uncertain and inconsistent due to the complex nature of the cultural heritage damage and the uncertainty of the crowd workers, as shown in Fig. 2.

Definition 6 (Crowd Query Ratio (θ)): We define θ to be an application-specific parameter that specifies the percentage of cultural heritage damage images that is sent in a crowd query, which is often decided by the performance and budget tradeoff of a CHDA application. In other words, a total of $\theta \cdot A$ images will be sent for the crowd to mark in a crowd query.

Definition 7 (Identified Damage Area by Crowd-AI Collaboration System ($\widehat{D}$)): We define $\widehat{D}$ to be the final identified damage area from our crowd-AI collaboration system by leveraging both the estimated damage area generated by AI module $\widehat{D^{\mathrm{AI}}}$ and marked damage areas returned by crowd query $\widehat{D^{\mathrm{CI}}}$. In particular, we define $\widehat{D_a}$ to represent the final identified damage area for the collected image X_a.

The goal of our article is to accurately assess the damage of the cultural heritage sites by identifying the damage areas of images of the sites through collective intelligence from both AI and crowd. Given the above-mentioned definitions, we formally define our problem as follows:

$$\arg\max_{\widehat{D_a}} (\Gamma(\widehat{D_a}, D_a) \mid X, Q, N, \theta) \quad \forall 1 \le a \le A \tag{2}$$

where $\Gamma(\cdot)$ represents the quantitative metrics (e.g., IoU and DSC [54]) to measure the similarity between the identified and actual damage area ($\widehat{D_a}$ and D_a) of an image. This problem is challenging due to the difficulty of effectively detecting the failure cases of AI in the absence of the ground truth labels and the imperfect knowledge obtained from crowdsourcing platforms. In this article, we develop a CollabLearn framework to address these challenges, which is elaborated in Section IV.

IV. Solution

A. Overview of CollabLearn Framework

CollabLearn is an uncertain-aware crowd-AI collaboration system to address the CHDA problem formulated earlier. The overview of the CollabLearn is shown in Fig. 4. In particular, it consists of two modules.

1) *Uncertainty-Aware Deep Damage Assessment:* First, the UDDA module designs a novel deep damage assessment model to accurately quantify the uncertainty of the estimated damage areas and detect the failure cases of AI in CHDA applications. In particular, the *UDDA* module designs a duo-branch deep estimation network that contains two parallel output branches to simultaneously generate the damage area estimations together with the quantification of the estimation uncertainty under a unified network architecture. More importantly, to ensure the accuracy of the uncertainty quantification, we design an uncertainty-aware loss function to model the error of the estimated damage area and accurately quantify the uncertainty of the estimation within the deep network optimization process.
2) *Imperfect Crowd Knowledge Fusion:* Second, the imperfect crowd knowledge fusion (ICKF) develops a confidence-aware estimation framework to explicitly model the uncertainty of both AI models and crowd responses to address the failure cases of AI and optimize the performance of CollabLearn. In particular, the ICKF module first designs a novel crowd annotation portal on AMT by allowing the crowd workers to document their confidence in their marked damaged areas, which is essential to obtain accurate CI given the complex

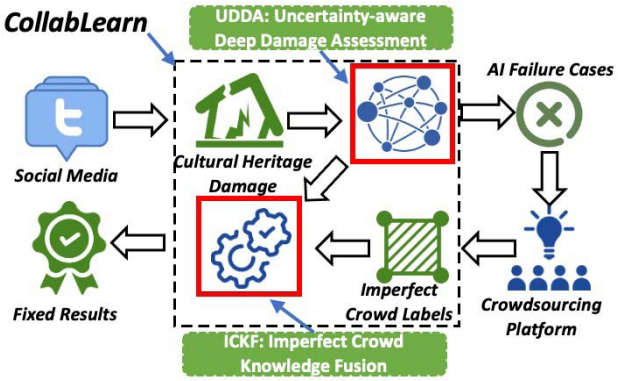


Fig. 4. Overview of CollabLearn framework.

cultural heritage damage and the unvetted nature of the crowd workers. More importantly, we further design a novel confidence-aware maximum likelihood estimation (MLE) model to leverage the inconsistent crowd responses with different confidence levels to derive the accurate damage areas of cultural heritage sites to fix the failure cases of AI.

B. Uncertainty-Aware Deep Damage Assessment

In this section, we present the UDDA network architecture in CollabLearn to estimate the damaged area in each cultural heritage image and quantify the uncertainty of the estimation results. In particular, our UDDA network architecture design consists of two network components: an encoder network (EN) and an assessment network (AN). In particular, the EN is first used to extract both high-level (e.g., objects and patterns) and low-level (e.g., colors and textures) damage-related visual features from the cultural heritage images. The AN is then used to explicitly identify the damaged areas and quantify the uncertainty of the estimation results using the multilevel visual features extracted by EN. To the best of our knowledge, the UDDA is the first end-to-end AI-based damage assessment approach that designs a multibranch uncertainty estimation network architecture to detect the failure cases of AI in CHDA applications in the absence of ground truth labels.

We first define a key concept for our UDDA module as follows.

Definition 8 (Damage Estimation Uncertainty Matrix (M)): We first consider the error between the actual and estimated damage area by AI as follows:

$$\mathcal{L}_{\text{CE}}\left(D_a, \widehat{D_a^{\text{AI}}}\right) \tag{3}$$

where $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss [55] that measures the error between the actual and estimated damage area of a cultural heritage image. D_a is the actual damage area for an image X_a (Definition 2) and $\widehat{D_a^{\text{AI}}}$ is the estimated damage area for X_a (defined in Definition 3). We observe that such an error often follows a Gaussian distribution [56]:

$$\mathcal{L}_{\text{CE}}\left(D_a, \widehat{D_a^{\text{AI}}}\right) \sim \mathcal{N}\left(\mathbf{0}, M_a^2\right) \tag{4}$$

where M_a represents the estimation uncertainty matrix that indicates the standard deviation of the cross-entropy loss at all pixels for the cultural heritage damage image X_a. Specifically, we define $M = \{M_1, M_2, \ldots, M_A\}$ to be a set of the damage estimation uncertainty matrices for all cultural heritage images in a CHDA application.

Given the above-mentioned definition, let us first formally define the EN and the AN in our UDDA module as follows.

Definition 9 (Encoder Network): We define EN as a mapping network to extract multilevel damage-related visual features from the cultural heritage imagery data as follows:

$$V^X = \text{EN}(X) \tag{5}$$

where V^X is used to represent the extracted damage-related visual features. We show an example of EN in Fig. 5(A). It contains a stack of ImageNet pretrained convolutional layers for damage-related visual feature extraction. This is done to ensure the mapping network is capable of accurately identifying the complex visual features for an input cultural heritage image. In addition, we enable the skip connection in the EN (i.e., the dotted lines in Fig. 5), which is used to forward different levels of damage-related visual features extracted by EN to AN. The different levels of visual features can then be utilized by AN to effectively identify the damaged area for each cultural heritage image.

Definition 10 (Assessment Network): We define AN as a generation network that estimates the damaged area for each cultural heritage image and infers the estimation uncertainty matrix using the damage-related visual feature V^X extracted by EN

$$\left(\widehat{D^{\text{AI}}}, M\right) = \text{AN}(V^X) \tag{6}$$

where $\widehat{D^{\text{AI}}}$ is the estimated damage area generated by the AN and M is the set of damage estimation uncertainty matrices defined earlier. We show an example of AN in Fig. 5(B). In particular, the AN consists of a set of deconvolutional layers that explicitly identify the damaged area of an image by gradually examining the damage-related visual features. In addition, AN also includes a set of convolutional layers that fuse different levels of visual features extracted by EN through skip-connections. This is done to ensure both high-level (e.g., objects and patterns) and low-level (e.g., colors and textures) damage-related visual features are successfully forwarded from EN to AN for accurate damage area estimation. The key novelty of the AN lies in the parallel output branch design where each branch contains a convolutional layer and a sigmoid layer, as shown in Fig. 5. This design provides the estimation of the damaged area together with the quantification of the estimation uncertainty under an end-to-end network architecture.

Given the two network architectures earlier, our next question is how to define a loss function for our network to generate the damage assessment results and the estimation uncertainty matrix to quantify the accuracy of the results. To that end, we define two sets of loss functions in our model. In particular, we first consider the assessment loss for the EN and AN as follows:

$$\mathcal{L}_{\text{EN,AN}}^{\text{Assess}}: \mathcal{L}_{\text{CE}}(\text{AN}(\text{EN}(X)), D) \tag{7}$$

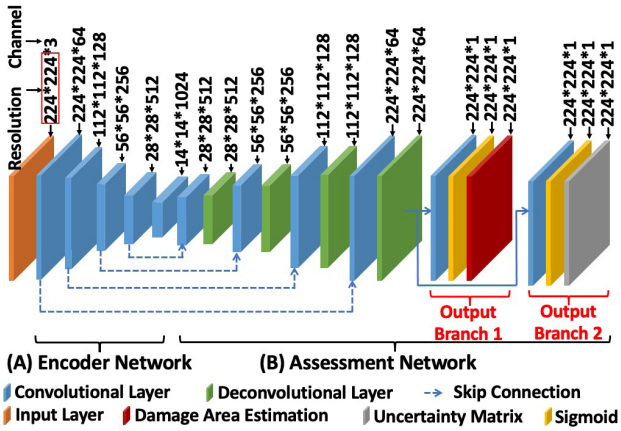


Fig. 5. Overall network architecture of UDDA. (A) Encoder network. (B) Assessment network.

where $\mathcal{L}_{\text{EN,AN}}^{\text{Assess}}$ represents the assessment loss function for EN and AN. $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss that measures the difference between the actual and estimated damage area of cultural heritage images. The goal of this loss function is to check if AN can accurately estimate the damage area of images using the visual features captured by EN.

Next, recall that the difference between the actual and estimated damage area [i.e., $\mathcal{L}_{\text{CE}}(\text{AN}(\text{EN}(X)), D)$] follows the Gaussian distribution [i.e., $\mathcal{N}(\mathbf{0}, M^2)$] in Definition 8. We can derive the log-likelihood function for $\mathcal{L}_{\text{CE}}(\text{AN}(\text{EN}(X)), D)$ as follows:

$$\begin{aligned}&\log \mathbb{L}(\mathbf{0}, M; \mathcal{L}_{\text{CE}}(\text{AN}(\text{EN}(X)), D))\\ &\quad = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log ||M||_2^2\\ &\qquad - \frac{1}{2||M||_2^2}||\mathcal{L}_{\text{CE}}(\text{AN}(\text{EN}(X)), D)||_2^2. \end{aligned} \tag{8}$$

Therefore, we define our uncertainty loss function as the negation of the log-likelihood function as follows:

$$\begin{aligned}\mathcal{L}_{\text{EN,AN}}^{\text{Uncertain}}: \min\Bigg(&\frac{1}{2}\big(\log 2\pi + \log||M||_2^2\\ &+\frac{1}{||M||_2^2}||\mathcal{L}_{\text{CE}}(\text{AN}(\text{EN}(X)), D)||_2^2\big)\Bigg). \end{aligned} \tag{9}$$

By minimizing the loss function $\mathcal{L}_{\text{EN,AN}}^{\text{Uncertain}}$ in our deep damage AN, we can obtain the uncertainty matrices M that maximize the likelihood function $\log \mathbb{L}(\mathbf{0}, M; \mathcal{L}_{\text{CE}}(\text{AN}(\text{EN}(X)), D))$ defined earlier.

We then combine the above-mentioned two sets of loss functions to derive the final loss function $\mathcal{L}_{\text{EN,AN}}^{\text{Final}}$ to generate the damage estimation results and the uncertainty matrix of the estimation for the UDDA module as follows:

$$\mathcal{L}_{\text{EN,AN}}^{\text{Final}}: \mathcal{L}_{\text{EN,AN}}^{\text{Access}} + \mathcal{L}_{\text{EN,AN}}^{\text{Uncertain}} \tag{10}$$

where $\mathcal{L}_{\text{EN,AN}}^{\text{Final}}$ is a summation of $\mathcal{L}_{\text{EN,AN}}^{\text{Access}}$ and $\mathcal{L}_{\text{EN,AN}}^{\text{Uncertain}}$. For the $\mathcal{L}_{\text{EN,AN}}^{\text{Access}}$, we follow the standard cross-entropy loss design that translates the matrix to a score by calculating the mean value of all the elements in the matrix. For $\mathcal{L}_{\text{EN, AN}}^{\text{Uncertain}}$, we translate the matrix to a score by calculating the L2 norm of the matrix.

Using the above-mentioned loss function, we can learn the optimal instances (i.e., EN* and AN*) using the RMSprop optimizer [57]. Finally, we use EN* and AN* to estimate the damage areas and the estimation uncertainty matrices for all input cultural heritage damage images X as follows:

$$(\widehat{D^{\text{AI}}}, M) = \text{AN}^*(\text{EN}^*(X)). \tag{11}$$

Given the estimated damage area and the associated uncertainty matrix learned by our UDDA module, our next step is to use them to determine the failure cases of the AI model and send the identified failure cases to the crowdsourcing platforms to obtain CI. In particular, a higher uncertainty value in the uncertainty matrix indicates the AI model is more uncertain about estimation results, where the estimation results on damage area are more likely to be inaccurate. Therefore, we define an uncertainty score Φ to determine which cultural heritage images should be added to the crowd query Q as follows.

Definition 11 (Uncertainty Score Φ): We define Φ_a to represent the uncertainty score of a cultural heritage image X_a as follows:

$$\Phi_a = \text{mean}(M_a) \tag{12}$$

where mean$(\cdot)$ indicates the mean value of all elements in a matrix. M_a is the uncertainty matrix of the image X_a.

Finally, we sort the uncertainty scores of all cultural heritage images and select the top $\theta \cdot A$ ranked images into the crowd query Q (θ refers to the crowd query ratio in Definition 6 and A is the number of studied images). For the images that are not added to the crowd query Q, we use the damaged area estimated by our AI module $\widehat{D^{\text{AI}}}$ as the output $\widehat{D}$ for those images.

C. Imperfect Crowd Knowledge Fusion

In Section IV-B, we present the UDDA module that identifies the failure cases of AI. Our next question is how to acquire accurate CI from the potentially unreliable crowd workers to fix the failure cases of AI. We note many current active learning-based crowd-AI approaches often retrain the AI models with additional labels from the annotators to improve their performance. However, we found that such a retraining mechanism does not work well with the imperfect labels on cultural heritage damage images obtained from the crowd. In particular, we compare the performance of three representative deep learning-based damage assessment baselines (i.e., UNet [48], FCN [58], and DFANet [49]) together with our UDDA module *with* and *without* retraining using the imperfect crowd labels. The results are shown in Fig. 6. We observe that the performance of all schemes decreases after they are retrained with the imperfect crowd labels. The reason for the decreased performance of AI models retrained by imperfect crowd labels is that the imperfect crowd labels could enforce the AI models to learn the incorrect visual characteristics about the damaged areas (e.g., mistakenly learn the visual characteristics of intact areas as the evidence for damaged areas). The results validate our hypothesis that simply retraining AI

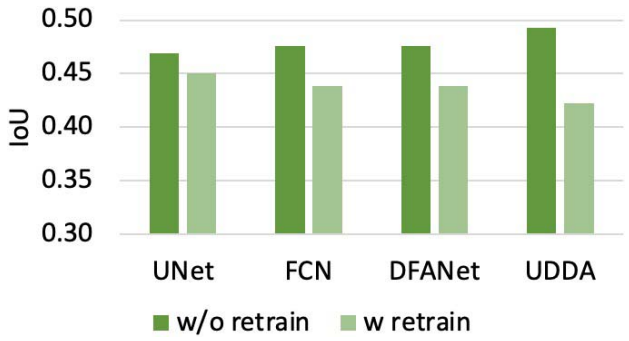


Fig. 6. Impact of imperfect CI on AI models.

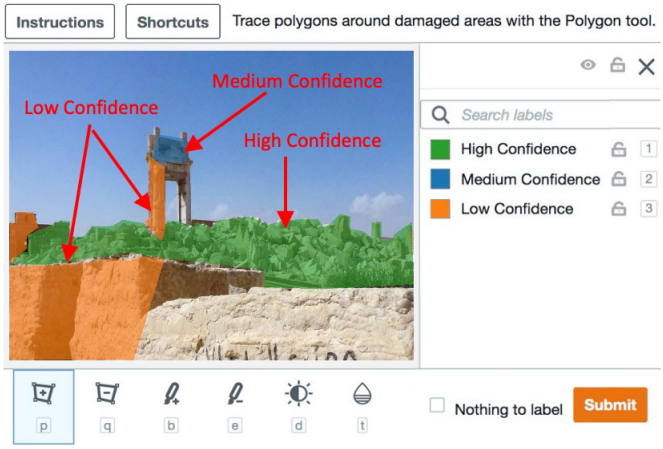


Fig. 7. Example of confidence-aware crowd annotation portal.

models with imperfect labels from the crowd may lead to suboptimal performance of the models.

To address the above-mentioned challenge, we design a crowd-AI fusion module that integrates the uncertainty estimation of the AI module and the imperfect crowd responses into a holistic estimation framework to improve the overall damage assessment accuracy. In particular, we first design a crowd annotation portal on AMT by allowing the crowd workers to document their *confidence* on their marked damage areas, as shown in Fig. 7. Such confidence-aware design is important due to the complex cultural heritage damage and imperfect nature of individual crowd workers. We note that different crowd workers could mark different areas as the damaged area and the same worker could express different levels of confidence for the marked areas. Our next question is how to obtain reliable CI by leveraging the inconsistent and uncertain responses from the individually unreliable crowd workers. To that end, we first define a key term as follows.

Definition 12 (Inferred Damage Area ($\overline{D^{\text{CI}}}$)): we define $\overline{D^{\text{CI}}}$ to represent the damage area inferred by our ICKF module using the responses from the crowd query Q. In particular, we define $\overline{D_a^{\text{CI}}}$ to be the inferred damage area for cultural heritage image X_a in the crowd query Q.

Given the definition earlier, our problem of deriving the inferred damage area using the inconsistent crowd response can be nicely formulated as an MLE problem as follows:

$$\Pr\left(\left(\widehat{D_a^{\text{CI}}}(1),\ldots,\widehat{D_a^{\text{CI}}}(N)\right),\left(\text{CF}_a(1),\ldots,\text{CF}_a(N)\right)|\overline{D_a^{\text{CI}}}\right) \quad (13)$$

where $\widehat{D_a^{\text{CI}}}(n)$ represents the damage area marked by a crowd worker in a crowd query (Definition 5) for cultural heritage image X_a. $\text{CF}_a(n)$ indicates the associated confidence-level for the $\widehat{D_a^{\text{CI}}}(n)$, as shown in Fig. 7. Our goal here is to estimate the likelihood of each pixel of an image being a part of the damage area in X_a given the crowd responses and associated confidence levels, which collectively help infer $\overline{D_a^{\text{CI}}}$. In particular, we first define a likelihood function $L(\Omega; O, Z)$ as follows:

$$\begin{aligned}
&\mathbb{L}(\Omega; O, Z)\\
&= \mathbb{L}\Big(\Omega; \Big(\big(\widehat{D_a^{\text{CI}}}(1),\ldots,\widehat{D_a^{\text{CI}}}(N)\big),\\
&\qquad \big(\text{CF}_a(1),\ldots,\text{CF}_a(N)\big)\Big), \overline{D_a^{\text{CI}}}\Big)\\
&= \prod_{p=1}^{P}\left(\prod_{n=1}^{N}\left(\prod_{c=1}^{C}\alpha_{n,c}^{U_{n,p}\&\&V_{n,p}^{c}}\right)\times\left(1-\sum_{c=1}^{C}\alpha_{n,c}\right)^{(1-U_{n,p})}\right.\\
&\qquad \times d\times z_p + \prod_{n=1}^{N}\left(\prod_{c=1}^{C}\beta_{n,c}^{U_{n,p}\&\&V_{n,p}^{c}}\right)\\
&\qquad \left.\times\left(1-\sum_{c=1}^{C}\beta_{n,c}\right)^{(1-U_{n,p})}\times(1-d)\times(1-z_p)\right). \quad (14)
\end{aligned}$$

The above-mentioned likelihood function represents the likelihood of the observed data O (i.e., damage areas marked by different workers) with different confidence levels and the values of hidden variables Z (i.e., the damage area of an image) given the estimated parameter Ω. The detailed explanations of the above-mentioned parameters of the likelihood function are summarized in Table I.

The objective of our problem is to infer the accurate damaged area $\overline{D_a^{\text{CI}}}$ by deriving the values of the hidden variable z_p that indicates whether a specific pixel p of an image belongs to a part of the damaged area. In particular, the formulated problem can be solved using expectation maximization (EM). However, one key issue for the EM algorithm is that the algorithm is often sensitive to the initialization of the model parameters, which may lead the algorithm to a suboptimal solution.

To address this problem, we leverage the uncertainty estimation generated by our UDDA module to help the EM algorithm with a better parameter initialization that maximizes the chance of the algorithm to reach an optimized solution. In particular, we first define a key term as follows.

Definition 13 (Reliable AI Estimation Area (δ_a)): We define δ_a to represent the subarea in a cultural heritage image X_a with top k percent lowest uncertainty values in the estimation uncertainty matrix M_a, which often indicates that the AI module is certain about the estimation results in δ_a The value of k is often set to be small (e.g., 10 in our experiments) in practice to ensure the estimation results from the AI module are reliable.

Leveraging δ_a, we then set the value of z_p for each pixel p within δ_a to be the same as the assessment result from the

TABLE I
NOTATIONS IN IMPERFECT CI FUSION

Notations	Definitions/Explanations
P	number of pixels in a cultural heritage image
N	number of crowd workers for each crowd query
C	number of confidence levels in our crowdsourcing portal
$\alpha_{n,c}$	conditional probability that a crowd worker n marks a pixel to be a part of the damage area with a confidence level of c given the pixel is a part of the damage area
$\beta_{n,c}$	conditional probability that a crowd worker n marks a pixel to be a part of the damage area with a confidence level of c given the pixel is not a part of the damage area
$U_{n,p}$	indicator variable that is set to be 1 when a crowd worker n marks a pixel p to be a part of the damage area and is set to be 0 otherwise
$V_{n,p}^{c}$	indicator variable that is set to be 1 when a crowd worker reports a pixel p to be a part of the damage area with a confidence level of c and is set to be 0 otherwise.
&&	"logical and" operation
d	prior probability that a randomly chosen pixel is a part of the damage area
z_p	probability that whether a specific pixel p is part of the damage area or not
Ω	estimation parameter of the model, where $\Omega = \{\alpha_{1,c}, \alpha_{2,c}, ..., \alpha_{N,c}; \beta_{1,c}, \beta_{2,c}, ..., \beta_{N,c}, d\}$ *for* $c = 1, 2..., C$
O	observed variable of the model, where $O = (\widehat{D_a^{CI}}(1), ..., \widehat{D_a^{CI}}(2), \widehat{D_a^{CI}}(N)), (CF_a(1), CF_a(2), ..., CF_a(N))$
Z	latent variable of the model, which indicates the inferred damage area $\overline{D_a^{CI}}$

UDDA model (i.e., set z_p to be 1 if the pixel is estimated by AI as a part of the damaged area and 0 otherwise) in the initialization and iterative process of the EM algorithm. We can then infer the damaged area of an image $\overline{D_a^{\mathrm{CI}}}$ in the crowd query from the learned z_p for each pixel. In particular, we exam the z_p for all pixels in an image and set all pixels p with $z_p > 0.5$ as the inferred damaged area. Finally, we use the inferred damage area $\overline{D^{\mathrm{CI}}}$ to replace the estimated damage area $\widehat{D^{\mathrm{AI}}}$ generated by the UDDA module for all images in the crowd query Q as the output $\widehat{D}$ of our CollabLearn framework, which fixes the failure cases of AI.

D. Summary of CollabLearn Framework

Finally, we summarize the CollabLearn framework in Algorithm 1. In particular, CollabLearn includes three main phases in performing the crowd-AI-based CHDA as follows.

1) *Model Training Phase*: The objective of this phase is to train an optimized UDDA network (i.e., EN* and AN*) that will be used in the later phases to detect the failure cases of AI and infer accurate labels on damage area from the crowd responses. In particular, our framework leverages labeled data to train the UDDA network by optimizing the final loss function [(10)] using the RMSprop optimizer [57].
2) *AI troubleshooting Phase*: Given the learned optimized EN* and AN*, our objective in this phase is to identify the failure cases of AI by selecting the images with a high uncertainty score Φ and adding those images to the crowd query Q. Note that our CollabLearn does not involve any network training during the AI troubleshooting phase. Instead, it utilizes the learned network instances (EN* and AN*) obtained from the model training phase to identify the damaged area of cultural heritage sites and generate the uncertainty estimation of the inferred damaged area. In addition, for the images that are not added to Q, we take the damaged area estimated by our AI module $\widehat{D^{\mathrm{AI}}}$ as the output $\widehat{D}$ of our CollabLearn framework.
3) *Crowd Knowledge Fusion Phase*: For the images in the crowd query Q, we first obtain the crowd responses $\widehat{D^{\mathrm{CI}}}$ from the crowdsourcing platform. Our objective of this phase is to integrate the uncertainty matrices M from the UDDA module and imperfect crowd response $\widehat{D^{\mathrm{CI}}}$ to infer the accurate damage area $\overline{D^{\mathrm{CI}}}$. The $\overline{D^{\mathrm{CI}}}$ will be used as the output $\widehat{D}$ of our CollabLearn framework for all cultural heritage images in the crowd query Q.

Algorithm 1 CollabLearn Framework Summary

▷ **Model Training Phase**
initialize EN (Definition 9)
initialize AN (Definition 10)
for each epoch **do**
  for each batch **do**
    optimize EN and AN (Equation (10))
  end for
end for
obtain EN^* and AN^*
▷ **AI Troubleshooting Phase**
obtain $\widehat{D^{AI}}$ and M using EN^* and AN^* (Equation (11))
for a in [1,2, . . . ,A] **do**
  if Φ_a in top $\theta \cdot A$ **then**
    add X_a to Q
  else
    set $\widehat{D_a^{AI}}$ as $\widehat{D_a}$
    add $\widehat{D_a}$ to $\widehat{D}$
  end if
end for
▷ **Crowd Knowledge Fusion Phase**
for each X_a in Q **do**
  obtain O from crowdsourcing platform
  drive $\overline{D_a^{CI}}$ by solving Equation (14) using EM
  set $\overline{D_a^{CI}}$ as $\widehat{D_a}$
  add $\widehat{D_a}$ to $\widehat{D}$
end for
output $\widehat{D}$

V. EVALUATION

In this section, we evaluate the performance of the CollabLearn framework using the real-world datasets on cultural heritage damages collected from seven different recently damaged cultural heritage sites. The results show that CollabLearn consistently outperforms the state-of-the-art AI-only and crowd-AI hybrid baselines in terms of CHDA accuracy under various application scenarios.

Fig. 8. Cultural heritage damage sites in our dataset.

Fig. 9. Examples of ground truth damage area annotation.

A. Dataset

1) Cultural Heritage Damage Dataset: In our evaluation, we use a real-world dataset on cultural heritage damage collected by [3].[1] In particular, the dataset consists of social media images collected from seven different recently damaged cultural heritage sites, as shown in Fig. 8. These cultural heritage damages have a diversified set of damage characteristics (e.g., damage types, affected areas, and building characteristics), which create a challenging evaluation scenario to study the cultural heritage assessment problem. The ground truth damage area in each cultural heritage image is manually annotated by domain experts using the image polygonal annotation tool Labelme.[2] In particular, for the seven recently damaged cultural heritage sites studied in our experiment, we first collect the validation images of intact cultural heritage sites using the online data sources (e.g., Google and Wikipedia). We then compare the damage images of the cultural heritage sites with the validation images to determine the ground truth damage area in each image. In addition, we show a few examples of such a ground truth damage area annotation process in Fig. 9. Please note that the above-mentioned ground truth dataset is used for the purpose of evaluation only and is often not available to the crowd-AI system due to the heavy labor costs and low efficiency of domain experts. In addition, we randomly sample the training and testing data from the dataset by setting the ratio of training to testing data as 1:1. Such a ratio is set to ensure that all compared schemes can be evaluated with a sufficient amount of testing data. A large testing set also makes it more challenging for all crowd-AI schemes (including CollabLearn) to identify the AI failure cases [59]. The training dataset is used to train all compared AI models for CHDA. In our experiments, we also study the robustness of the CollabLearn scheme and the baselines by varying the ratio between the training and testing data.

2) Amazon Mechanical Turk Platform: To obtain the CI, we utilize the AMT.[3] In our experiment, each image in a crowd query is marked by three independent crowd workers. To ensure the crowd label quality, we select the crowd workers who have an overall task approval rate greater than 95% and have completed at least 1000 approved tasks to participate in our crowd query task. We pay \$0.20 to each worker per image in our experiment. In each crowd query task, we ask the crowd workers to mark the damaged area for each image in the crowd query together with their *confidence* on their marked damage areas as shown in Fig. 7 in Section IV-C. In our experiments, we vary the crowd query ratio (Definition 6) from 10% to 25% in our experiments. We also set the number of crowd workers who respond to each queried image to be 3 for all compared schemes, which achieves a reasonable balance between the number of crowd labels and the query cost. We have followed the corresponding IRB protocol of this research.

B. Baselines and Experiment Settings

We compare CollabLearn with a set of representative AI and crowd-AI baselines that are widely used in the literature for the damage assessment using social media data images.

1) *AI Baselines:*
   a) *UNet [48]:* A widely used deep neural network approach that utilizes both contracting and expanding paths to enable cross-layer information transmission for desirable CHDA accuracy.
   b) *FCN [58]:* A deep learning model that utilizes the fully convolutional neural networks to map the cultural heritage imagery data into a latent deep feature space to infer the damaged area.
   c) *Attention UNet [60]:* A recent deep convolutional model that integrates the U-Net model with an attention gate mechanism to improve the model sensitivity in detecting the cultural heritage damages.
   d) *DFANet [49]:* A deep segmentation network that aggregates discriminative features from different subnetworks to achieve a fast model convergence speed for the CHDA task.

[1] https://crisisnlp.qcri.org/heritage
[2] https://github.com/wkentaro/labelme
[3] https://www.mturk.com/

TABLE II
PERFORMANCE COMPARISONS ON DAMAGE ASSESSMENT ACCURACY

		θ = 10%		θ = 15%		θ = 20%		θ = 25%	
Category	Algorithm	IoU	DSC	IoU	DSC	IoU	DSC	IoU	DSC
Random	Random	0.2050	0.3277	0.1989	0.3198	0.2154	0.3294	0.2023	0.3251
AI-Only	UNet	0.4115	0.5540	0.4500	0.5915	0.4244	0.5562	0.4928	0.6311
	FCN	0.4069	0.5507	0.4384	0.5797	0.4662	0.6058	0.4879	0.6269
	AttentionUNet	0.3329	0.4535	0.3633	0.4835	0.3684	0.4911	0.3962	0.5179
	DFANet	0.3839	0.5539	0.3692	0.5382	0.3677	0.5371	0.3810	0.5510
Crowd-AI	Hybrid Para	0.4963	0.6278	0.5027	0.6337	0.5162	0.6457	0.5205	0.6489
	Deep Active	0.3856	0.5263	0.3970	0.5389	0.4273	0.5706	0.4580	0.5970
	CrowdLearn	0.4894	0.6248	0.4987	0.6322	0.5074	0.6406	0.5171	0.6504
Our Model	**CollabLearn**	**0.5298**	**0.6580**	**0.5490**	**0.6763**	**0.5581**	**0.6838**	**0.5686**	**0.6924**

2) *Crowd-AI Hybrid Baselines:*
   a) *Hybrid Para [15]:* An elastic crowd-AI learning architecture that allocates the imagery data with complex image property (e.g., the images with large size and complex color distributions) for the crowd to label in order to improve the overall assessment accuracy of CHDA application.
   b) *Deep Active [43]:* A deep active learning-based crowd-AI system that utilizes the deep features extracted from each cultural heritage image to identify the representative ones for crowd labeling to retrain the AI models for performance optimization.
   c) *CrowdLearn [11]:* A recent crowd-AI framework that explores the CI and AI by directly combining crowd labels with AI outputs to improve the accuracy of the estimated labels on cultural heritage damage.

To ensure a fair comparison, the inputs to all compared schemes are set to be the same, which include: 1) the collected social media images; 2) the ground truth labels of images in the training dataset; and 3) the labeled images from crowd workers. In particular, we retrain the AI baselines using the labels returned by the crowd for a fair comparison. In addition, we also consider the *random* baseline, which estimates the damaged area for each image by randomly determining whether each pixel in the image is a part of the damaged area or not. In our experiments, we implement our model using PyTorch 1.1.0 libraries[4] and train our model using the NVIDIA Quadro RTX 6000 GPUs. In our experiments, all hyperparameters are optimized using the RMSprop optimizer [57]. In particular, we set the learning rate to be 10^{-4}. We also set the batch size to be 10 and the model is trained over 300 epochs.

To evaluate the performance of all compared schemes, we adopt two representative metrics that are widely used to study the performance of the object detection in image processing and computer vision community [54]. In particular, the two metrics measure the overlap between the estimated and actual damage area as: 1) intersection over union (IoU) = (Intersection/Union) and 2) dice similarity coefficient (DSC) = $(2 * \text{Intersection}/\text{Intersection} + \text{Union})$, where *intersection* and *union* represent the intersection and union between the inferred damage area and the actual damage area, respectively. Intuitively, a higher IOU or DSC value indicates a better performance in identifying the damage area of a cultural heritage image.

[4]https://pytorch.org/

C. Evaluation Results

1) Performance Comparisons on Cultural Heritage Damage Assessment: In the first set of experiments, we evaluate the accuracy of all compared schemes in estimating the damaged area of a cultural heritage image. In particular, we study the performance of all compared schemes by varying the crowd query ratio θ (the percentage of images that are sent to the crowdsourcing platform for labeling defined in Definition 6) from 10% to 25%, which achieves a reasonable tradeoff between the number of crowd responses and the query cost. In particular, we set the lower bound of the crowd query ratio in our experiment to be 10% to ensure that the CollabLearn can acquire sufficient crowd labels to fix the failure cases of AI. In addition, we set the upper bound of the crowd query ratio to be 25% because the performance of CollabLearn plateaus when the crowd query ratio reaches 25% and further increasing the crowd query ratio and query cost will not further improve the performance of CollabLearn. The evaluation results are presented in Table II. We observe that the CollabLearn scheme consistently outperforms all compared baselines. For example, the performance gain of CollabLearn compared with the best-performing baseline (i.e., hybrid para) when the crowd query ratio θ = 25% on IoU and DSC are 4.81% and 4.35%, respectively. Such performance gains mainly come from the fact that our CollabLearn scheme carefully explores the uncertainty of both the deep learning model and CI under

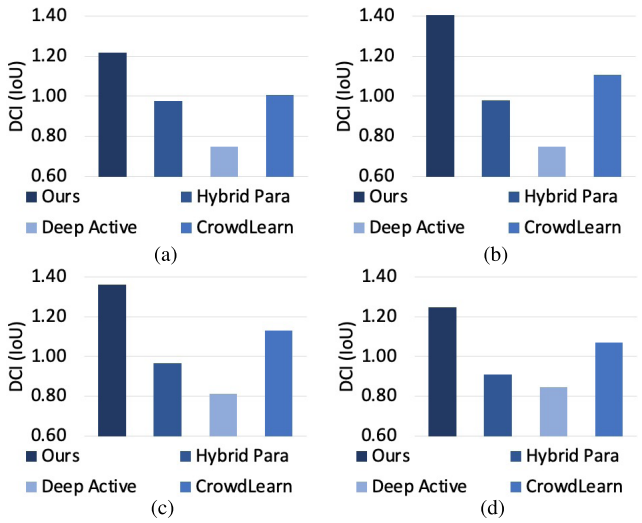


Fig. 10. Effectiveness of uncertainty-aware damage assessment [DCI (IoU)]. (a) Crowd query ratio = 10% with DCI (IoU). (b) Crowd query ratio = 15% with DCI (IoU). (c) Crowd query ratio = 20% with DCI (IoU). (d) Crowd query ratio = 25% with DCI (IoU).

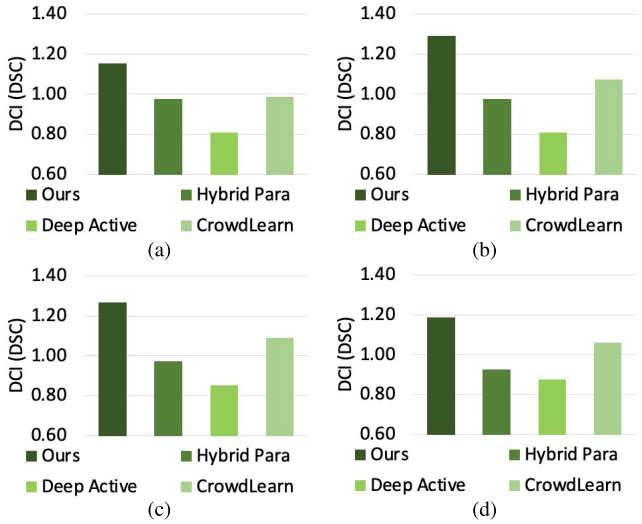


Fig. 11. Effectiveness of uncertainty-aware damage assessment [DCI (DSC)]. (a) Crowd query ratio = 10% with DCI (DSC). (b) Crowd query ratio = 15% with DCI (DSC). (c) Crowd query ratio = 20% with DCI (DSC). (d) Crowd query ratio = 25% with DCI (DSC).

a holistic estimation framework and collectively improve the overall damage assessment accuracy. We also observe that the performance of our CollabLearn scheme improves when the crowd query ratio increases. This is because, with a larger crowd query ratio, more problematic AI cases will be identified by the UDDA module and fixed by the crowd via the ICKF module of our solution. The above-mentioned results demonstrate the effervesces of our CollabLearn in leveraging the imperfect CI to carefully address the failure cases of AI to boost the accuracy of CHDA applications.

2) Effectiveness of AI Failure Detection: In the second set of experiment, we evaluate the effectiveness of our CollabLearn in identifying the failure cases of the AI model. In particular, we first introduce a new metric—detection efficiency index (DCI). In particular, we have

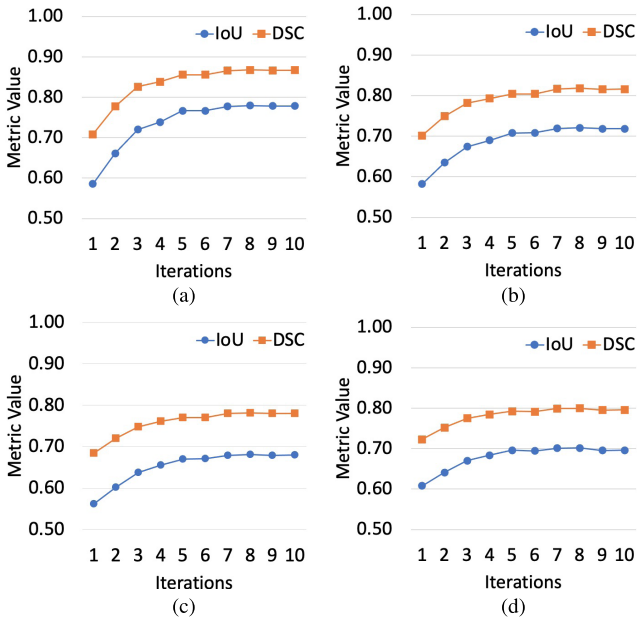


Fig. 12. Convergence of CollabLearn scheme. (a) Crowd query ratio = 10%. (b) Crowd query ratio = 15%. (c) Crowd query ratio = 20%. (d) Crowd query ratio = 25%.

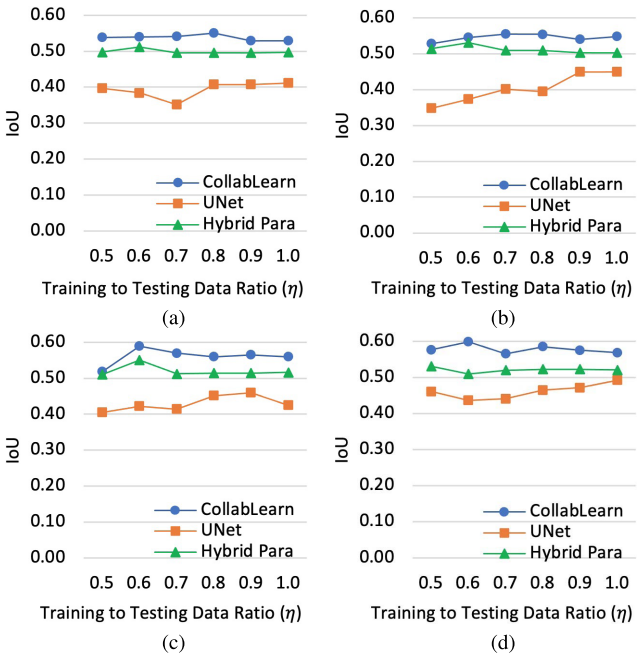


Fig. 13. Robustness of CollabLearn scheme (IoU). (a) Crowd query ratio = 10%. (b) Crowd query ratio = 15%. (c) Crowd query ratio = 20%. (d) Crowd query ratio = 25%.

$\mathrm{DCI}(\Delta) = (\Delta(\text{nonselected})/\Delta(\text{selected}))$. The *selected* indicates the set of images that the scheme estimated as the failure cases of AI, which are selected for the crowd query. The *nonselected* indicates the remaining images that are not selected for the crowd query. Δ indicates the mean value of the IoU or DSC of the images in the set. Intuitively, a higher DCI value indicates that the crowd-AI schemes are more effective in identifying the AI failure cases (i.e., the images selected for the crowd query have much lower IoU/DSC values compared to the ones that are not selected). The results are

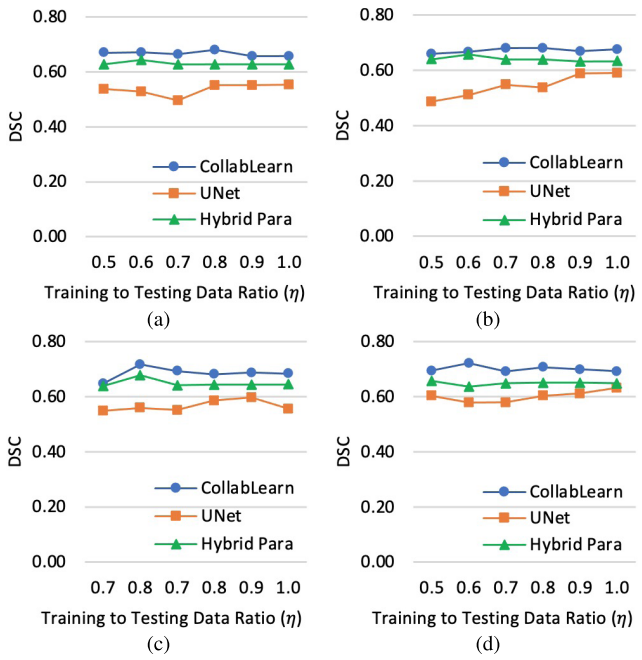


Fig. 14. Robustness of CollabLearn scheme (DSC). (a) Crowd query ratio = 10%. (b) Crowd query ratio = 15%. (c) Crowd query ratio = 20%. (d) Crowd query ratio = 25%.

shown in Figs. 10 and 11. Note that we only compare our CollabLearn with the crowd-AI Hybrid baselines because the current AI-only solutions often do not have a mechanism for detecting their failure cases. We observe that the CollabLearn clearly outperforms all compared schemes by achieving the highest DCI in all evaluation scenarios. The above-mentioned results further validate the effectiveness of the UDDA module in our CollabLearn framework.

3) Convergence Study of Imperfect Crowd Knowledge Fusion: In the third set of experiments, we evaluate the convergence of the ICKF module of our CollabLearn framework in inferring the accurate damage area of images from the imperfect crowd responses. In particular, we show the convergence of our ICKF module in learning the inferred damaged areas over different iterations. The results are shown in Fig. 12. We observe that our CollabLearn can quickly boost the assessment performance and remain stable afterward. The results are similar across different metrics and crowd query ratios. Such results illustrate the effectiveness of the ICKF module in CollabLearn in leveraging the uncertainty estimation from both AI and crowd responses to derive accurate labels on damaged areas of the queried images to improve the overall performance of CHDA applications.

4) Robustness Study of CollabLearn Scheme: In the last set of experiments, we study the robustness of the CollabLearn scheme by varying the ratio between the training and testing data. In particular, we first define the training to testing data ratio: η = (number of testing images/number of training images). We compare the performance of the CollabLearn with the best-performing baselines from both AI-only and crowd-AI hybrid categories (i.e., UNet from the AI-only and hybrid para from the crowd-AI). The results are shown in Figs. 13 and 14.[5] We observe that the performance of our CollabLearn scheme is relatively stable as training to testing data ratio changes under different crowd query settings. The results demonstrate the robustness of our scheme over various evaluation settings. We also observe that CollabLearn consistently outperforms the best-performing baselines on different evaluation metrics, which further demonstrates the effectiveness of CollabLearn in optimizing the performance of CHDA applications under a unified crowd-AI analytical framework.

VI. Conclusion

This article presents a CollabLearn framework to address a CHDA problem by exploring the collective intelligence from both AI and crowd under a unified analytical framework. CollabLearn addresses two key challenges: the identification of AI failures cases without ground truth labels and the imperfect CI fusion. We develop an uncertainty-aware crowd-AI collaboration system to explicitly model the uncertainty of both AI models and crowd responses in a principled estimation framework and explore their complementary strengths to improve the overall performance of the CHDA applications. The results on the real-world CHDA applications show that CollabLearn consistently outperforms both AI-only and crowd-AI hybrid baselines in terms of the CHDA accuracy. We believe CollabLearn will provide useful insights to explore the integrated power of uncertain AI models and imperfect CI to boost the performance of a diversified set of complex intelligent computing systems (e.g., intelligent transportation, smart health, and social AI) where AI and CI are melded into a collaborative and mutually beneficial paradigm.

Acknowledgment

This research is supported in part by the National Science Foundation under Grant No. IIS-2008228, CNS-1845639, CNS-1831669, Army Research Office under Grant W911NF-17-1-0409. 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 Army Research Office or 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

[1] D. H. R. Spennemann, "Cultural heritage conservation during emergency management: Luxury or necessity?" *Int. J. Public Admin.*, vol. 22, no. 5, pp. 745–804, Jan. 1999.

[2] J. Taboroff, "Natural disasters and urban cultural heritage: A reassessment," in *Building Safer Cities*. Washington, DC, USA: World Bank, 2003, p. 233.

[3] P. Kumar, F. Ofli, M. Imran, and C. Castillo, "Detection of disaster-affected cultural heritage sites from social media images using deep learning techniques," *J. Comput. Cultural Heritage*, vol. 13, no. 3, pp. 1–31, Oct. 2020.

[4] D. Wang *et al.*, "Using humans as sensors: An estimation-theoretic perspective," in *Proc. 13th Int. Symp. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2014, pp. 35–46.

[5]Note that we start from $\eta = 0.5$ to ensure that both CollabLearn and the compared baselines can be evaluated with a sufficient amount of testing data.

[5] D. Wang, T. Abdelzaher, and L. Kaplan, *Social Sensing: Building Reliable Systems on Unreliable Data*. San Mateo, CA, USA: Morgan Kaufmann, 2015.

[6] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2017, pp. 569–576.

[7] X. Li, D. Caragea, H. Zhang, and M. Imran, "Localizing and quantifying damage in social media images," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 194–201.

[8] L. Wang, X. Qian, Y. Zhang, J. Shen, and X. Cao, "Enhancing sketch-based image retrieval by CNN semantic re-ranking," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3330–3342, May 2020.

[9] B. Xie *et al.*, "Machine learning on satellite radar images to estimate damages after natural disasters," in *Proc. 28th Int. Conf. Adv. Geographic Inf. Syst. (SIGSPATIAL)*. New York, NY, USA: ACM, 2020, pp. 461–464, doi: 10.1145/3397536.3422349.

[10] L. Zhu, J. Shen, H. Jin, R. Zheng, and L. Xie, "Content-based visual landmark search via multimodal hypergraph learning," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2756–2769, Dec. 2015.

[11] D. Zhang, Y. Zhang, Q. Li, T. Plummer, and D. Wang, "CrowdLearn: A crowd-AI hybrid system for deep learning-based damage assessment applications," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2019, pp. 1221–1232.

[12] N. Wang, X. Zhao, L. Wang, and Z. Zou, "Novel system for rapid investigation and damage detection in cultural heritage conservation based on deep learning," *J. Infrastruct. Syst.*, vol. 25, no. 3, Sep. 2019, Art. no. 04019020.

[13] A. Klamer, A. Mignosa, and L. Lyudmila, "Cultural heritage policies: A comparative perspective," in *Handbook on the Economics of Cultural Heritage*. Cheltenham, U.K.: Edward Elgar, 2013.

[14] D. McDuffie, "Using Amazon's mechanical Turk: Benefits, drawbacks, and suggestions," *APS Observer*, vol. 32, no. 2, pp. 34–35, 2019.

[15] B. Blake, J. Jarrett, I. Saleh, R. Malcolm, S. Thorpe, and T. Grandison, "Combining human and machine computing elements for analysis via crowdsourcing," in *Proc. 10th IEEE Int. Conf. Collaborative Comput., Netw., Appl. Worksharing*, Oct. 2014, pp. 312–321.

[16] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[17] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learn.*, vol. 28, nos. 2–3, pp. 133–168, 1997.

[18] N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer, "An evaluation of aggregation techniques in crowdsourcing," in *Proc. Int. Conf. Web Inf. Syst. Eng.* Berlin, Germany: Springer, 2013, pp. 1–15.

[19] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 399–407.

[20] D. Y. Zhang, Y. Huang, Y. Zhang, and D. Wang, "Crowd-assisted disaster scene assessment with human-ai interactive attention," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 3, pp. 2717–2724. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5658

[21] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," 2017, *arXiv:1705.10694*. [Online]. Available: http://arxiv.org/abs/1705.10694

[22] A. Fedorov, J. Johnson, E. Damaraju, A. Ozerin, V. Calhoun, and S. Plis, "End-to-end learning of brain tissue segmentation from imperfect labeling," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 3785–3792.

[23] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101693. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184152030058X

[24] D. C. Brabham, *Crowdsourcing*. Cambridge, MA, USA: MIT Press, 2013.

[25] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *Computer*, vol. 52, no. 1, pp. 36–45, Jan. 2019.

[26] B. Predic and D. Stojanovic, "Enhancing driver situational awareness through crowd intelligence," *Expert Syst. Appl.*, vol. 42, no. 11, pp. 4892–4909, Jul. 2015.

[27] R. Chunara, M. S. Smolinski, and J. S. Brownstein, "Why we need crowdsourced data in infectious disease surveillance," *Current Infectious Disease Rep.*, vol. 15, no. 4, pp. 316–319, Aug. 2013.

[28] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan, "Crowdsourcing cybersecurity: Cyber attack detection using social media," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 1049–1057.

[29] M. Srivastava, T. Abdelzaher, and B. Szymanski, "Human-centric sensing," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 370, no. 1958, pp. 176–197, 2012.

[30] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," *Commun. ACM*, vol. 54, no. 4, pp. 86–96, Apr. 2011.

[31] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal, "On credibility estimation tradeoffs in assured social sensing," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 6, pp. 1026–1037, Jun. 2013.

[32] D. Zhang, Y. Ma, X. Sharon Hu, and D. Wang, "Toward privacy-aware task allocation in social sensing-based edge computing systems," *IEEE Internet Things J.*, vol. 7, no. 12, pp. 11384–11400, Dec. 2020.

[33] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *Proc. IEEE 33rd Int. Conf. Distrib. Comput. Syst.*, Jul. 2013, pp. 530–539.

[34] Y. Zhang, H. Wang, D. Zhang, and D. Wang, "DeepRisk: A deep transfer learning approach to migratable traffic risk estimation in intelligent transportation using social sensing," in *Proc. 15th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, May 2019, pp. 123–130.

[35] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proc. ACM/IEEE 11th Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2012, pp. 233–244.

[36] J. Gao, Q. Li, B. Zhao, W. Fan, and J. Han, "Truth discovery and crowdsourcing aggregation: A unified perspective," in *Proc. VLDB Endowment*, vol. 8, no. 12, pp. 2048–2049, 2015.

[37] W. Li *et al.*, "Crowd intelligence in AI 2.0 era," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, no. 1, pp. 15–43, 2017.

[38] H. Mouzannar, Y. Rizk, and M. Awad, "Damage identification in social media posts using multimodal deep learning," in *Proc. ISCRAM*, 2018, pp. 529–543.

[39] M. T. Rashid, D. Y. Zhang, and D. Wang, "DASC: Towards a road damage-aware social-media-driven car sensing framework for disaster response applications," *Pervas. Mobile Comput.*, vol. 67, Sep. 2020, Art. no. 101207.

[40] Y. Zhang, R. Zong, and D. Wang, "A hybrid transfer learning approach to migratable disaster assessment in social media sensing," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Dec. 2020, pp. 131–138.

[41] Y. Zhang, R. Zong, Z. Kou, L. Shang, and D. Wang, "A crowd-driven dynamic neural architecture searching approach to quality-aware streaming disaster damage assessment," in *Proc. IEEE/ACM 29th Int. Symp. Qual. Service (IWQOS)*, Jun. 2021, pp. 1–6.

[42] M. T. Rashid, D. Y. Zhang, and D. Wang, "SocialDrone: An integrated social media and drone sensing system for reliable disaster response," in *Proc. IEEE Conf. Comput. Commun. (IEEE INFOCOM)*, Jul. 2020, pp. 218–227.

[43] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[44] A. Guo, A. Jain, S. Ghose, G. Laput, C. Harrison, and J. P. Bigham, "Crowd-AI camera sensing in the real world," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–20, Sep. 2018.

[45] J. Yang, A. Smirnova, D. Yang, G. Demartini, Y. Lu, and P. Cudre-Mauroux, "Scalpel-CD: Leveraging crowdsourcing and deep probabilistic modeling for debugging noisy training data," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 2158–2168.

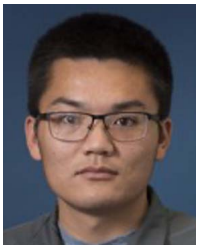
[46] Z. Kou, D. Y. Zhang, L. Shang, and D. Wang, "ExFaux: A weakly supervised approach to explainable fauxtography detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 631–636.

[47] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 17, 2021, doi: 10.1109/TCSVT.2021.3080920.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[49] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9522–9531.

[50] B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K.-K.-R. Choo, and M. Jamshidi, "Toward artificial emotional intelligence for cooperative social human–machine interaction," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 1, pp. 234–246, Feb. 2019.
[51] X. Liu *et al.*, "Collaborative edge computing with FPGA-based CNN accelerators for energy-efficient and time-aware face tracking system," *IEEE Trans. Computat. Social Syst.*, early access, Feb. 25, 2021, doi: 10.1109/TCSS.2021.3059318.
[52] H. Zhu, G. Han, L. Shu, and H. Zhao, "ArvaNet: Deep recurrent architecture for PPG-based negative mental-state monitoring," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 1, pp. 179–190, Feb. 2021.
[53] R. Dong, L. Li, Q. Zhang, and G. Cai, "Information diffusion on social media during natural disasters," *IEEE Trans. Computat. Social Syst.*, vol. 5, no. 1, pp. 265–276, Mar. 2018.
[54] L. Wang, H. Liu, Y. Lu, H. Chen, J. Zhang, and J. Pu, "A coarse-to-fine deep learning framework for optic disc segmentation in fundus images," *Biomed. Signal Process. Control*, vol. 51, pp. 82–89, May 2019.
[55] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. ICML*, 2016, vol. 2, no. 3, p. 7.
[56] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.
[57] A. Graves, "Generating sequences with recurrent neural networks," 2013, *arXiv:1308.0850*. [Online]. Available: http://arxiv.org/abs/1308.0850
[58] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[59] C. Coleman *et al.*, "Selection via proxy: Efficient data selection for deep learning," 2019, *arXiv:1906.11829*. [Online]. Available: http://arxiv.org/abs/1906.11829
[60] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: http://arxiv.org/abs/1804.03999

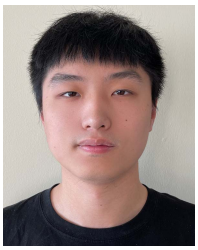


Yang Zhang (Graduate Student Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in 2013, and the M.S. degree from Indiana University at Bloomington, Bloomington, IN, USA, in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA.

His research interests include social sensing, deep learning, and human-centered artificial intelligence.

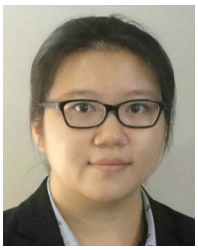
Ruohan Zong (Student Member, IEEE) received the B.S. degree in computer science from Sichuan University, Chengdu, China, in 2020. She is currently pursuing the master's degree in computer science with Columbia University, New York, NY, USA.

Her research interests include machine learning, deep learning, and image processing.



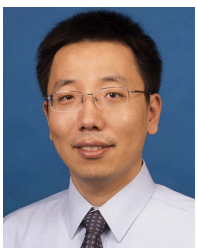
Ziyi Kou (Student Member, IEEE) received the B.S. degree in software engineering from Chongqing University, Chongqing, China, in 2018, and the M.S. degree in computer science from the University of Rochester, Rochester, NY, USA, in 2020. He is currently pursuing the Ph.D. degree with the School of Information Sciences, University of Illinois Urbana–Champaign, Champaign, IL, USA.

His research interest lies in the intersection of social sensing, computer vision, and human-in-the-loop systems.



Lanyu Shang (Student Member, IEEE) received the B.S. degree in applied mathematics from the University of California at Los Angeles, Los Angeles, CA, USA, in 2014, and the M.S. degree in data science from New York University, New York, NY, USA, in 2017. She is currently pursuing the Ph.D. degree with the School of Information Sciences, University of Illinois Urbana–Champaign, Champaign, IL, USA.

Her research interest primarily lies in online misinformation detection using social media data.



Dong Wang (Member, IEEE) received the Ph.D. degree in computer science from the University of Illinois Urbana–Champaign (UIUC), Champaign, IL, USA, in 2012.

He is currently an Associate Professor with the School of Information Sciences, UIUC. His research interests lie in the area of reliable social sensing, human-centric AI, cyber-physical computing, and smart city applications.

Dr. Wang received the Best Paper Award of the IEEE Real-Time and Embedded Technology and Applications Symposium in 2010, the Wing-Kai Cheng Fellowship from UIUC in 2012, the Army Research Office Young Investigator Program Award in 2017, the Google Faculty Research Award in 2018, and the NSF CAREER Award in 2019. He is a member of the ACM.